Open Science and reproducible research

Thursday 12th December 2019

Brendan Palmer,

Clinical Research Facility - Cork &

School of Public Health

@B_A_Palmer

Can we believe what we see in the literature?

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?

🔓 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

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.

somee cards
user card

7:55 AM - 19 Mar 2019

p-values do not define a study

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

Jeffrey T. Leek & Roger D. Peng
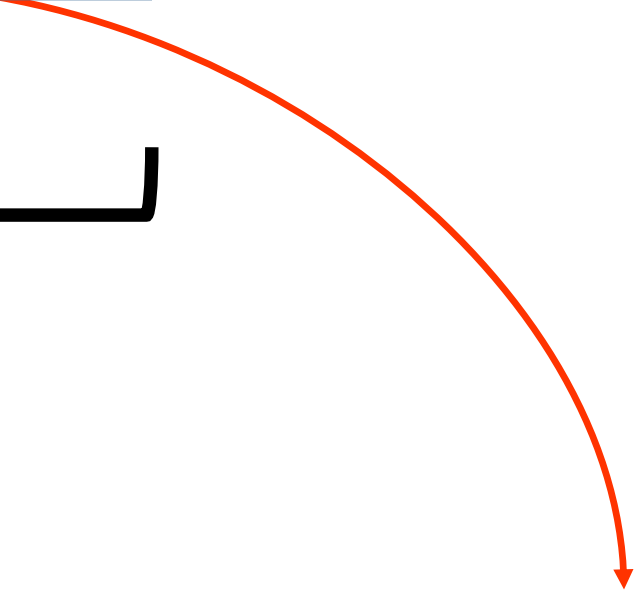
28 April 2015
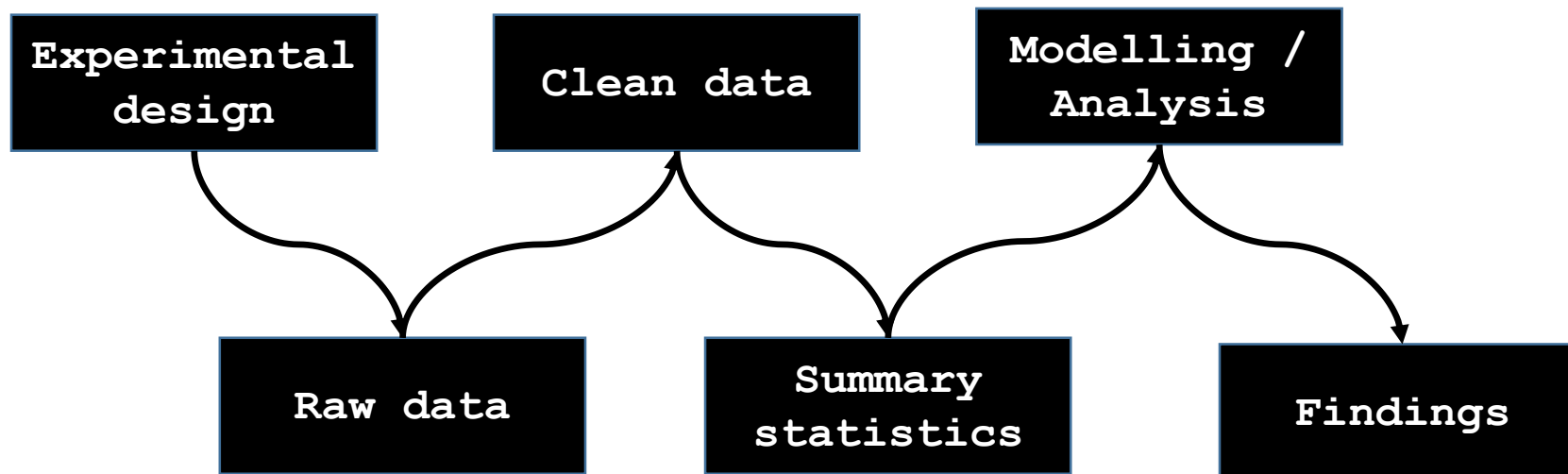
Experimental design → **Raw data**

Clean data ← **Raw data**

Clean data → **Summary statistics**

Modelling / Analysis ← **Summary statistics**

Modelling / Analysis → **Findings**

Experimental design → Raw data

Clean data → Summary statistics

Modelling / Analysis → Findings

Little debate

Extreme
scrutiny

p-value

Ig Noble research - 2012 Neuroscience winner

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. Baird [2], Michael B. Miller [1] and George L. Wolford [3]

[1]Department of Psychology, University of California at Santa Barbara, Santa Barbara, CA 93106
[2]Department of Psychology, Blodgett Hall, Vassar College, Poughkeepsie, NY 12604
[3]Department of Psychological and Brain Sciences, Moore Hall, Dartmouth College, Hanover, NH 03755
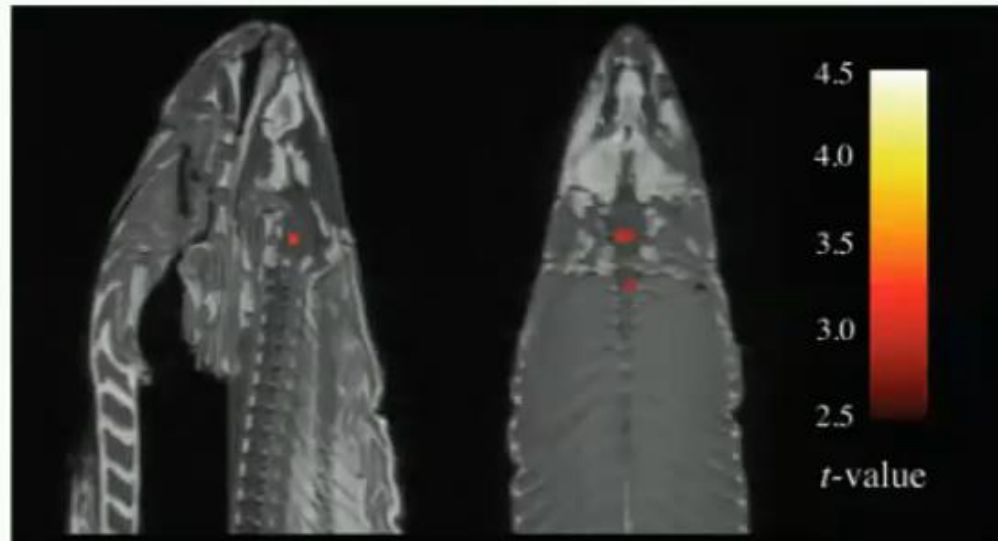
One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon measured approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning. It is not known if the salmon was male or female, but given the post-mortem state of the subject this was not thought 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.

What is your research hypothesis?

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

Daniël Lakens[1] and Alexander J. Etz[2]

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

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
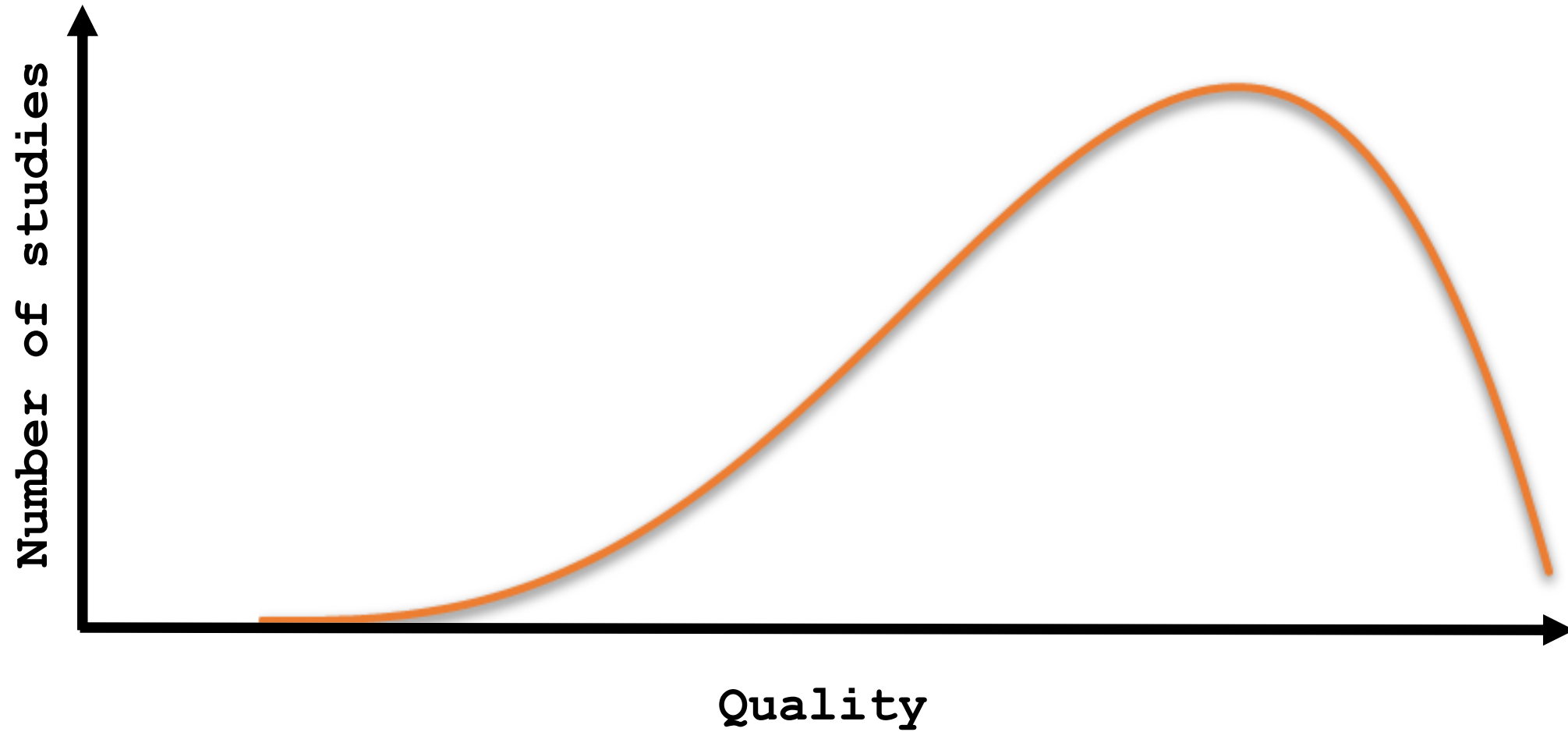
Don't do what Donny Dont does!

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

James Heathers,
Northwestern University

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"
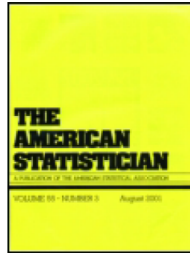
The winds of change

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

Ronald L. Wasserstein & Nicole A. Lazar

ESSAY

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

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]
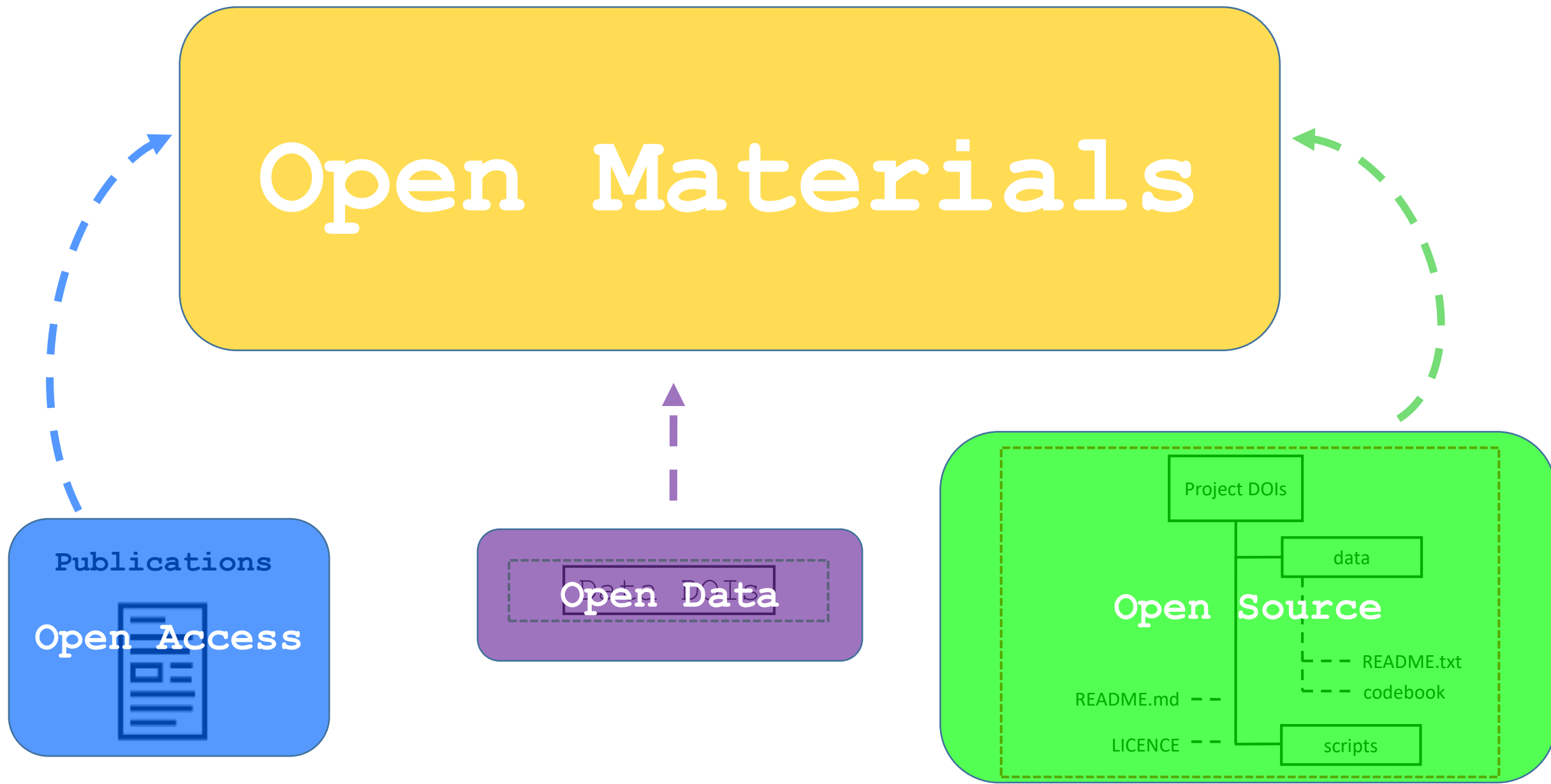
Can we shift the distribution?
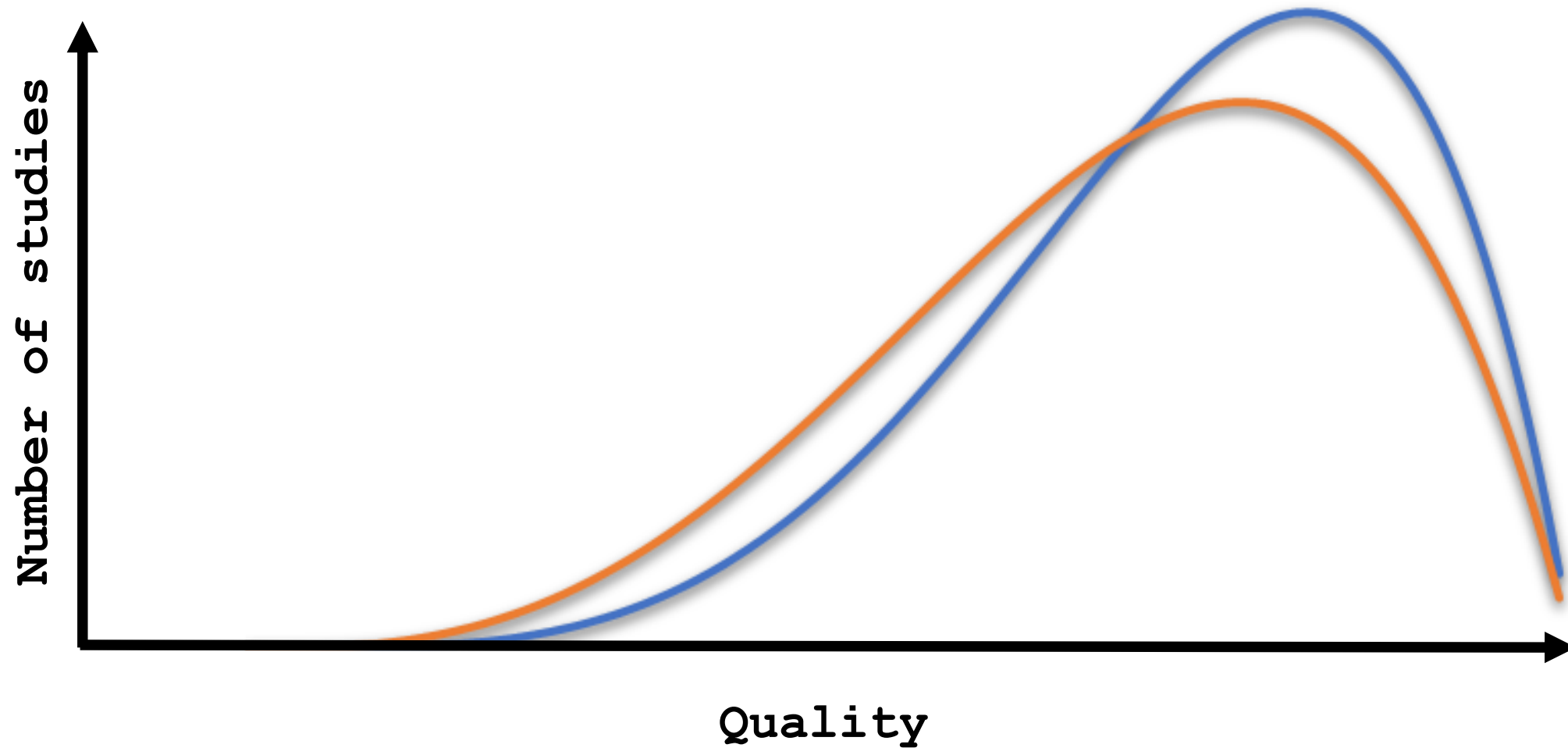
The 'Closed Science Picture'

Open Materials

Publications

Data DOIs

OSF / GitHub Project DOIs

data

README.txt

codebook

README.md

LICENCE

scripts
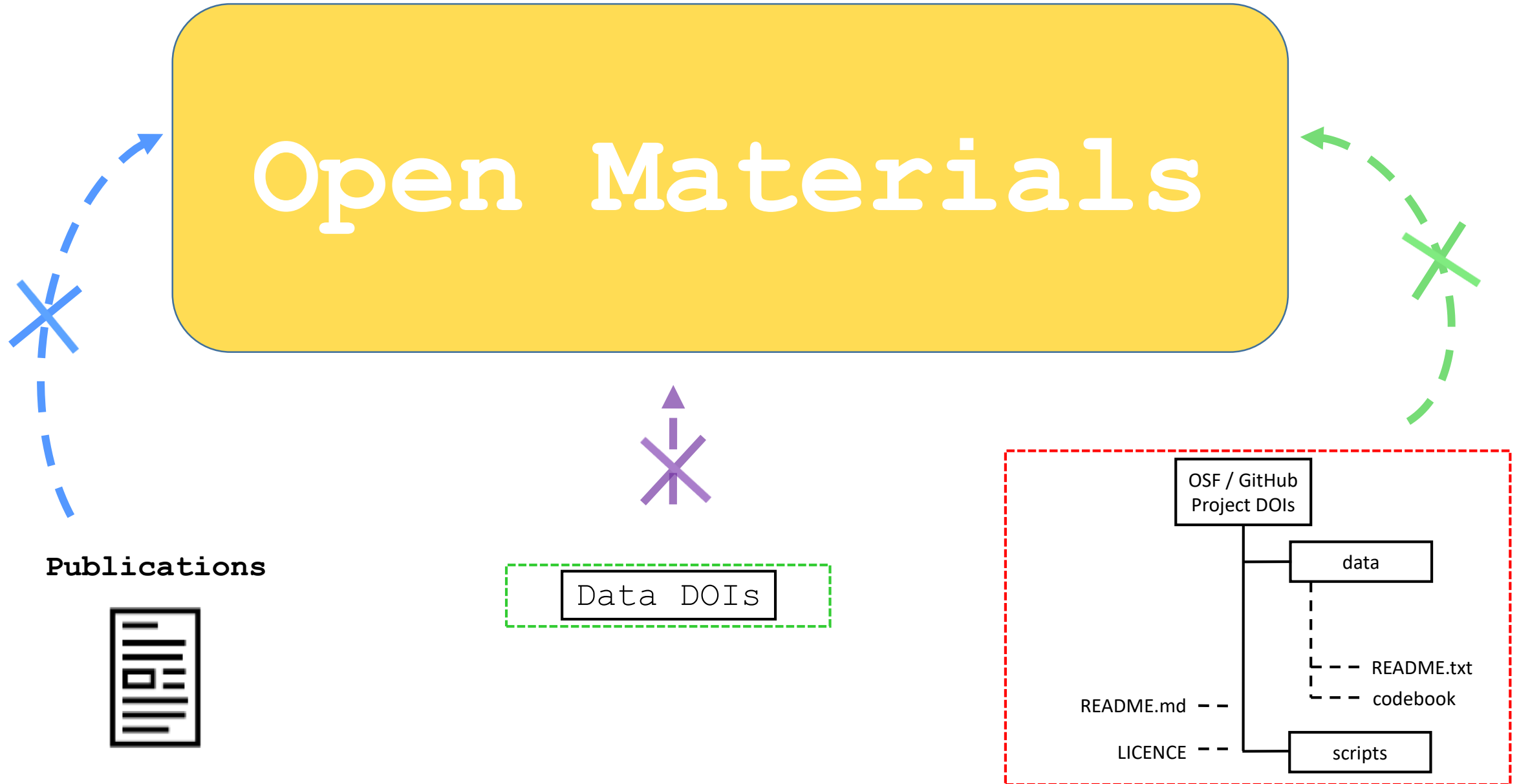
The 'Open Science Picture'

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

AMERICAN SOCIETY FOR MICROBIOLOGY

Journal of Virology

search

Advanced Search

Home   Articles   For Authors   About the Journal   Subscribe

Genetic Diversity and Evolution | Spotlight

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."

NCBI   Resources ⊙   How To ⊙

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Format: Abstract ▾

Send to ▾

Autophagy. 2016;12(1):1-222. doi: 10.1080/15548627.2015.1100356.

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

Klionsky DJ, Abdelmohsen K, Abe A, Abedin MJ, Abeliovich H, Acevedo Arozena A, Adachi H, Adams CM, Adams PD, Adeli K, Adhihetty PJ, Adler SG, Agam G, Agarwal R, Aghi MK, Agnello M, Agostinis P, Aguilar PV, Aguirre-Ghiso J, Airoldi EM, Ait-Si-Ali S, Akematsu T, Akporiaye ET, Al-Rubeai M, Albaiceta GM, Albanese C, Albani D, Albert ML, Aldudo J, Algül H, Alirezaei M, Alloza I, Almasan A, Almonte-Beceril M, Alnemri ES, Alonso C, Altan-Bonnet N, Altieri DC, Alvarez S, Alvarez-Erviti L, Alves S, Amadoro G, Amano A, Amantini C, Ambrosio S, Amelio I, Amer AO, Amessou M, Amon A, An Z, Anania FA, Andersen SU, Andley UP, Andreadi CK, Andrieu-Abadie N, Anel A, Ann DK, Anoopkumar-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, Askanas V, Asselin E, Atarashi R, Atherton SS, Atkin JD, Attardi LD, Auberger 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, Bagniewska-Zadworna A, Bai H, Bai J, Bai XY, Bailly 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, Baumgarner BL, Bayer KU, Beale R, Beaulieu JF, Beck GR Jr, Becker C, Beckham JD, Bédard PA, Bednarski PJ, Begley TJ, Behl C, Behrends C, Behrens GM, Behrns KE, Bejarano E, Belaid A, Belleudi F, Bénard G, Berchem G, Bergamaschi D, Bergami M, Berkhout B, Berliocchi L, Bernard A, Bernard M, Bernassola F, Bertolotti A, Bess AS, Besteiro S, Bettuzzi S, Bhalla S, Bhattacharyya S, Bhutia SK, Biagosch 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 E, Böckler S, Boes M, Boesze-Battaglia K, Boise LH, Bolino A, Boman A, Bonaldo P, Bordi M, Bosch J, Botana LM, Botti J, Bou G, Bouché M, Bouchecareilh M, Boucher MJ, Boulton ME, Bouret 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, Bultynck G, Bumbasirevic V, Burelle Y, Burke RE, Burmeister M, Bütikofer P, Caberlotto L, Cadwell K, Cahova M, Cai D, Cai J, Cai Q, Calatayud S, Camougrand N, Campanella M, Campbell GR, Campbell M, Campello S, Candau R, Caniggia I, Cantoni L, Cao L, Caplan AB, Caraglia M, Cardinali C, Cardoso SM, Carew JS, Carleton LA, Carlin CR, Carloni S, Carlsson SR, Carmona-Gutierrez D, Carneiro LA, Carnevali O, Carra S, Carrier A, Carroll B, Casas C, Casas J, Cassinelli G, Castets P, Castro-Obregon S, Cavallini G, Ceccherini I, Cecconi F, Cederbaum AI, Ceña V, Cenci S, Cerella C, Cervia D, Cetrullo S, Chaachouay H, Chae HJ, Chagin AS, Chai CY, Chakrabarti G, Chamilos G, Chan EY, Chan MT, Chandra D, Chandra P, Chang CP, Chang RC, Chang TY, Chatham JC, Chatterjee S, Chauhan S, Che

<> Code   ⊘ Issues 0   ⑆ Pull requests 0   ▥ Projects 0   ▦ Wiki   ⛉ Security   � Insights

Branch: master ▾   **IFNsorted_flu_single_cell** / monocle_analysis.ipynb

jbloom saved plot of association between co-infection / IFN

1 contributor

12.8 MB

Table of Contents

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_flu_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_flu, labels="C", scale=0.95, label_size=18, vjust=1),
    plot_grid(p_nflu_genes, labels="D", scale=0.95, label_size=18, vjust=1),
    plot_grid(p_flu_rel_expr, p_coinfect,
            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)
```
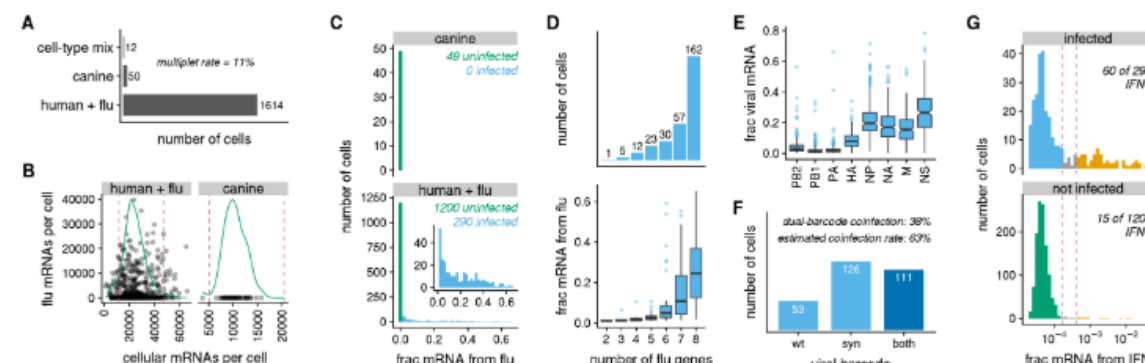

Now a supplementary figure to the one above with the single-cell transcriptomic data:

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)
```
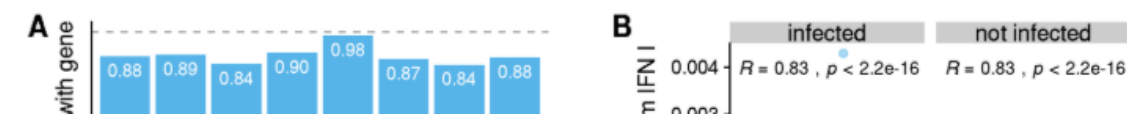
FAIR is a part of our life now!

Data Guidelines

1. Background
   1.1 Open Data Policy
   1.2 Fair Data Principles

2. Share Your Data in 3 Steps
   2.1 Prepare Your Data for Sharing
   2.2 Select a Repository
   2.3 Add a Data Availability Statement to Your Manuscript
   2.4 Linking your datasets to your article

Some types of data benefit from visualization within the article. Wellcome Open Research welcomes the submission of manuscripts featuring Plot.ly interactive figures and Code Ocean compute capsules. For further detail, please contact us.

wellcome

Research Data Management

Good data governance and stewardship are key components of good research practice. In this regard, Science Foundation Ireland supports that research data should be Findable, Accessible, Interoperable and Reusable (FAIR)*. Appropriate data management and data sharing are fundamental to all stages of the research process and support high quality, reproducible research. As such, access to research data arising in whole or in part from SFI funding should be as open as possible.

Science Foundation Ireland **SFI** For what's next

HRB Health Research Board

FAIR Data Management
Describe the approach to data management that will be taken during and after the project, including who will be responsible for data management and data stewardship. The word limit is **500 words**.

UCC University College Cork, Ireland / Coláiste na hOllscoile Corcaigh

Social Research Ethics Committee (SREC)
ETHICS APPROVAL FORM
✉ srec@ucc.ie
https://www.ucc.ie/en/research/about/ethics/

[4] Data management should follow the FAIR guiding principles (Findability, Accessibility, Interoperability & Reusability). See, for example, Wilkinson, M. D. et al. (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship. Full text: http://www.nature.com/articles/sdata201618. It is required that all staff and student researchers store those data which are required to replicate research findings, and the information required to enable re-use of data. Details of the UCC policy on research data storage can be found in section 8 of the Code of Research Conduct (2016): https://www.ucc.ie/en/media/research/researchatucc/documents/UCCCodeofResearchConduct.pdf. SREC advises against storing research data on non UCC approved cloud-based storage services. Physical data must be stored in a locked cabinet and you must specify who has permission to access this data.

NIH National Institutes of Health
Turning Discovery Into Health

A set of Digital Object Compliance principles that describes the properties of digital objects that enables them to be findable, accessible, interoperable and reproducible (FAIR).

Beware of default settings

Genome Biology

COMMENT   **Open Access**

CrossMark

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

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

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and .xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene sym-

- Why did it take us until 2016 to discover this?

2004

But it doesn't end there

Date and time expressed according
to ISO 8601 [refresh]

Date	2019-10-15
Date and time in UTC	2019-10-15T19:49:52+00:00
	2019-10-15T19:49:52Z
	20191015T194952Z
Week	2019-W42
Date with week number	2019-W42-2
Date without year	--10-15[1]
Ordinal date	2019-288

- YYYY-MM-DD or YYYYMMDD

- Type this into Excel

	A	B
1	2019-10-15	
2		

- And hit return

	A	B
1	15/10/2019	
2		

- DD/MM/YYYY

Beware of default settings

The Reinhart-Rogoff error – or how not to Excel at economics

April 22, 2013 9.40pm BST

Data and computer code should be made publicly available at an early stage – or else ... esarastudillo

Last week we learned a famous [2010 academic paper](#), relied on by political big-hitters to bolster arguments for austerity cuts, contained significant errors; and that those errors came down to misuse of an Excel spreadsheet.

Sadly, these are not the first mistakes of this size and nature when handling data. So what on Earth went wrong, and can we fix it?

Harvard's [Carmen Reinhart](#) and [Kenneth Rogoff](#) are two of the most respected and influential academic economists active today.

Email

Twitter   88

Facebook   453

LinkedIn

Print

What can you do?

Casey Greene
@GreeneScientist

Follow

Reproducibility is important because the you of 3 months ago is terrible at answering email! - @tracykteal at #2016dssummit

1:17 PM - 26 Oct 2016 from Manhattan, NY

11 Retweets 25 Likes

1   11   25

Get organised

- Help your future-self

Define a generic project structure
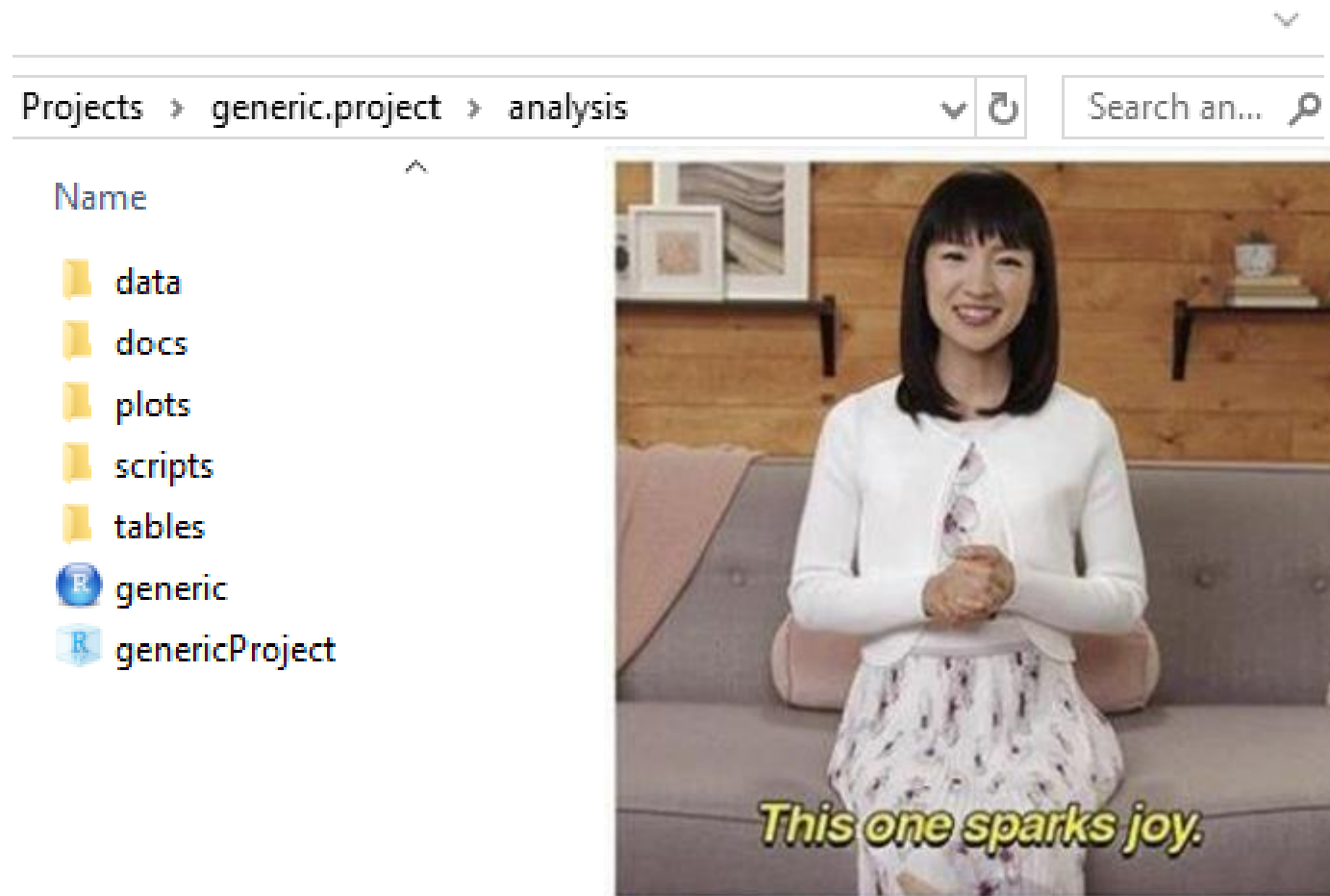
Give your files and folders informative names

- 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_Unique_Clonal_AA

Yes

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

Name

- ⓡ 01_clean_data
- ⓡ 02_plots
- ⓡ 03_tables
- ⓡ 04_stats_analysis
- ⓡ 05_post_hoc_stats
- ⓡ functions
- ⓡ randomization
- ⓡ tables

Putting the pieces together

A: Define a project structure

B: Set a naming convention

C: Use scripted workflows

D: Digital notebooks

Reproducible research

E: Version control

F: Data packaging

Coming to a UCC near you in 2020

Brendan Palmer
@B_A_Palmer

🚨 BIG NEWS: We've just received approval for a new postgraduate module where we'll teach reproducible scripted workflows through #rstats!

Thanks to all who helped get it to this point, but special mention to @statsepi who set the ball rolling 🏝️

@UCC > Modules > PG6030 🥳

Module Content: The module will introduce fundamental concepts of reproducible research alongside hands-on training in the R programming language. Students will be instructed on data collation, curation and management techniques that will serve as a foundation towards downstream visualisation, analysis and reporting via scripted, reproducible workflows.

- Email b.palmer@ucc.ie to register your interest
- First class likely to run in April 2020

Questions…

Richard Feynman
@ProfFeynman

There is no learning without having to pose a question.
And a question requires doubt.