

A Hitchhikers Guide to Reproducible Research

Thursday 20th February 2020

Brendan Palmer,

Clinical Research Facility - Cork &
School of Public Health

 @B_A_Palmer

Talk materials

github.com/bapalmer/lunchtime_sessions

Search or jump to... / Pull requests Issues Marketplace Explore

bapalmer / lunchtime_sessions Watch 2 Star 3 Fork 0

Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security Insights Settings

Short 1 hour introductions to R-related topics such as creating R projects, using GitHub through RStudio and more Edit

Manage topics

50 commits 1 branch 0 packages 0 releases 1 contributor MIT

Branch: master New pull request Create new file Upload files Find file Clone or download

File	Commit	Time
bapalmer Binder link	57af84c	2 minutes ago
Session_1-R_projects	PG6015	Dec 2019 8 minutes ago
Session_2-Reproducible_reports	PG6015	Dec 2019 8 minutes ago
Session_3-Git_and_RStudio	PG6015	Dec 2019 8 minutes ago
Session_4-OS_and_reproducible_research	PG6015	Dec 2019 8 minutes ago

Can we believe what we see in the literature?



PERSPECTIVE



The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications

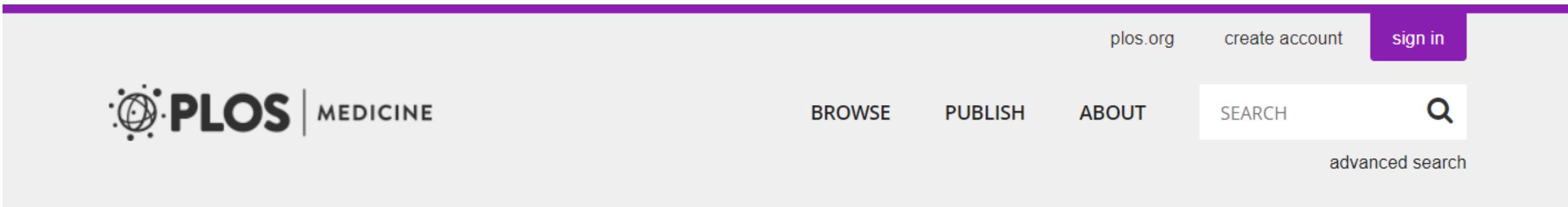
Elisabeth M. Bik,^a Arturo Casadevall,^{b,c} Ferric C. Fang^d

Department of Medicine, Division of Infectious Diseases, Stanford School of Medicine, Stanford, California, USA^a; Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA^b; Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA^c; Departments of Laboratory Medicine and Microbiology, University of Washington School of Medicine, Seattle, Washington, USA^d

ABSTRACT Inaccurate data in scientific papers can result from honest error or intentional falsification. This study attempted to determine the percentage of published papers that contain inappropriate image duplication, a specific type of inaccurate data.

The images from a total of 20,621 papers published in 40 scientific journals from 1995 to 2014 were visually screened. Overall, 3.8% of published papers contained problematic figures, with at least half exhibiting features suggestive of deliberate manipulation. The prevalence of papers with problematic images has risen markedly during the past decade. Additional papers written by authors of papers with problematic images had an increased likelihood of containing problematic images as well. As this analysis focused only on one type of data, it is likely that the actual prevalence of inaccurate data in the published literature is higher. The marked variation in the frequency of problematic images among journals suggests that journal practices, such as prepublication image screening, influence the quality of the scientific literature.

Can we believe what we read in the literature?



The image shows the header of the PLOS Medicine website. At the top right are links for "plos.org", "create account", and "sign in". Below that is a search bar with a magnifying glass icon and a link to "advanced search". The main navigation menu includes "BROWSE", "PUBLISH", and "ABOUT". To the left is the PLOS logo with "MEDICINE" next to it. A purple horizontal bar spans the width of the header.

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>

68,836 Save	2,931 Citation
2,768,586 View	10,482 Share

And publication bias is very real



Rink Hoekstra
@RinkHoekstra

Follow ▾

Elsevier editor Spada acknowledging that null results are not even considered for Addictive Behaviors, seemingly not realizing how problematic that is. Offering a lower prestige alternative journal doesn't make that right.



Professor M. M. Spada said:

"Articles that may not traditionally be considered by Addictive Behaviors, including negative/null data papers, studies using smaller samples and cross-sectional designs, replication studies, cross-cultural research, and case reports will be welcome by its sister journal Addictive Behaviors Reports."

Editor-in-Chief
Professor M. M. Spada
London South Bank University

Journal Metrics

> CiteScore: 3.10 ⓘ

Impact Factor: 2.686 ⓘ

Journal Metrics

> CiteScore: 2.11 ⓘ

...and this is where we put the non-significant results.



someecards
user card

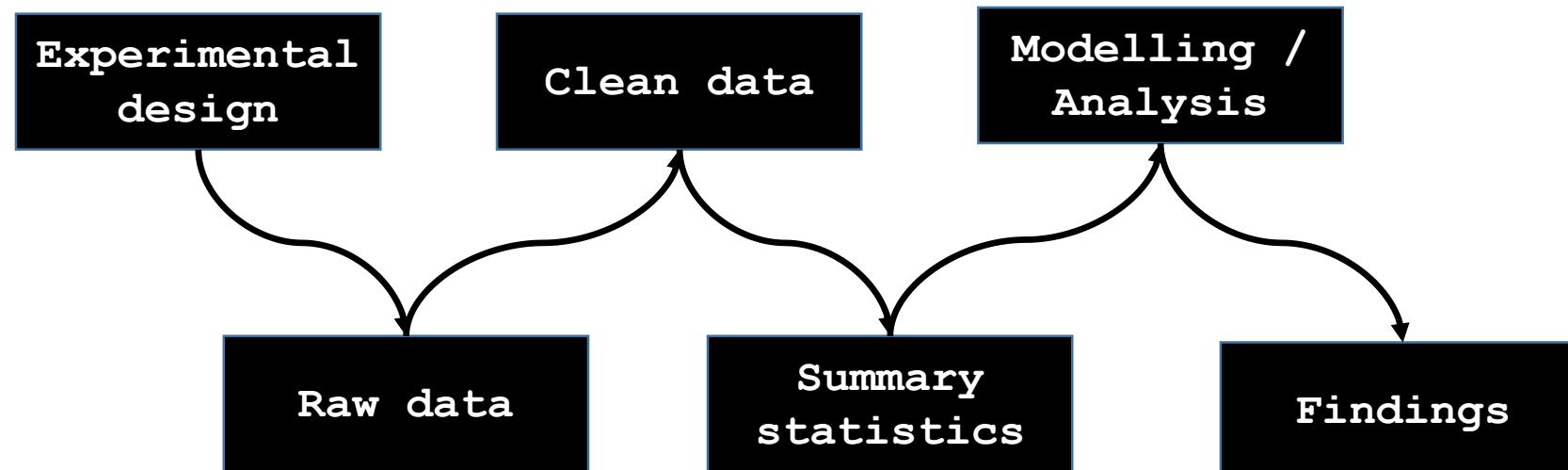
p-values do not define a study

The screenshot shows the homepage of the journal **nature**, which is described as an "International weekly journal of science". The top navigation bar includes links for Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and For Authors. Below this, a breadcrumb navigation shows the path: Archive > Volume 520 > Issue 7549 > Comment > Article. On the right side of the header are search fields for "Search" and "Advanced search", along with social media links for E-alert, RSS, Facebook, and Twitter. The main content area features a large, bold title: "Statistics: *P* values are just the tip of the iceberg". Below the title, the authors are listed as "Jeffrey T. Leek & Roger D. Peng", and the publication date is given as "28 April 2015".

Statistics: *P* values are just the tip of the iceberg

Jeffrey T. Leek & Roger D. Peng

28 April 2015



Experimental
design

Clean data

Modelling /
Analysis

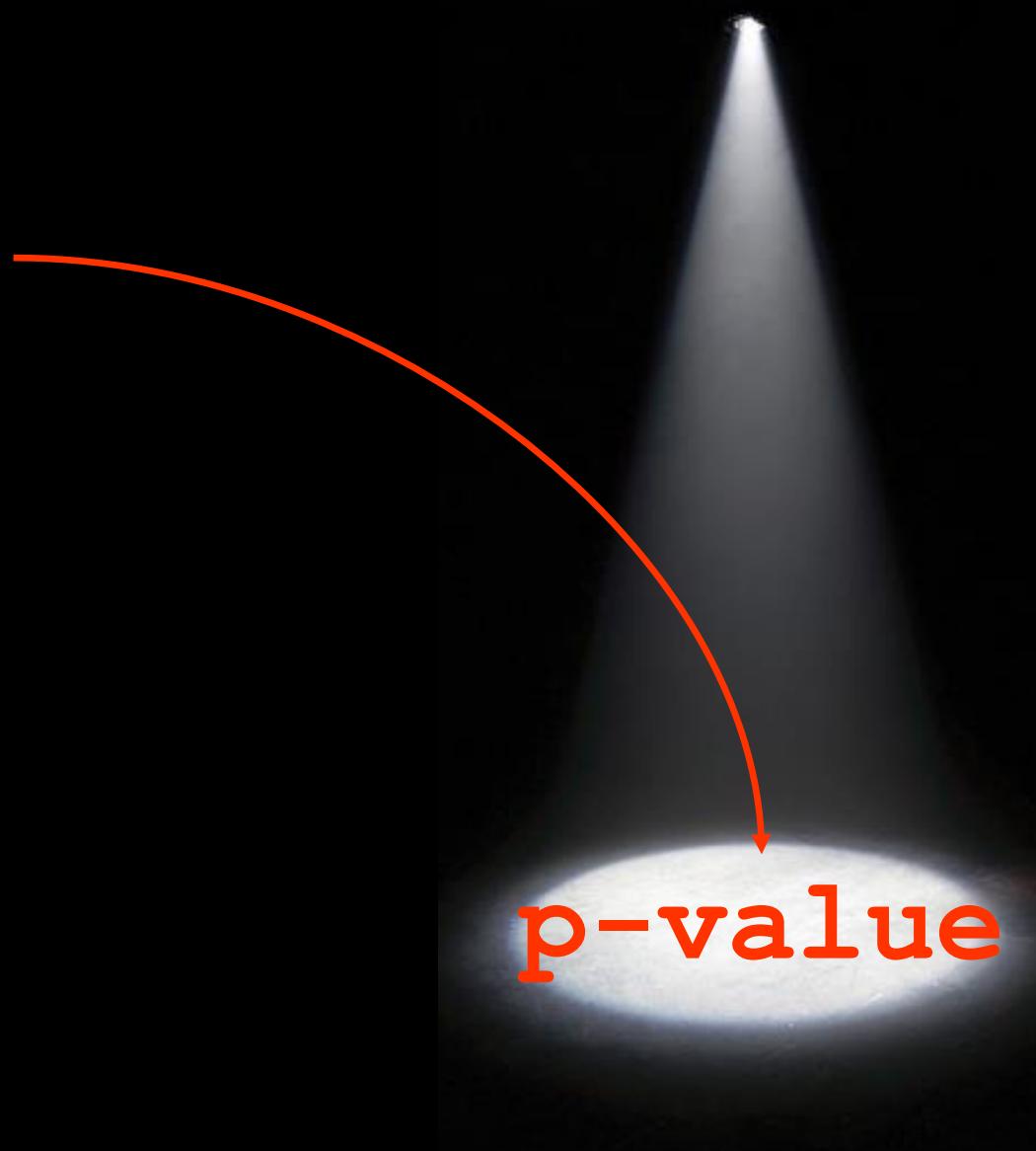
Raw data

Summary
statistics

Findings

Little debate

Extreme
scrutiny



Ig Nobel research - 2012 Neuroscience winner



Journal of Serendipitous and Unexpected Results

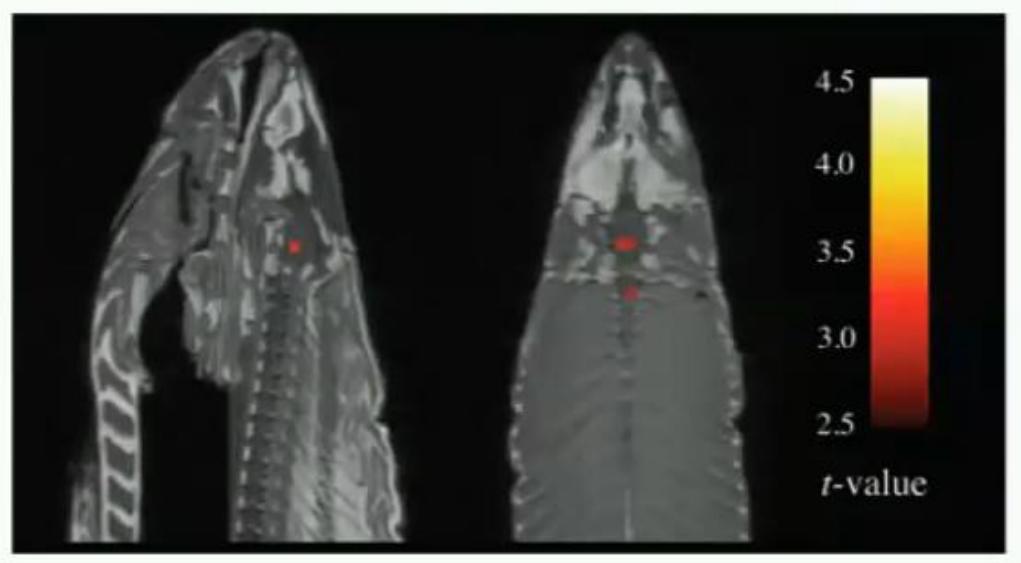
Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett^{1*}, Abigail A. Baird², Michael B. Miller¹ and George L. Wolford³

¹Department of Psychology, University of California at Santa Barbara, Santa Barbara, CA 93106

²Department of Psychology, Blodgett Hall, Vassar College, Poughkeepsie, NY 12604

³Department of Psychological and Brain Sciences, Moore Hall, Dartmouth College, Hanover, NH 03755



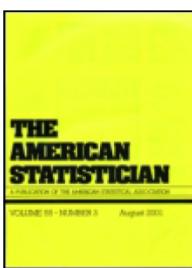
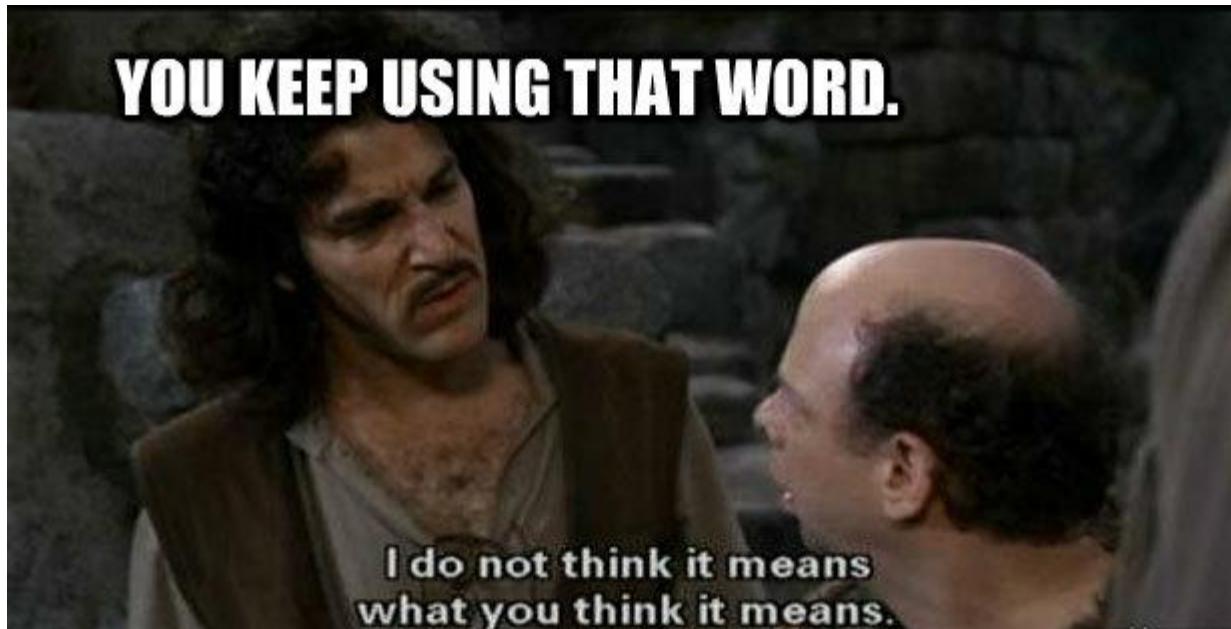
One mature Atlantic Salmon (*Salmo salar*) participated in the fMRI study. The salmon measured approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning. It is not known if the salmon was male or female, but given the post-mortem state of the subject this was not thought to be a critical variable.

The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence, either socially inclusive or socially exclusive. The salmon was asked to determine which emotion the individual in the photo must have been experiencing.

Several active voxels were observed in a cluster located within the salmon's brain cavity (see Fig. 1). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$.

Either we have stumbled onto a rather amazing discovery in terms of post-mortem ichthyological cognition, or there is something a bit off with regard to our uncorrected statistical approach.

Significant



The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: <http://amstat.tandfonline.com/loi/utas20>

The ASA's Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar



Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3



ESSAY

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland¹ · Stephen J. Senn² · Kenneth J. Rothman³ · John B. Carlin⁴ · Charles Poole⁵ · Steven N. Goodman⁶ · Douglas G. Altman⁷

Received: 9 April 2016/Accepted: 9 April 2016/Published online: 21 May 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific

literature. In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the *P* values they produce) can lead to small *P* values even if the declared test hypothesis is correct, and can lead to large *P* values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of *P* values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.

Editor's note This article has been published online as supplementary material with an article of Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process and purpose. *The American Statistician* 2016.

Albert Hofman, Editor-in-Chief EJE.

✉ Sander Greenland
lesdomes@ucla.edu

Stephen J. Senn
stephen.senn@lih.lu

John B. Carlin
john.carlin@mcri.edu.au

Charles Poole
cpoole@unc.edu

Steven N. Goodman
steve.goodman@stanford.edu

Douglas G. Altman
doug.altman@csm.ox.ac.uk

¹ Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA

² Competence Center for Methodology and Statistics, Luxembourg Institute of Health, Strassen, Luxembourg

³ RTI Health Solutions, Research Triangle Institute, Research Triangle Park, NC, USA

⁴ Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, School of Population Health, University of Melbourne, Melbourne, VIC, Australia

⁵ Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA

⁶ Meta-Research Innovation Center, Departments of Medicine and of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA

⁷ Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

What is your research hypothesis?

Article

Too True to be Bad: When Sets of Studies With Significant and Nonsignificant Findings Are Probably True

Daniël Lakens¹ and Alexander J. Etz²

Consider the following example:

- Starting out, H_0 and H_1 are equally likely
 - α is controlled for 0.05
 - The study has 80% power
 - What is the most likely outcome?
A - True positive, B - True negative,
C - False Positive, D - False negative

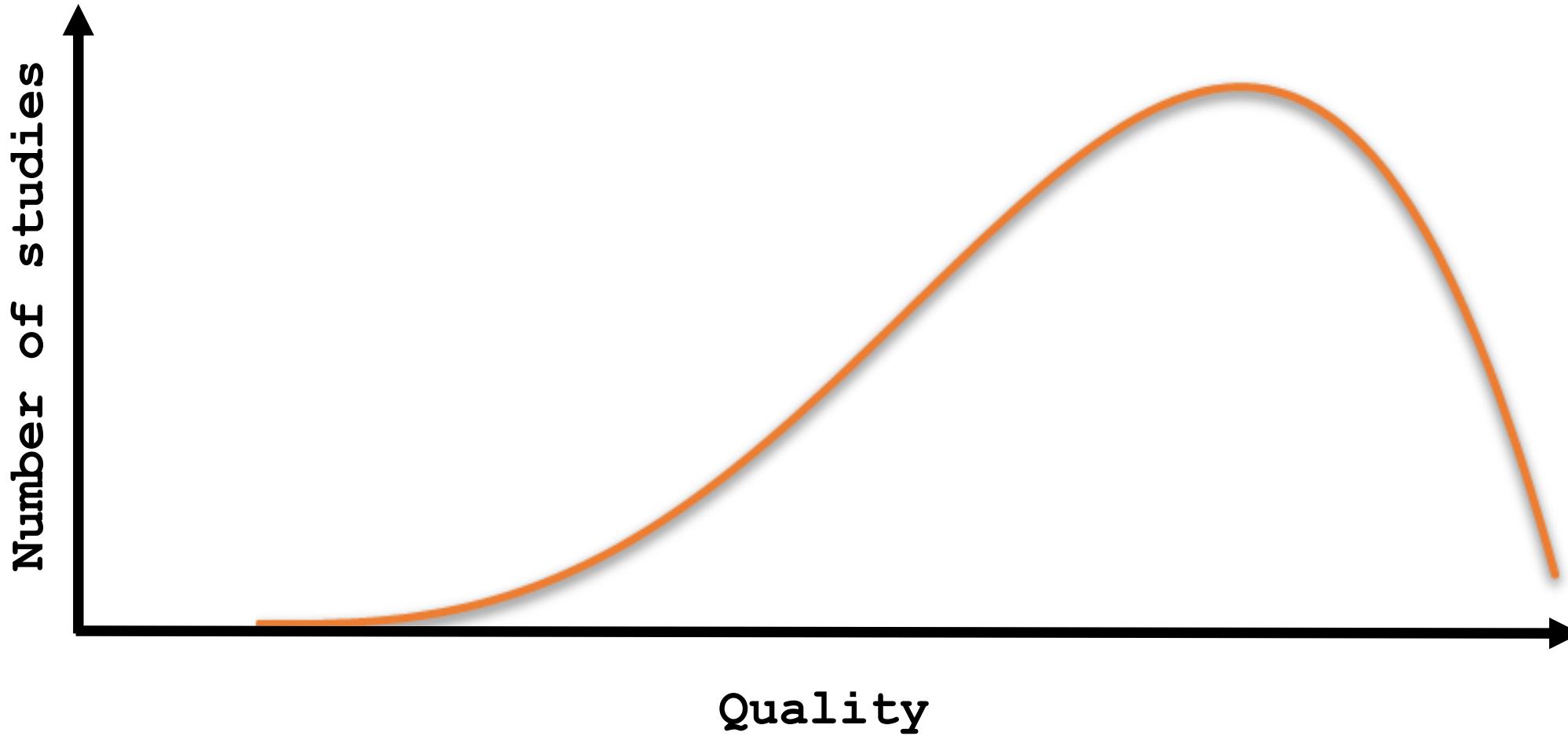
Social Psychological and Personality Science
2017, Vol. 8(8) 875-881
© The Author(s) 2017
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1948550617693058
journals.sagepub.com/home/spp

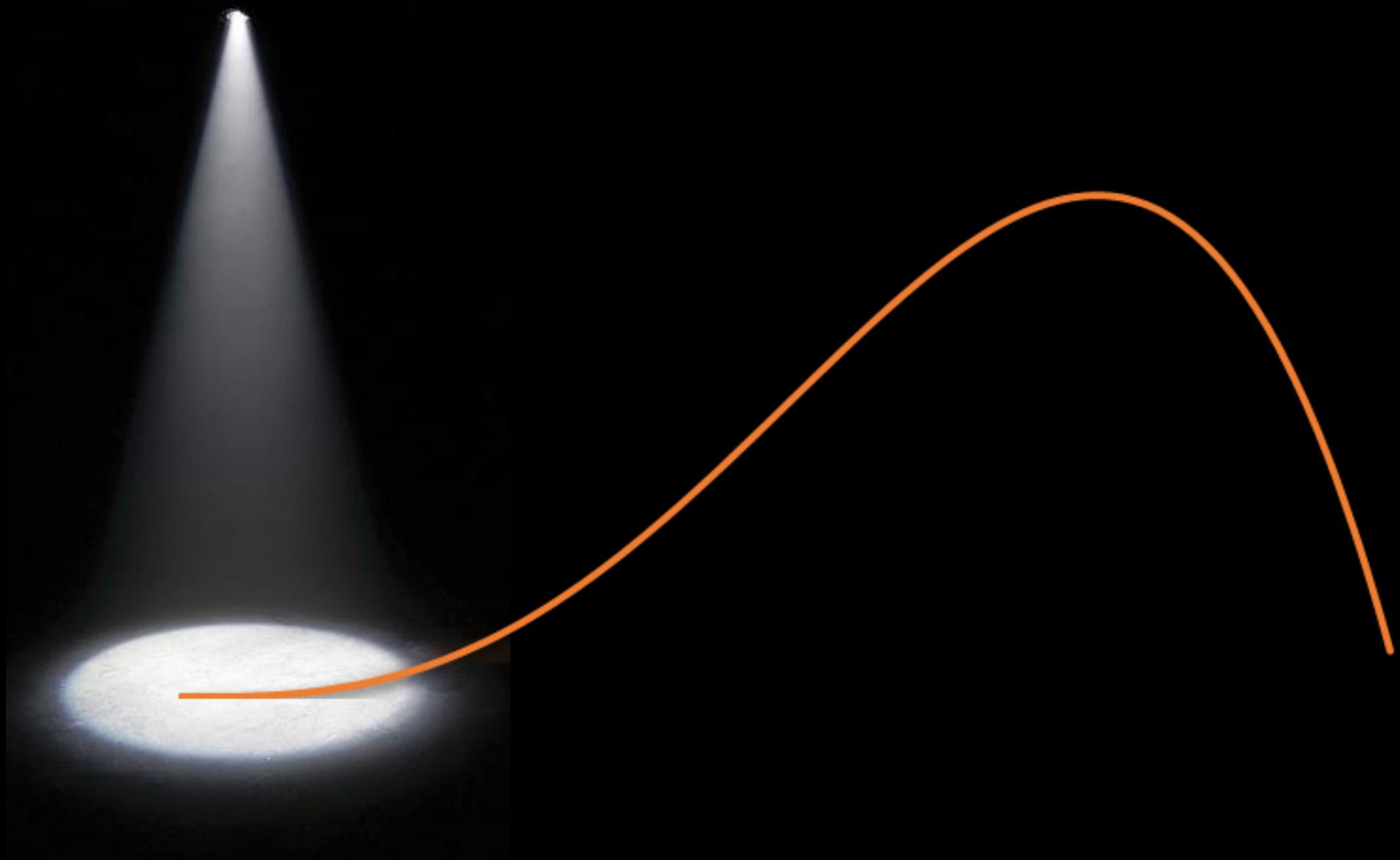


Result

- What is the most likely outcome?
A – True positive – 40%
B – True negative – 47.5%
C – False Positive – 2.5%
D – False negative – 10%
- What is the best strategy to improve this outcome?
 - Pick a better hypothesis
 - If our alternate hypothesis is more likely...
 - $H_0 = 40\%$ and $H_1 60\%$
A – True positive – 48%
B – True negative – 38%
C – False Positive – 2%
D – False negative – 12%

Today





Don't do what Donny Dont does!



"In short, peer review misses all the hard stuff, and a worrying amount of the easy stuff"

James Heathers,
Northwestern University

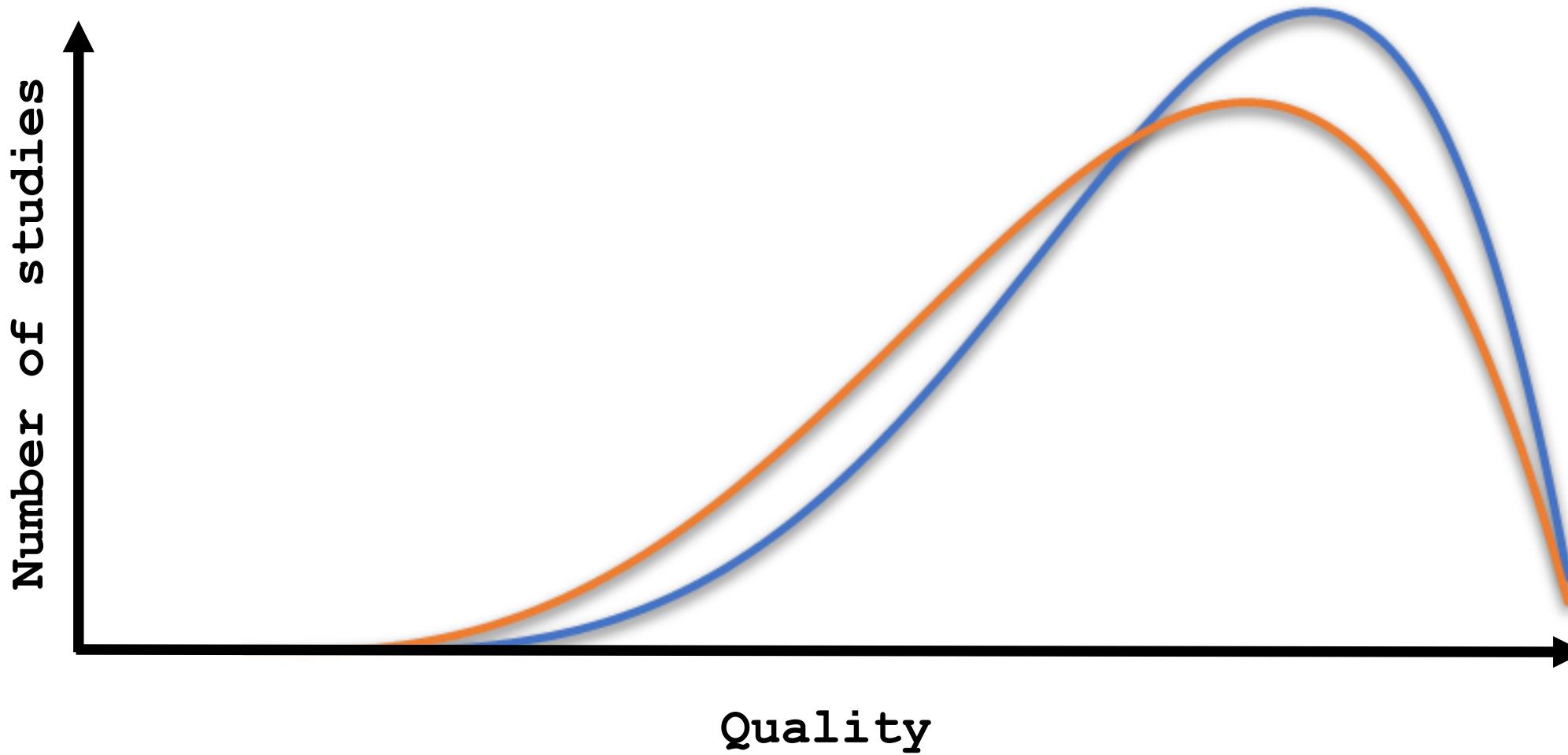
#datathugs



Brian Wansink: The grad student who never said no

"Every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions"

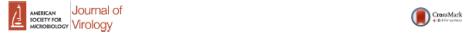
Can we shift the distribution?



How is research presented?

Theses

Papers



Network Analysis of the Chronic Hepatitis C Virome Defines Hypervariable Region 1 Evolutionary Phenotypes in the Context of Humoral Immune Responses

Brendan A. O'Farrior,¹ Daniel Schmidt-Martin,² Zoya Dimitrova,² Pavel Skums,³ Orla Crosbie,⁴ Elizabeth Kenny-Walsh,⁵ Liam J. Fanning⁶

Molecular Virology Diagnostic & Research Laboratory, Department of Medicine, University College Cork, Cork, Ireland; Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia, USA; Department of Reproductive, Cork University Hospital, Cork, Ireland

ABSTRACT
Hypervariable region 1 (HVR1) of hepatitis C virus (HCV) comprises the first 27 N-terminal amino acid residues of E2. It is classically recognized as the major antigenic target of the humoral immune response. HVR1 undergoes rapid sequence evolution in chronically infected patients over a short, 10-week period. Organization of the sequence set into connected components that represented single nucleotide substitution events revealed a network dominated by highly connected, centrally positioned master sequences. HVR1 phenotypes were observed to be under strong purifying (stationary) and strong positive (antigenic drift) selection pressures, which were consistent with advancing patient age and disease course of HCV infection. It was found that the HVR1 network structure was dominated by stationary variants, which were composed from conservative single amino acid substitution events. We present evidence to suggest that neutralization antibody efficacy was diminished for stationary-virome HVR1 variants. Our results identify the HVR1 network structure during chronic hepatitis C as the predominant dominance of a single variant within a narrow sequence space.

IMPORTANCE
Hepatitis C virus (HCV) is often asymptomatic, and chronic infection is generally well established in advance of initial diagnosis and subsequent treatment. HVR1 can undergo rapid sequence evolution during acute infection, and the variant pool is typically seen to diverge away from ancestral sequences as infection progresses from the acute to the chronic phase. In this report, we describe HVR1 viromes in chronically infected patients that are defined by a dominant epitope located centrally within a narrow variant pool. Our findings suggest that weakened humoral immune activity, as a consequence of persistent chronic infection, allows for the acquisition and maintenance of host-specific adaptive mutations at HVR1 that reflect virus fitness.

Hepatitis C virus (HCV) infection is a global health burden and a major etiological agent of liver related diseases worldwide. In the United States alone, the estimated prevalence of HCV represents approximately 2% of the global adult (15 years of age and older) population (1). Following transmission, HCV infection is characterized by a long incubation period, with many infections initially passing undetected (2). It is estimated that up to 4 million Americans are living with the virus, the majority of whom became infected prior to the first clinical identification of the virus (3, 4). Currently, the U.S. Centers for Disease Control and Prevention now recommend that Americans born from 1945 to 1965 be screened for the presence of the virus, as this is the largest cohort of chronically infected individuals (5).

HCV is a single-stranded positive-sense RNA virus of considerable genomic heterogeneity. A recent reclassification defines the major genotypes 1a and 1b and 67 subtypes within genotypes 1 and 5 accounting for the majority of infections worldwide (6, 7). An error-prone RNA-dependent RNA polymerase, together with an inherent propensity of defining hypervariable regions (HVRs), is responsible for much of this variability. These HVRs are located within the envelope glycoprotein E2 (residues 456–656), the terminal end of the E2 glycoprotein (8). Recent studies indicated that the central region of E2 (residues 456 to 656) is globular and surprisingly compact, whereas the first 80 amino acids (including

Received 25 November 2015; Accepted 22 December 2015
Accepted manuscript posted online 10 December 2015
Editorial decision received 10 December 2015
Editorial decision received 10 December 2015
Editor: M. S. Diamond
Associate Editor: L. J. Fanning
Editorial Review Committee: L. J. Fanning, H. Hammill, G. L. Jackson, R. A. P. and D. S. M. contributed equally to this article
Copyright © 2016, American Society for Microbiology. All Rights Reserved.

3318 J. Virol. 2016; 90:3318–3326

April 2016

Volume 90 Number 7

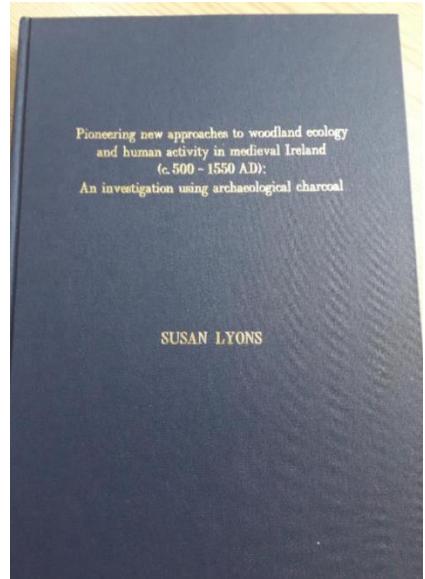
Books



Talks



Posters



But what does it really look like?



Putting the pieces together

A: Define a project structure

B: Set a naming convention

C: Use scripted workflows

D: Digital notebooks

E: Version control

F: Data packaging

Reproducible
research

Still haven't found what I'm looking for

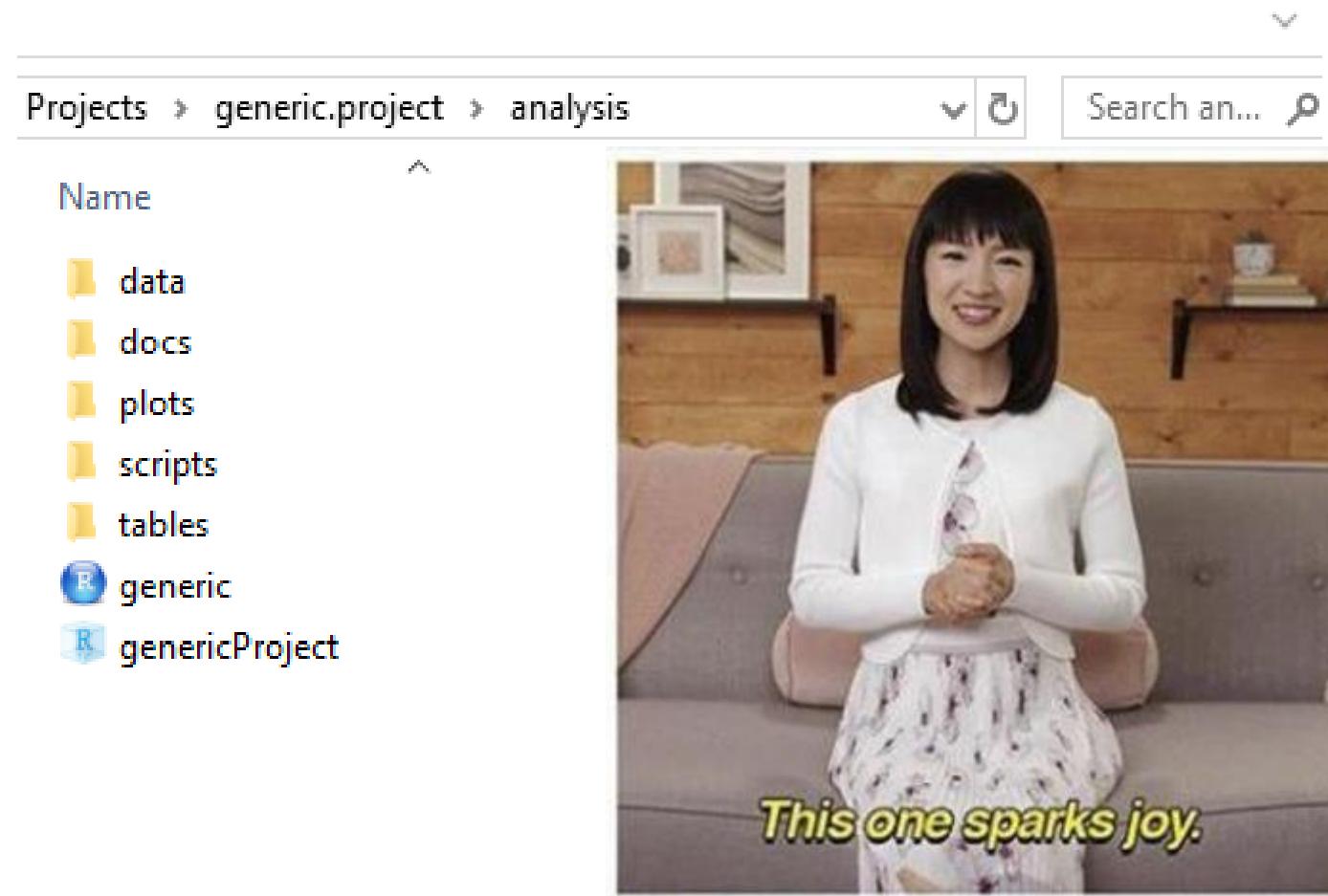
- Help your future-self

B_Palmer_Medicine_Files > 4a Project > Pyrosequencing_analysis > Pyrosequencing_Paper > Draft_Paper_incl_Figs > Submission > JVI_Resubmission > JVI_resubmission_files > Final Final version

Name	Date modified
Cover_letter_B_A_Palmer_Sept_2014	10/09/2014 17:05
Fig_1_Sept_14	11/09/2014 10:31
Fig_1_Sept_14	10/09/2014 23:07
Fig_2_Sept_14	11/09/2014 10:31
Fig_2_Sept_14	10/09/2014 23:07
Fig_3_Sept_14	11/09/2014 10:31
Fig_3_Sept_14	10/09/2014 23:07
Fig_4_Sept_14	11/09/2014 10:31
Fig_4_Sept_14	10/09/2014 23:07
Fig_5_Sept_14	11/09/2014 10:33
Fig_5_Sept_14	10/09/2014 23:07
HCV_UDPS_B_A_Palmer_Sept_14	17/09/2014 12:21
Response_to_Reviewer_Sept_14	10/09/2014 22:42
Supplementary_Figure_B_A_Palmer_Sept_14	29/08/2014 13:21
Supplementary_Figure_B_A_Palmer_Sept_14	10/09/2014 22:31
Tables_B_A_Palmer_Sept_2014	10/09/2014 22:09



A: Define a generic project structure



B: Give your files and folders informative names

This PC > Documents > Projects > **2016-08-08_RespPCT** > analysis > data

Name	Date modified
raw_data	21/01/2019 21:06
2018-11-06_abx	06/11/2018 13:10
2018-11-06_monitoring	06/11/2018 13:09
2018-11-06_pct	06/11/2018 13:08
2018-11-06_pt_info	06/11/2018 13:07

Everything in its right place

- Make your file names:
 1. Machine readable
 2. Human readable
 3. Work with default ordering

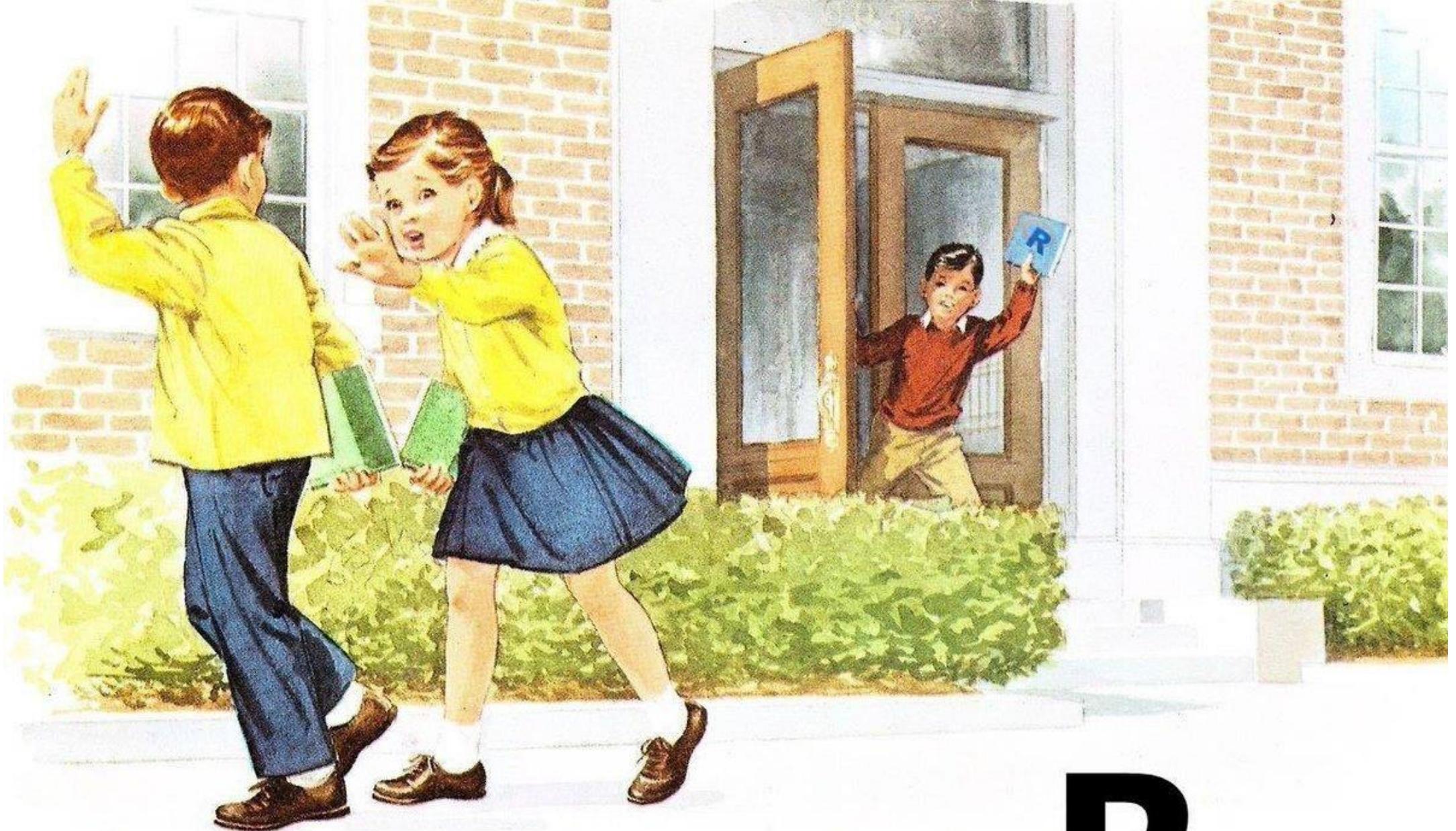
NO

Name
All unique 4a amino acid Sequences (B-N).fas
All unique 4a amino acid Sequences (B-N).meg
All_AA_haplotypes.meg
All_AA_haplotypes_with_clonal_sequences.meg
BS100_AA_with_clones
BS100_AA_with_clones.nwk
BS1000_AA_pyro&clones
BS1000_AA_pyro&clones.nwk
BS1000_AA_pyro_only
BS1000_AA_pyro_only.nwk
BS1000_Uncle_Clonal_AA

Yes

Projects > 2016-08-08_RespPCT > analysis > scripts

Name
R 01_clean_data
R 02_plots
R 03_tables
R 04_stats_analysis
R 05_post_hoc_stats
R functions
R randomization
R tables



Run, or he's going to tell us about
again!

R

C: Use scripted workflows

- The R scripts should also be human readable
 - Annotate the code
 - Break up the scripts into dedicated tasks
 - Interlink to other project scripts

```
1 # Data ----
2 # Eight tibbles returned from the 01_data_import_and_tidying_master_file.R
3 # 1. fgf23_data => FGF23 readings from study centres 01-03
4 # 2. food_level_data => Food diary entries
5 # 3. grouped_data => Dialysis and nondialysis diary entries by component
6 # 4. k_data => Serum potassium
7 # 5. master_data_clean => all the clean master file data if required
8 # 6. p_data => Serum phosphate
9 # 7. pth_data => Parathyroid hormone readings
10 # 8. pulses_nuts_data
11
12 source("scripts/01_data_import_and_tidying_master_file.R")
```

Work from the raw data ALWAYS!!



Tom Webb @tomjwebb · 16 Jan 2015

If you could tell a new PhD student one thing to help make their data more useful/shareable, what would it be?

27

11

7



Dr Gavin Simpson

@ucfagls

Follow

Replying to @tomjwebb

@tomjwebb don't, not even with a barge pole, not for one second, touch or otherwise edit the raw data files. Do any manipulations in script

7:15 AM - 16 Jan 2015

D: R Markdown

R ~/Open_Science/Digital_Badge/RCR - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

lettuce_report.Rmd* Go to file/function Addins

```
1 ---  
2 title: "This is a reproducible document"  
3 author: "Dr. Brendan Palmer"  
4 date: "18th June 2019"  
5 output:  
6   word_document:  
7     fig_height: 4  
8     fig_width: 6  
9 ---  
10 # This is the beginning of the project  
11  
12 our initial reports might be restricted to lab meetings etc. We can use `R  
13 Markdown` to show the code we are using, so that the meetings are not just a  
14 demonstration of the results, but also an examination of the `code` used to obtain  
15 them.  
16  
17 knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)  
18  
19 # Load your packages here  
20 library(tidyverse)  
21 library(knitr)  
22  
23  
24 The plot below is call from the ggplot object entitled `report_plot` created in  
25 the script `03_final_analysis.R`.  
26  
27 {r Plots from script, echo = FALSE}  
28 source("scripts/03_final_analysis.R")  
29  
30 # The location of the Rmd file dictates whether the path to other files is intact
```

This is a reproducible document

Dr. Brendan Palmer

18th June 2019

This is the beginning of the project

Our initial reports might be restricted to lab meetings etc. We can use R Markdown to show the code we are using, so that the meetings are not just a demonstration of the results, but also an examination of the code used to obtain them.

Data overview

The plot below is call from the ggplot object entitled report_plot created in the script 03_final_analysis.R.

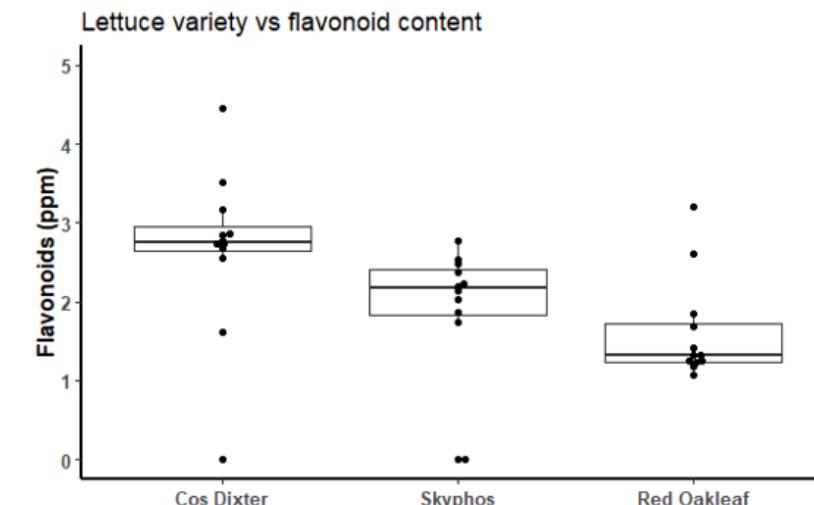
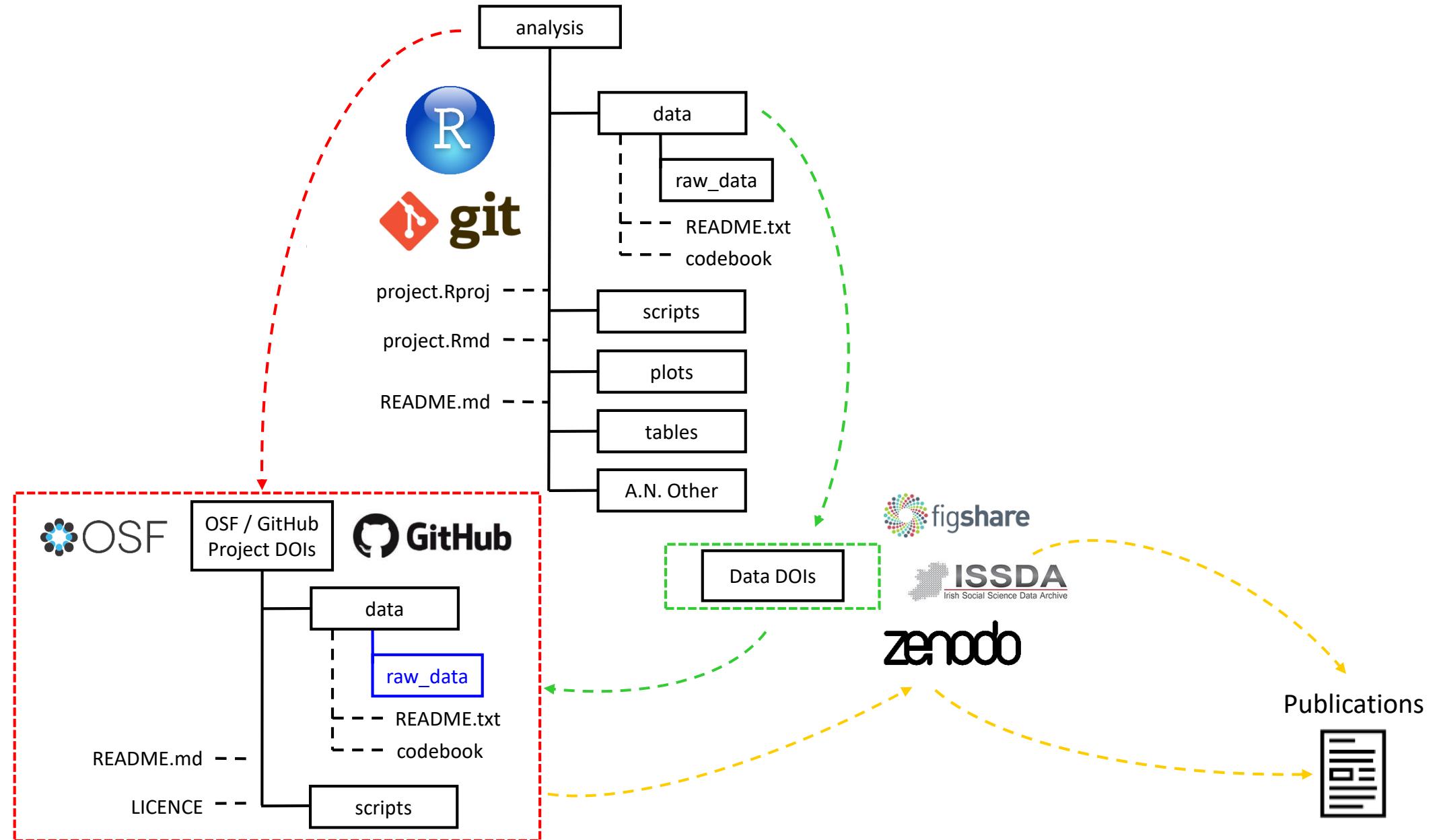


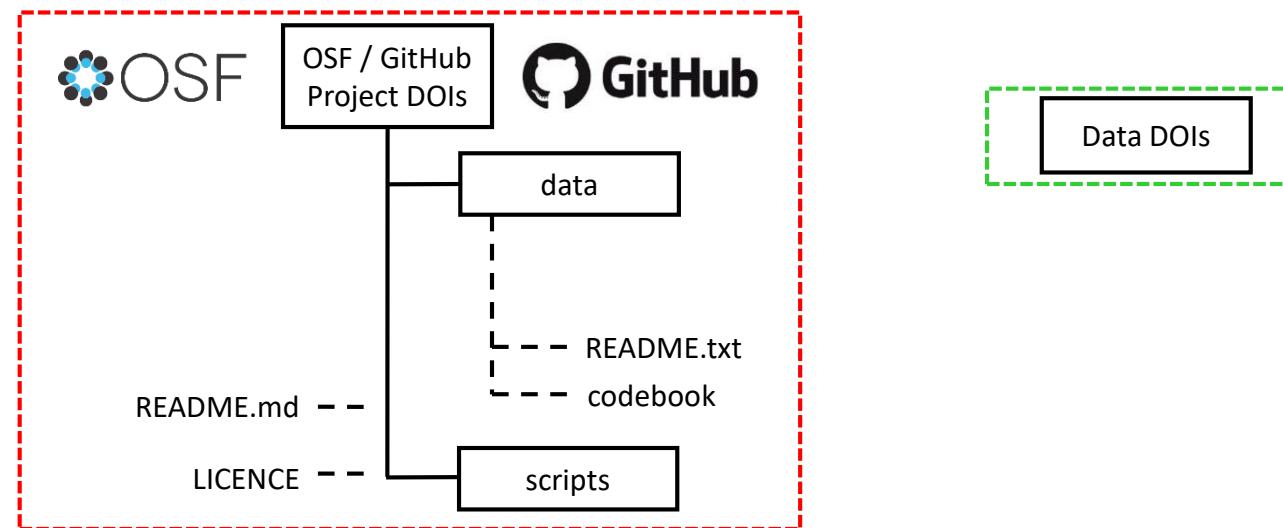
Fig. 1. Flavonoid content of three lettuce varieties under three experimental conditions.

Or we can also recreate the code within the R Markdown document as seen below.

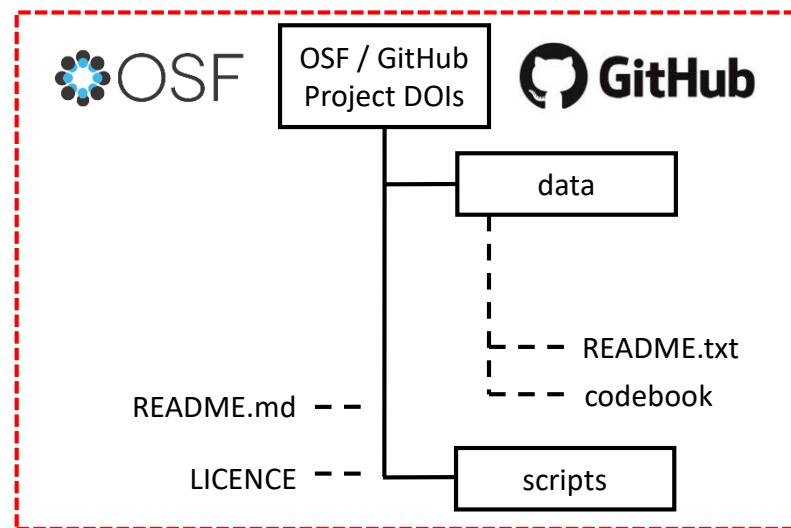
What does this allow us to do?



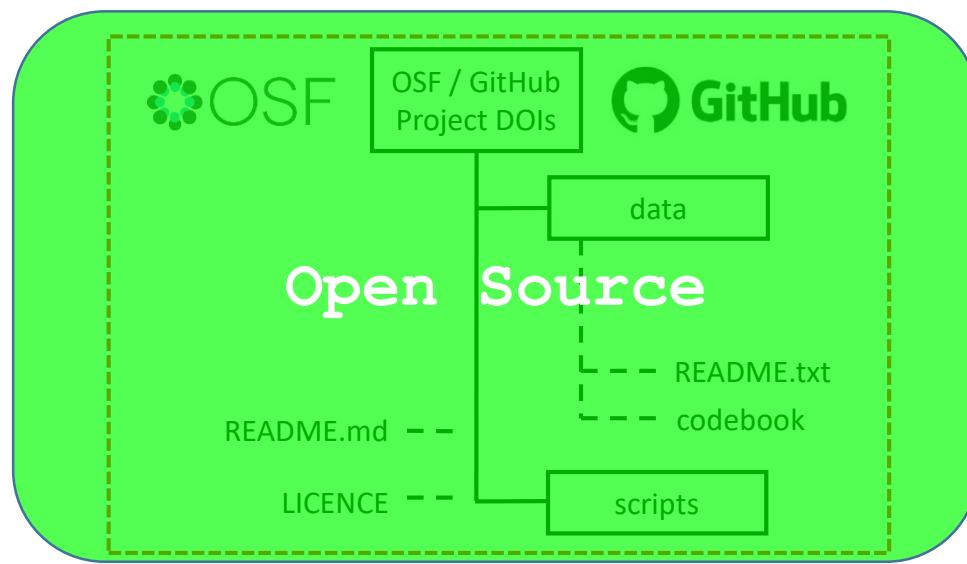
What does this allow us to do?



What does this allow us to do?

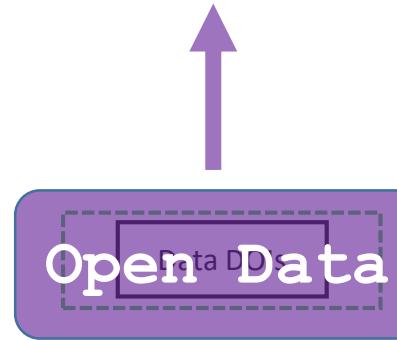
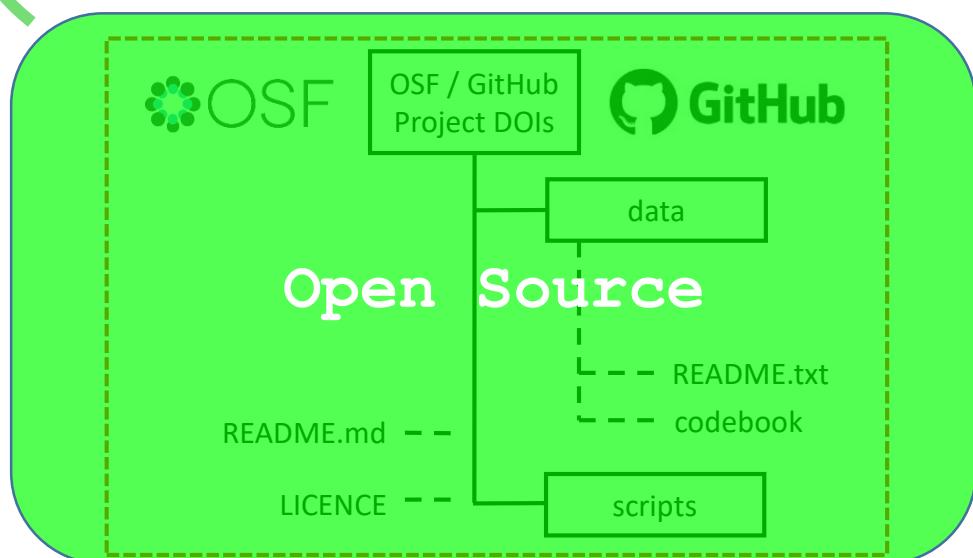


What does this allow us to do?



What does this allow us to do?

Open Materials



Install the Chrome plugin PubPeer

NCBI Resources ▾ How To ▾ Sign in to NCBI

PubMed ▾ Is the Power Threshold of 0.8 Applicable to Surgical Science? Search

PubMed.gov US National Library of Medicine National Institutes of Health Create RSS Create alert Advanced Help

Format: Abstract ▾ Send to ▾

See 1 citation found by title matching your search:

J Surg Res. 2019 Apr 26;241:235-239. doi: 10.1016/j.jss.2019.03.062. [Epub ahead of print]

23 comments on PubPeer (by: Andrew D. Althouse, Thom Baguley, Guillaume A. Rousselet, Timothy Feeney, Paul M Brown, Frank E. Harrell, David Nunan, Samantha R. Seals, Raj Mehta, Yevgeniy Feyman, Ionomidotis Irregularis, Andrew Gelman, Aleksi Reito, Daniel E. Leisman, Pavlos Msaouel, Ryan Miller, Maarten Van Smeden, Zad Rafi Chow)

Is the Power Threshold of 0.8 Applicable to Surgical Science?-Empowering the Underpowered Study.

Bababekov YJ¹, Hung YC², Hsu YT², Udelsman BV², Mueller JL², Lin HY², Stapleton SM², Chang DC².

Author information

Abstract

BACKGROUND: Many articles in the surgical literature were faulted for committing type 2 error, or concluding no difference when the study was "underpowered". However, it is unknown if the current power standard of 0.8 is reasonable in surgical science.

Full text links ELSEVIER FULL-TEXT ARTICLE

Save items

Similar articles

Review Interventions to Prevent Falls in Community-L Agency for Healthcare Research...]

Review Is There Truly "No Significant Difference"? Underl J Bone Joint Surg Am. 2015]

Review Randomized controlled trials and neurosurgery: the ideal fit or : [J Neurosurg. 2016]

Review Low-Dose Aspirin for the Prevention of Morbidity anc Agency for Healthcare Research...]

Further reading



Sam Westwood

@westwoodsam1

Following



I am embarking on my own [#PaperPerDayChallenge](#) where I read at least one paper, well, per day for a whole year. To kick start, nature.com/articles/43573... inspired by [@ukrepro](#) Reproducibility Workshop [@CumberlandLodge](#) and a talk by [@MarcusMunafo](#)



Scientists behaving badly

In a questionnaire-based survey of US biomedical researchers, respondents admitted to a range of dubious practices. Transgressions included failing to present data nature.com

Can I see your data and code?

1989

* Corresponding author.

1999

* Corresponding author. Mailing address: Institute of Human Virology, 725 West Lombard St., Rm. N649, University of Maryland, Baltimore, MD 21201. Phone: (410) 706-4680. Fax: (410) 706-4694. E-mail: devico@umbi.umd.edu.

2009

* Corresponding author. Mailing address: Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 201 Althouse Laboratory, University Park, PA 16802. Phone: (814) 863-8705. Fax: (814) 865-7927. E-mail: cec9@psu.edu.

Also 2019



The screenshot shows the homepage of the Journal of Virology. It features the journal's logo, "AMERICAN SOCIETY FOR MICROBIOLOGY", and the title "Journal of Virology". There is a search bar and an "Advanced Search" link. A horizontal menu includes "Home", "Articles", "For Authors", "About the Journal", and "Subscribe". Below the menu, a link to "Genetic Diversity and Evolution | Spotlight" is visible.

Single-Cell Virus Sequencing of Influenza Infections That Trigger Innate Immunity

Finally, we process the annotated cell-gene matrix in R to generate the plots shown in this paper. This analysis utilized a variety of R and Bioconductor ([90](#)) packages, including Monocle ([91](#), [92](#)) and ggplot2. A Jupyter notebook that performs these analyses is at

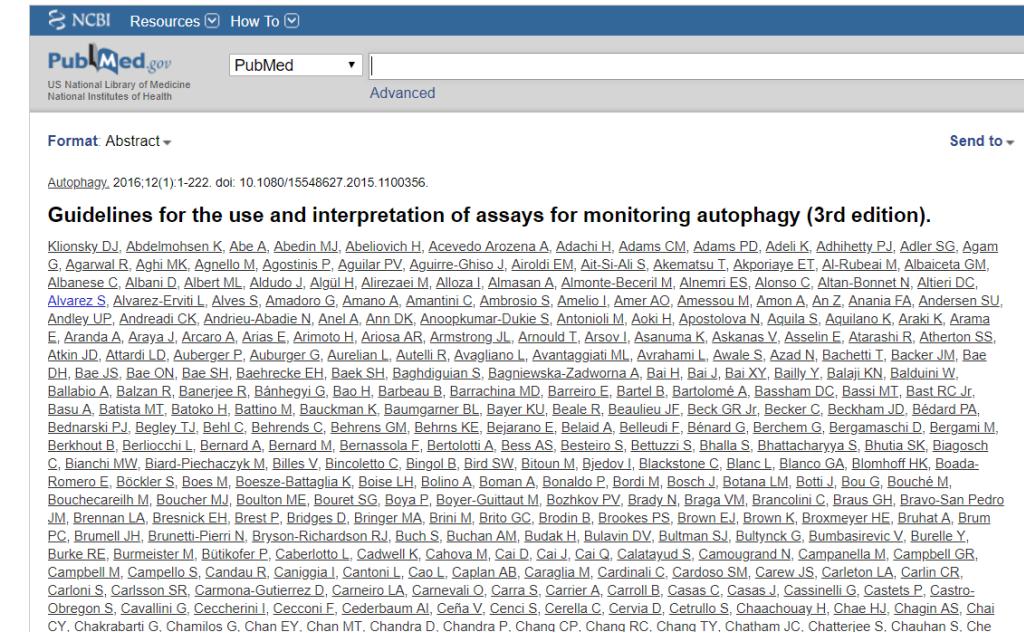
https://github.com/jbloomlab/IFNsorted_flu_single_cell/blob/master/monocle_analysis.ipynb,

2019



Dr Mark Burnley
@DrMarkBurnley

"I'm the 38th author..."
"Wow, that sucks."



The screenshot shows a PubMed search results page for the article "Autophagy, 2016;12(1):1-222. doi: 10.1080/15548627.2015.1100356." The search bar contains "Autophagy". The results list the article title and authors. The page includes a navigation bar with links for NCBI Resources, How To, PubMed, Advanced, Format (Abstract), and Send to.

Guidelines for the use and interpretation of assays for monitoring autophagy (3rd edition).

Klionsky DJ, Abdelmohsen K, Abe A, Abedin MJ, Abieliovich H, Acevedo Arozena A, Adachi H, Adams CM, Adams PD, Adeli K, Adhiketty PJ, Adler SG, Agam G, Agarwal R, Agha MK, Agnello M, Agostinis P, Aguilar PV, Aguirre-Ghiso J, Airoldi EM, Ait-Si-Ali S, Akematsu T, Akporiaye ET, Al-Rubeai M, Albacete GM, Albanese C, Albani D, Albert ML, Aldudo J, Alguí H, Alirezai M, Alloza I, Almasan A, Almonte-Becerril M, Almenr E, Alonso C, Altan-Bonnet N, Althier DC, Alvarez S, Alvarez-Erviti L, Alves S, Amadoro G, Amano A, Amanti C, Ambrosio S, Amelio I, Amer AO, Amessou M, Amon A, An Z, Ananias FA, Andersen SU, Andley UP, Andreadi CK, Andreu-Abadie N, Anel A, Ann DK, Anopukumar-Dukie S, Antonioli M, Aoki H, Apostolova N, Aquila S, Aquilano K, Araki K, Arama E, Aranda A, Araya J, Arcaro A, Arias E, Arimoto H, Ariosa AR, Armstrong JL, Arnould T, Arsov I, Asanuma K, Askarans V, Asselin E, Atarashi R, Atherton SS, Atkin JD, Attardi LD, Auburger P, Auburger G, Aurelian L, Autelli R, Avagliano L, Avantaggiati ML, Avrahami L, Awale S, Azad N, Bachetti T, Backer JM, Bae DH, Bae JS, Bae ON, Bae SH, Baehrecke EH, Baek SH, Baghdiguian S, Baghowska-Zadworna A, Bai H, Bai J, Bai XY, Bally Y, Balaji KN, Balduini W, Ballabio A, Balzan R, Banerjee R, Bánhegyi G, Bao H, Barbeau B, Barrachina MD, Barreiro E, Bartel B, Bartolomé A, Bassham DC, Bassi MT, Bast RC Jr, Basu A, Batista MT, Batoko H, Battino M, Bauckman K, Baumgartner BL, Bayer KU, Beale R, Beaulieu JF, Beck GR Jr, Becker C, Beckham JD, Bédard PA, Bednarski PJ, Begley TJ, Behl C, Behrends GM, Behrens KE, Bejarano E, Belaid A, Belleudi F, Bénard G, Berchem G, Bergamaschi D, Bergami M, Berkhouit B, Berliocchi L, Bernard A, Bernard M, Bernassola E, Bertolotti A, Bess AS, Besteiro S, Bettuzzi S, Bhalla S, Bhattacharyya S, Bhutia SK, Biagusch C, Bianchi MW, Biard-Piechaczyk M, Billes V, Bincoletto C, Bingol B, Bird SW, Bitoun M, Bjedov I, Blackstone C, Blanc L, Blanco GA, Blomhoff HK, Boada-Romero I, Böcker S, Boes M, Boesze-Battaglia K, Boisse L, Bolino A, Roman A, Bonaldo P, Bordi M, Bosch J, Botana LM, Bott J, Bou G, Bouché M, Bouchebareilh M, Boucher MJ, Boulton ME, Bourret SG, Boya P, Boyer-Guittaut M, Bozhkov PV, Brady N, Braga VM, Brancolini C, Braus GH, Bravo-San Pedro JM, Brennan LA, Bresnick EH, Brest P, Bridges D, Bringer MA, Brini M, Brito GC, Brodin B, Brookes PS, Brown EJ, Brown K, Broxmeyer HE, Bruhat A, Brum PC, Brumell JH, Brunetti-Pierri N, Bryson-Richardson RJ, Buch S, Buchan AM, Budak H, Bulavin DV, Bultman SJ, Bulyntck G, Bumbasirevic V, Burelle Y, Burke RE, Burmeister M, Butikofer P, Caberlotto L, Cadwell K, Cahoya J, Cai D, Cai J, Cai Q, Calafatayud S, Camougrand N, Campanella M, Campbell GR, Campbell M, Campello S, Candau R, Caniggià J, Cantoni L, Cao L, Caplan AB, Caraglia M, Cardinali C, Cardoso SM, Carew JS, Carlton LA, Carlton CR, Carloni S, Carlsson J, Carmona-Gutiérrez D, Carneiro LA, Carnevali O, Carras S, Carrier A, Carroll B, Casas C, Casas J, Cassinelli G, Castets P, Castro-Obregon S, Cavallini G, Ceccherini I, Ceconi F, Cederbaum AI, Cefá V, Censi S, Cerella C, Cervia D, Cetrullo S, Chaachouay H, Chae HJ, Chagin AS, Choi CY, Chakrabarti G, Chamilos G, Chan EY, Chan MT, Chandra D, Chandra P, Chano RC, Chang TY, Chatham JC, Chatterjee S, Chauhan S, Che

jbloom saved plot of association between co-infection / IFN

1 contributor

12.8 MB

Table of Contents

- [Analyze viral features associated with IFN induction](#)
 - [Setup for analysis](#)
 - [Load / install packages](#)
 - [Notebook-wide variables / functions](#)
 - [Get cell-gene matrices](#)
 - [Specify cell types](#)
 - [Load cell-gene matrix](#)
 - [Count cells and annotate multiplets](#)
 - [Annotate cross-celltype multiplets](#)
 - [Number of cells and multiplet frequency](#)
 - [Plot summarizing cell counts and multiplets](#)
 - [Filter multiplets and low-quality cells](#)
 - [Remove cross-celltype multiplets](#)
 - [Number of cellular and flu mRNAs, bounds for filtering](#)
 - [Plot cellular / flu mRNAs with filters](#)
 - [Filter cells with extreme mRNA amounts](#)
 - [Call infection / gene presence from canine cell thresholds](#)
 - [Constant fraction or number of mRNAs from flu?](#)
 - [Confirm equal mix of flu barcodes in canine cells](#)
 - [Look at segment frequencies](#)
 - [Get human cells for infection-status calling](#)
 - [Compute P-value flu is above background](#)
 - [Call infected cells by amount of total flu](#)
 - [Call gene presence/absence](#)

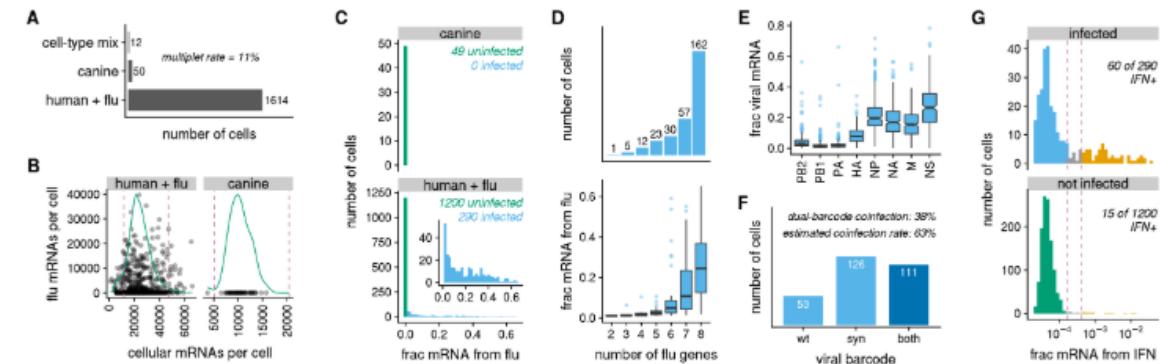
Figures for paper

We have made all of the plots above, and saved some of them to the figures directory already by using the `isfig=TRUE` argument to `saveShowPlot`. However, there are others that we want to assemble into multi-panel figures. We do that here.

First, we assemble a figure that shows the calling of cells, infected cells, and IFN+ cells:

```
In [101]: p_cell_summary <- plot_grid(
  plot_grid(p_cellcounts, p_fiu_vs_cell,
            ncol=1, rel_heights=c(1, 1.5), scale=0.9,
            labels=c("A", "B"), label_size=18, vjust=1),
  plot_grid(p_frac_fiu, labels="C", scale=0.95, label_size=18, vjust=1),
  plot_grid(p_nfui_genes, labels="D", scale=0.95, label_size=18, vjust=1),
  plot_grid(p_fiu_rel_expr, p_cinfect,
            scale=0.95, ncol=1,
            labels=c("E", "F"), label_size=18, vjust=1),
  plot_grid(p_ifn_dist, labels="G", scale=0.95, label_size=18, vjust=1),
  nrow=1, scale=0.95, rel_widths=c(1, 0.7, 0.6, 0.75, 0.7), align="h"
) +
  theme(plot.margin=unit(c(t=0, r=0, b=-0.3, l=0), "in"))

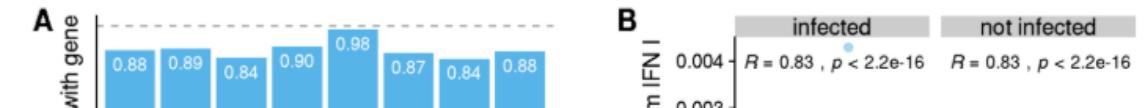
saveShowPlot(p_cell_summary, width=15.5, height=4.9, isfig=TRUE)
```



Now a supplementary figure to the one above with the single-cell transcriptomic data:

```
In [102]: p_cell_summary_supp <- plot_grid(
  p_frac_has_gene,
  p_ifn_genes_corr,
  p_isg_dist,
  p_isg_corr,
  ncol=2,
  scale=0.9,
  rel_heights=c(0.68, 1),
  labels=c("A", "B", "C", "D"), label_size=18, vjust=2, hjust=-1
)

saveShowPlot(p_cell_summary_supp, width=9.5, height=7.5)
```



Data management is a part of your life now!

HRB Health Research Board

Funding Data collections & evidence Publications Success stories News About

Home > Funding > Funding schemes > Before you apply > All grant policies > HRB Policy on Management and Sharing of Research Data

Funding schemes

- All funding schemes
- Before you apply
- How we assess applications
- Tips for writing a grant application
- Useful Links
- All grant policies
- HRB Guidelines on Intellectual Property
- Health Research Charities Ireland
- HRB Policy on alleged misconduct in research grants

HRB Policy on Management and Sharing of Research Data

[HRB_Policy_on_sharing_of_research_data.pdf 233 KB](#)

Note: The HRB will host an engagement event in Q1 of 2020 to discuss this new policy with Host Institutions. Once a date has been confirmed, notification of the event will be sent to Research Offices.

For data gathered and generated in whole or in part from HRB-funded research, the following policy will be adhered to with effect from 1st of January 2020.

NIH Data Management and Sharing Activities Related to Public Access and Open Science

Validation and progress in biomedical research – the cornerstone of developing new prevention strategies, treatments, and cures – is dependent on access to scientific data. Sharing scientific data helps validate research results, enables researchers to combine data types to strengthen analyses, facilitates reuse of hard to generate data or data from limited sources, and accelerates ideas for future research inquiries. Central to sharing scientific data is the recognized need to make data as available as possible while ensuring that the privacy and autonomy of research participants are respected, and that confidential/proprietary data are appropriately protected.

Scientific Data Sharing

> Genomics and Health

> Scientific Data Management



SERVICES SUPPORT

Open Research Data Pilot in H2020

SHARE

Date(s): 08 June 2015

OpenAIRE Public webinar on the "Open Research Data Pilot in H2020", supported by FOSTER project.

Target audience: researchers, project coordinators and research administrator

★★★★★
0.0/5 rating (0 votes)

Twitter



UK Reproducibility Network

@ukrepro

UK Reproducibility Network: a peer-led consortium to investigate factors which contribute to robust research, provide training, and disseminate best practice.



Malcolm Macleod #FBPE

@Macломaclee Follows you

clinical neurologist, stroke trialist, and interested in improving the quality of laboratory research



Open Science MOOC

@OpenScienceMOOC Follows you

A community designed for students and researchers to help make 'Open' the default setting for the future of research. Slack: osmooc.herokuapp.com

⌚ Everywhere



Brian Nosek

@BrianNosek

Executive Director @ Center for Open Science, Professor @ University of Virginia, and co-Founder of Project Implicit



Kate Button

@ButtonKate Follows you

Academic. Psychologist. Cognitive mechanisms of depression & anxiety. Meta-science & scientific rigour. Sporadic Twitterer.



Darren L Dahly

@statsep1 Follows you

Principal Statistician, Epidemiologist, Sr Lecturer | @HRBIreland Clinical Research Facility @CRF_CORK | Cork #Rstats Users Group [meetup.com/Cork-Ireland-R...](https://meetup.com/Cork-Ireland-R-/)



Dorothy Bishop

@deevybee

Professor of developmental neuropsychology. Blog on deevybee.blogspot.com Main focus #devlangdis, see: youtube.com/radld



Elisabeth Bik

@MicrobiomDigest

Science consultant, PhD. Harbers-Bik LLC. Microbiome, research integrity & misconduct. Ex @Stanford. MicrobiomeDigest/Bik's Picks. Dutch/USA. My views.



Retraction Watch

@RetractionWatch

Tracking retractions as a window into the scientific process. Sign up for our daily newsletter: eepurl.com/bNR1Un Tips? team@retractionwatch.com



Jenny Bryan

@JennyBryan

Software engineer @rstudio, humane #rstats, adjunct prof @UBC where I created @STAT545, part of @ropensci

Lets try it out

github.com/bapalmer/lunchtime_sessions

Search or jump to... / Pull requests Issues Marketplace Explore

bapalmer / lunchtime_sessions Watch 2 Star 3 Fork 0

Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security Insights Settings

Short 1 hour introductions to R-related topics such as creating R projects, using GitHub through RStudio and more Edit

Manage topics

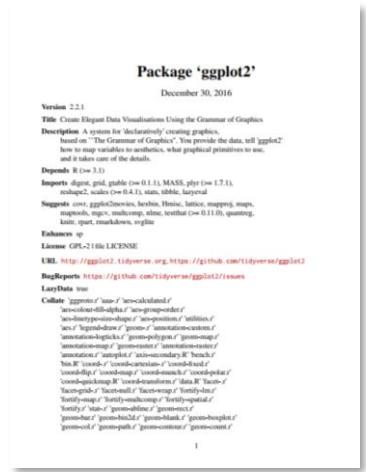
50 commits 1 branch 0 packages 0 releases 1 contributor MIT

Branch: master New pull request Create new file Upload files Find file Clone or download

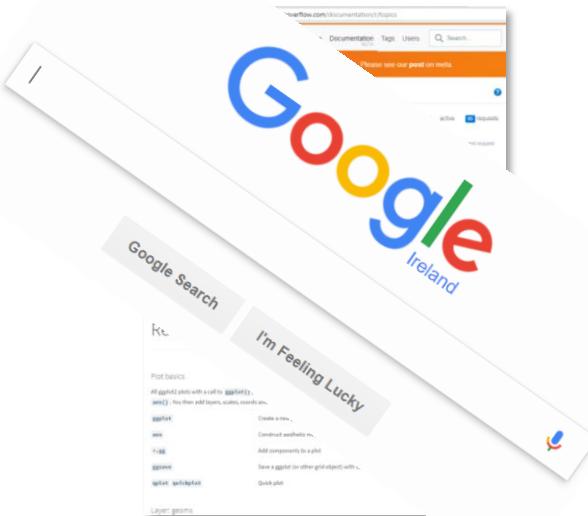
File	Commit	Time
bapalmer Binder link	57af84c	2 minutes ago
Session_1-R_projects	PG6015	Dec 2019 8 minutes ago
Session_2-Reproducible_reports	PG6015	Dec 2019 8 minutes ago
Session_3-Git_and_RStudio	PG6015	Dec 2019 8 minutes ago
Session_4-OS_and_reproducible_research	PG6015	Dec 2019 8 minutes ago

R is for Resources

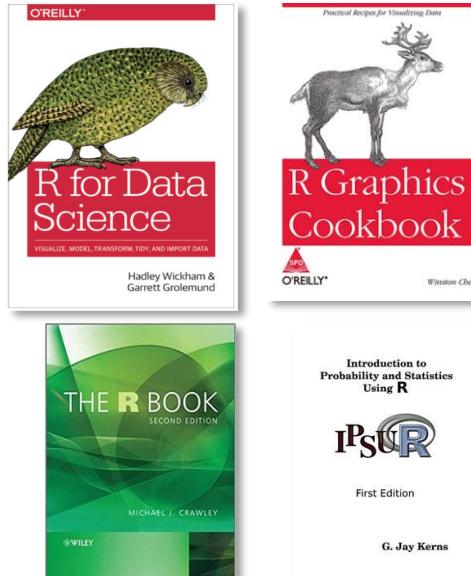
Vignettes



Webpages



eBooks



Cheatsheets



Twitter



Mara Averick

@dataandme

tidyverse @rstudio, hoop head, gnashgb, blatherskite, lesser ½ of @batpigandme

Massachusetts

maraaverick@uidaho.edu



One R Tip a Day

@RLangTip

One tip per day M-F on the R programming language #stats. Brought to you by the R community team at Microsoft.



Hadley Wickham

@hadleywickham

R, data, visualisation.

Houston, TX

hadley.nz



David Robinson

@drob

Data Scientist at @StackOverflow, #stats fan/evangelist

New York, NY

varianceexplained.org



Jenny Bryan

@JennyBryan

Software engineer @rstudio, humane #stats, adjunct prof @UBC where I created @STAT545, part of @ropensci

Vancouver, BC

jennybryan.org



Darren L Dahly

@statsepi Follows you

Principal Statistician, Epidemiologist, Sr Lecturer | @HRBIreland Clinical Research Facility @CRF_CORK | Cork #Rstats Users Group meetup.com/Cork-Ireland-R...
 Cork, Ireland

darrendahly.github.io



Data Scientists IRL

@DataSci_Ireland Follows you

Promoting the Data Science professions in Ireland.

Ireland

facebook.com/DataScientists...



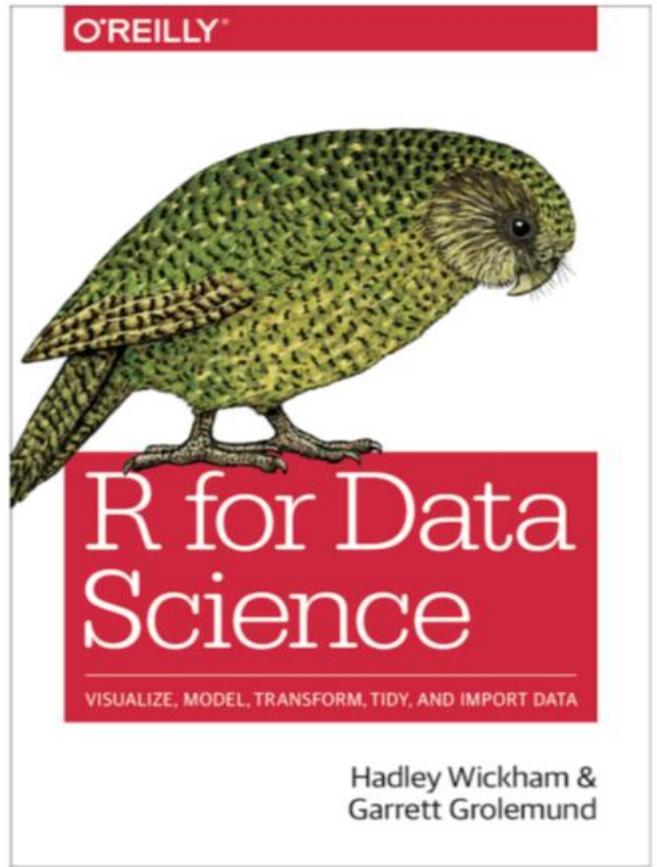
Kara Woo

@kara_woo

Research scientist at @sagebio. Data curation, visualization, #rstats, reproducibility, open science, ballet

karawoo.com

R is for Resources



Hadley Wickham @hadleywickham



Garrett Grolemund @StatGarrett



Course Books



Level 1: Grassroots

Our first-year undergraduate course covers current state of psychological science and what Open Science is as well as its importance. It also aims to make students confident and competent at using RStudio as a tool to achieve good data management skills.

Authors: Emily Nordmann, Heather Woods

Contact: Emily Nordmann

Contributors: Jack Taylor, Shannon McNee



Level 2: Practical

Our second-year undergraduate course covers data skills such as R Markdown, data wrangling with tidyverse, and data visualisation with ggplot2. It also introduces statistical concepts such as permutation tests, NHST, alpha, power, effect size, and sample size. Semester 2 focusses on correlations and the general linear model.

Authors: Phil McAleer, Helena Paterson

Contact: Phil McAleer



Level 3: Statistical Models (Coming Soon)

This third-year undergraduate course teaches students how to specify, estimate, and interpret statistical models corresponding to various study designs, using a General Linear Models approach.

Author: Dale Barr



MSc Conversion

This book contains materials for students on the MSc Conversion in Psychological Studies/Science, a one-year postgraduate degree for students with a non-psychology undergraduate degree. This research methods course covers core data skills that allow you to manipulate and analyse quantitative data.



Emily Nordmann @emilynordmann

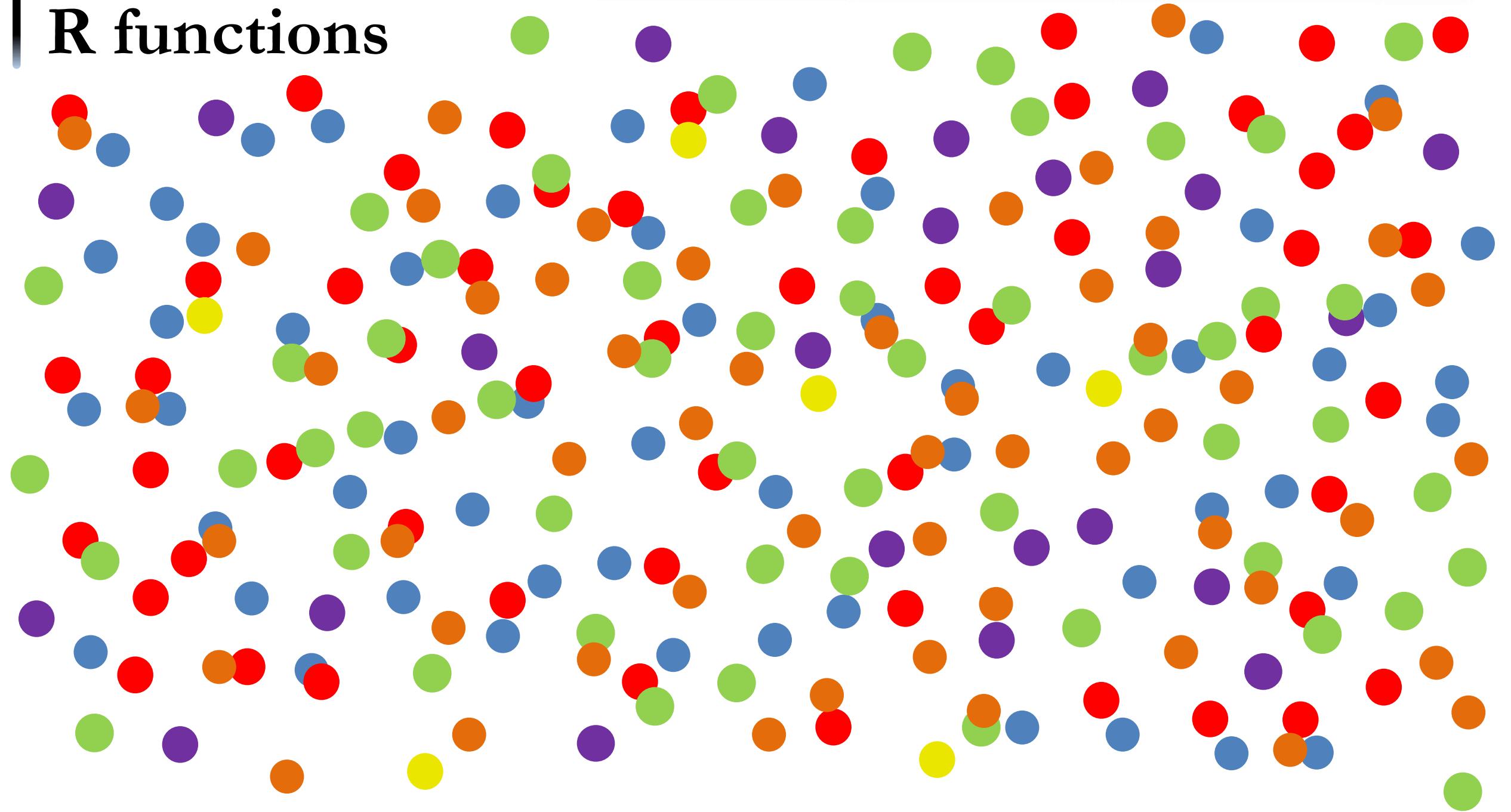


Lisa DeBruine @LisaDeBruine

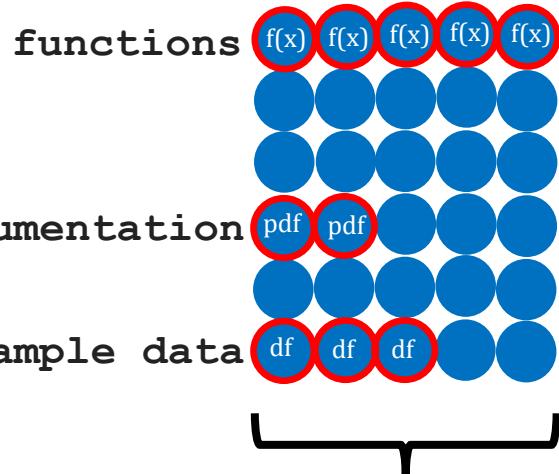


Phil McAleer @McAleerP

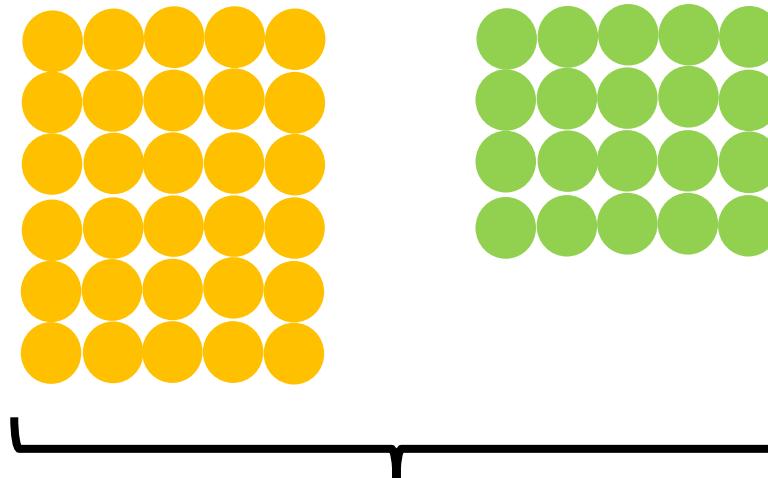
R functions



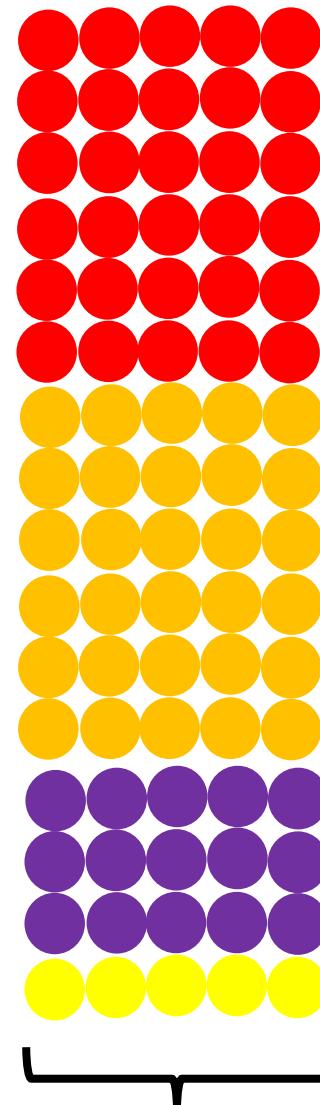
R packages



R comes pre-loaded with ~30 other packages (e.g. base, stats, graphics etc.)



Other packages:
Install once
Update regularly
Load each session



tidyverse

What is the tidyverse?

Tidyverse

Packages Articles Learn Help Contribute

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")
```

- Joined up collection of packages for data analysis
 - Consistent functions
 - Uses (tidy) data
 - Supports end-to-end workflows

R user interface versus RStudio

RGui (64-bit)

File Edit View Misc Packages Windows Help

[Icons: File, Open, Save, Print, Copy, Paste, Find, Stop]

R Console

```
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> print("Hello world!!!")
[1] "Hello world!!!"
> object <- "Hello world!!!"
> object
[1] "Hello world!!!"
> x <- 1:15
> y <- 16:30
> z <- mean(x + y)
> z
[1] 31
> plot(x,y)
> |
```

R Graphics: Device 2 (ACTIVE)

A scatter plot titled 'R Graphics: Device 2 (ACTIVE)'. The x-axis is labeled 'x' and ranges from 2 to 14. The y-axis is labeled 'y' and ranges from 16 to 30. There are 15 data points represented by open circles, forming a clear upward-sloping line.

x	y
2	16
3	18
4	19
5	20
6	21
7	22
8	23
9	24
10	25
11	26
12	27
13	28
14	29
15	30

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Pre-workshop

test_script.R

```
31 gather(sample, expression, G0.05:U0.3) %>%
32 separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE) %>%
33 mutate(nutrient = plyr::revalue(nutrient, nutrient_names)) %>%
34 filter(!is.na(expression), systematic_name != "")
```

Plot the clean data

```
41 cleaned_genes_tb1 %>%
42 filter(BP == "leucine biosynthesis") %>%
43 ggplot(mapping = aes(x = rate, y = expression, color = nutrient)) +
44 geom_point() +
45 geom_smooth(method = "lm", se = FALSE) +
46 facet_wrap(~ name)
```

Code editor

Console Terminal

```
-/R_Users_Workshop/PG_module/course_notes/R-A_Hitchhikers_Guide_to_Reproducible_Research/Pre-workshop/ >
GID = col_character(),
YORF = col_character(),
NAME = col_character()
```

See spec(...) for full column specifications.

```
> cleaned_genes_tb1 %>%
+   filter(BP == "leucine biosynthesis") %>%
+   ggplot(mapping = aes(x = rate, y = expression, color = nutrient)) +
+   geom_point() +
+   geom_smooth(method = "lm", se = FALSE) +
+   facet_wrap(~ name)
>
```

R console

Environment

Name	Type	Length	Size	Value
cleaned_g...	tbl_df	7	11.3 ... 198430 obs. of 7...	
nutrient_...	character	6	984 B Named chr [1:6] "G...	
url	character	1	168 B "http://varianceex...	

Environment



To understand R, remember the following

- Everything that exists is an object
- Everything that happens is a function

Creating objects

For most of us, R is simply the creation of and manipulation of objects:

```
new_object <- c(1, 2, 3)
```

- the objects are then fed into functions to create amazing new objects

```
amazing_new_object <- function(new_object)
```

Broadly speaking the following is true in R:

- information

```
> data_frame <- function(information)
> plot          <- function(data_frame)
> model         <- function(data_frame)
```

Give it a go

An example scripted, reproducible workflow

Link to packaged data, code and reports



launch binder