Computational reproducibility and the struggle for reliable science

Mark Ziemann PhD, GCHELT
mark.ziemann@burnet.edu.au

LES Graduate Researcher Conference
2025-08-14

Burnet
reach for the many

AT BURNET INSTITUTE, WE PROUDLY ACKNOWLEDGE THE BOON WURRUNG PEOPLE OF THE KULIN NATIONS AS THE TRADITIONAL CUSTODIANS OF THE LAND ON WHICH OUR OFFICE IS LOCATED AND RECOGNISE THEIR CONTINUING CONNECTION TO LAND, WATERS AND COMMUNITY. WE ACKNOWLEDGE ABORIGINAL AND TORRES STRAIT ISLANDER PEOPLES AS AUSTRALIA’S FIRST PEOPLES AND ACKNOWLEDGE THAT SOVEREIGNTY WAS NEVER CEDED. WE PAY OUR RESPECT TO ELDERS PAST AND PRESENT, AND EXTEND THAT RESPECT TO ALL FIRST NATIONS PEOPLE.

Overview

- Defining reliable science
- Scale of the problem
- Forces at play
- The state of reproducibility in bioinformatics
- Case study
- What you can do
- Our work on enrichment analysis

What is reliable research?

Validity	Appropriateness of the tools, processes and data.
Transparency	Methods, raw data, and code are fully shared to enable reproduction.
Reproducibility	The ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.
Methods reproducibility	Sufficient methodological detail is provided to enable experimental replication.
Results reproducibility	Repeating methods yields similar data/results.
Inferential reproducibility	Independent replication yields similar conclusions.
Computational reproducibility	Reanalysis of the original raw data yields similar results.
Reliable	Research is valid and reproducible.

Research quality is crucial for society to successfully navigate crises like social problems, disease outbreaks, environmental challenges and to translate progress in science to new technological advances and improve standard of living

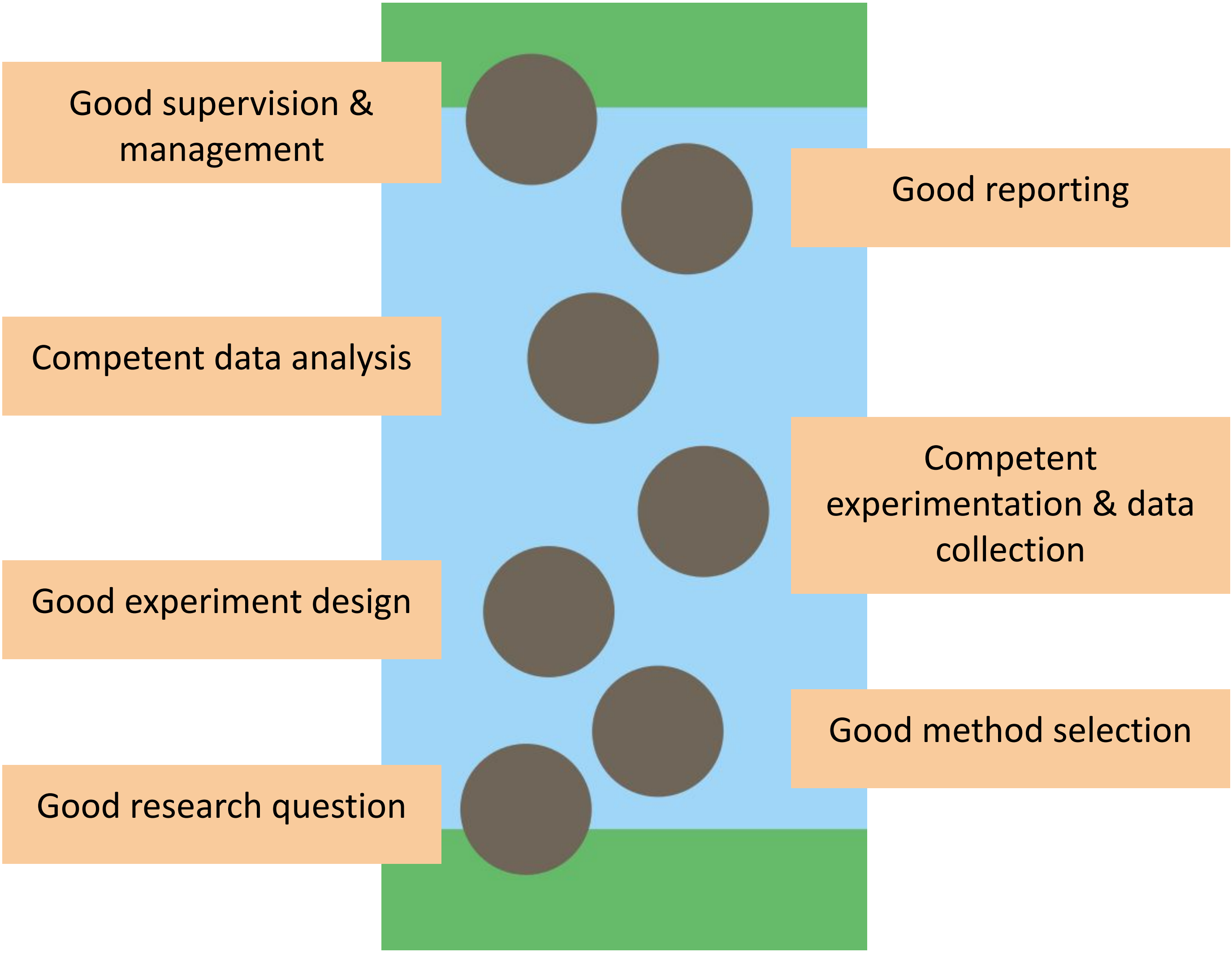
1. Goodman et al, 2016; 2. Gundersen 2021

Towards reliable research

Like most things in life, reliable research requires a series of tasks to be completed to a high degree of quality to be successful.

Any breakdown in quality can lead to problems:

- Wasted resources
- Misleading results
- False/inflated claims
- Irreproducible findings
- Reputational damage



Good supervision & management

Good reporting

Competent data analysis

Competent experimentation & data collection

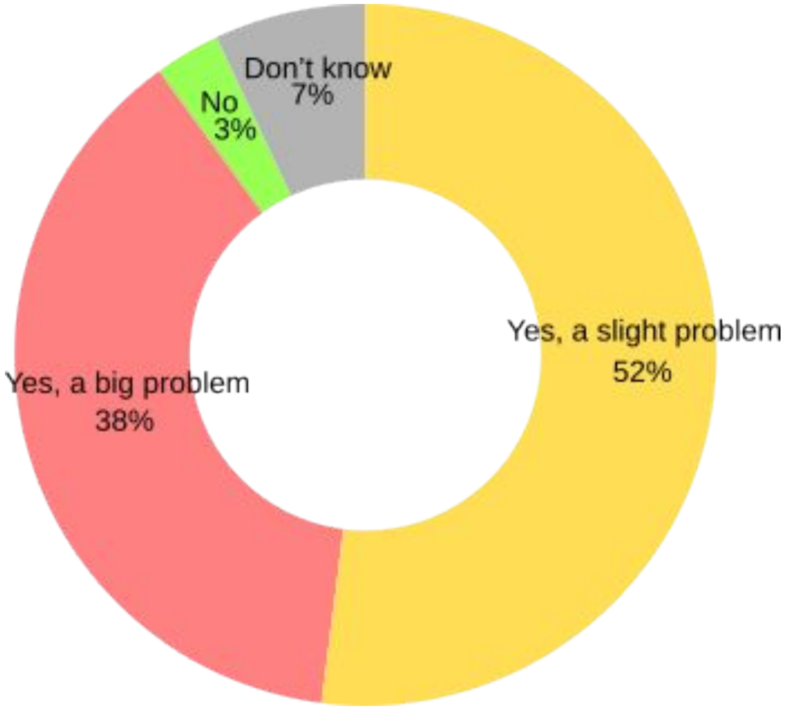
Good experiment design

Good method selection

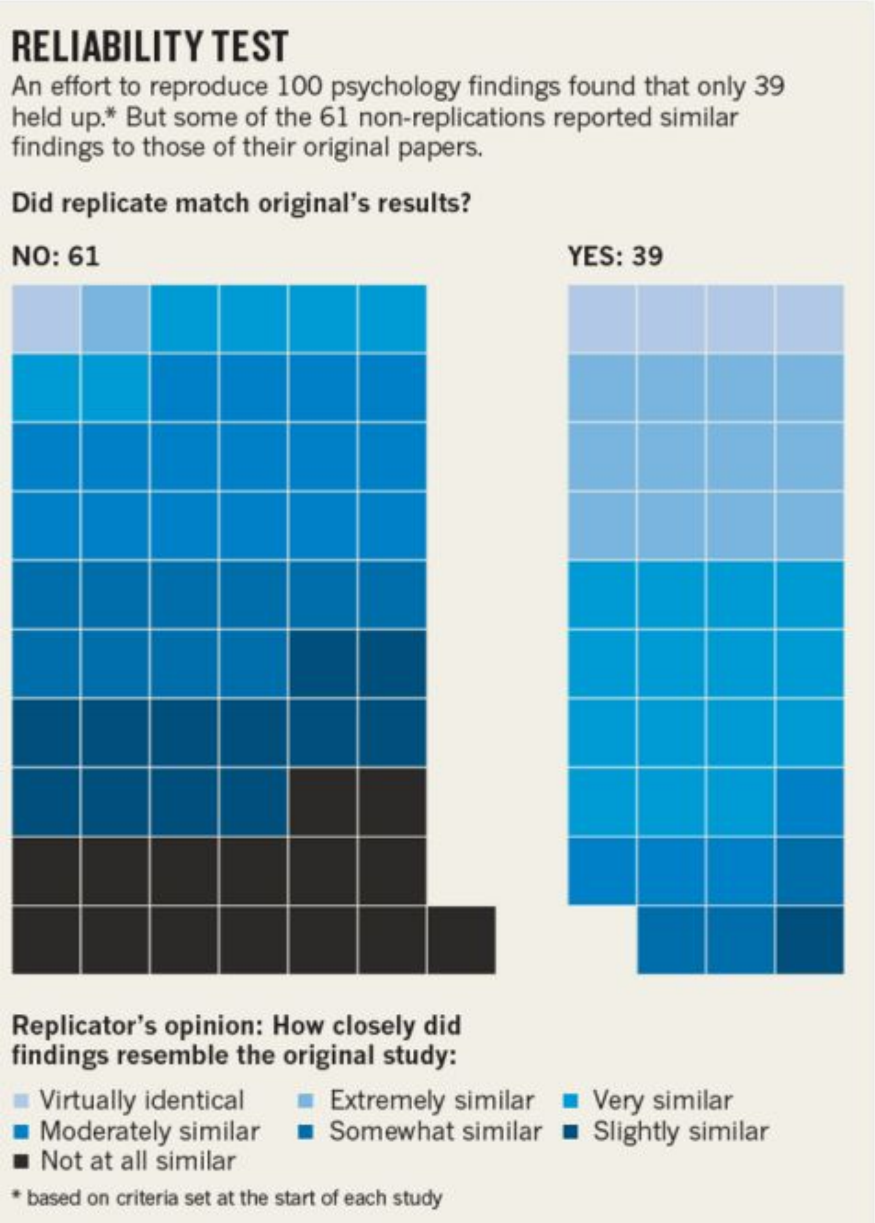
Good research question

Scale of the problem

In 2016, 1,576 researchers were asked whether there is a reproducibility crisis in science



3. Baker 2016



RELIABILITY TEST

An effort to reproduce 100 psychology findings found that only 39 held up.* But some of the 61 non-replications reported similar findings to those of their original papers.

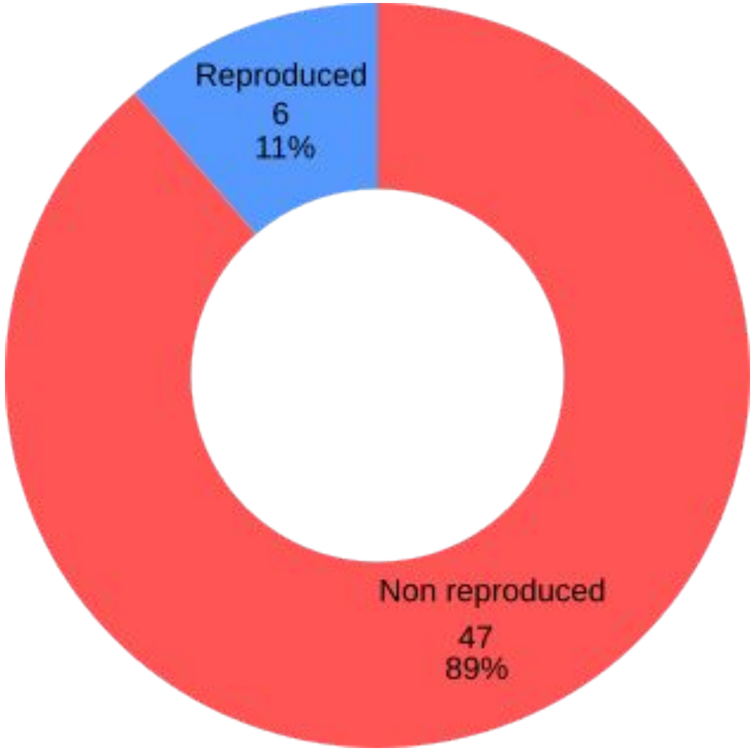
Did replicate match original's results?

Replicator's opinion: How closely did findings resemble the original study:

* based on criteria set at the start of each study

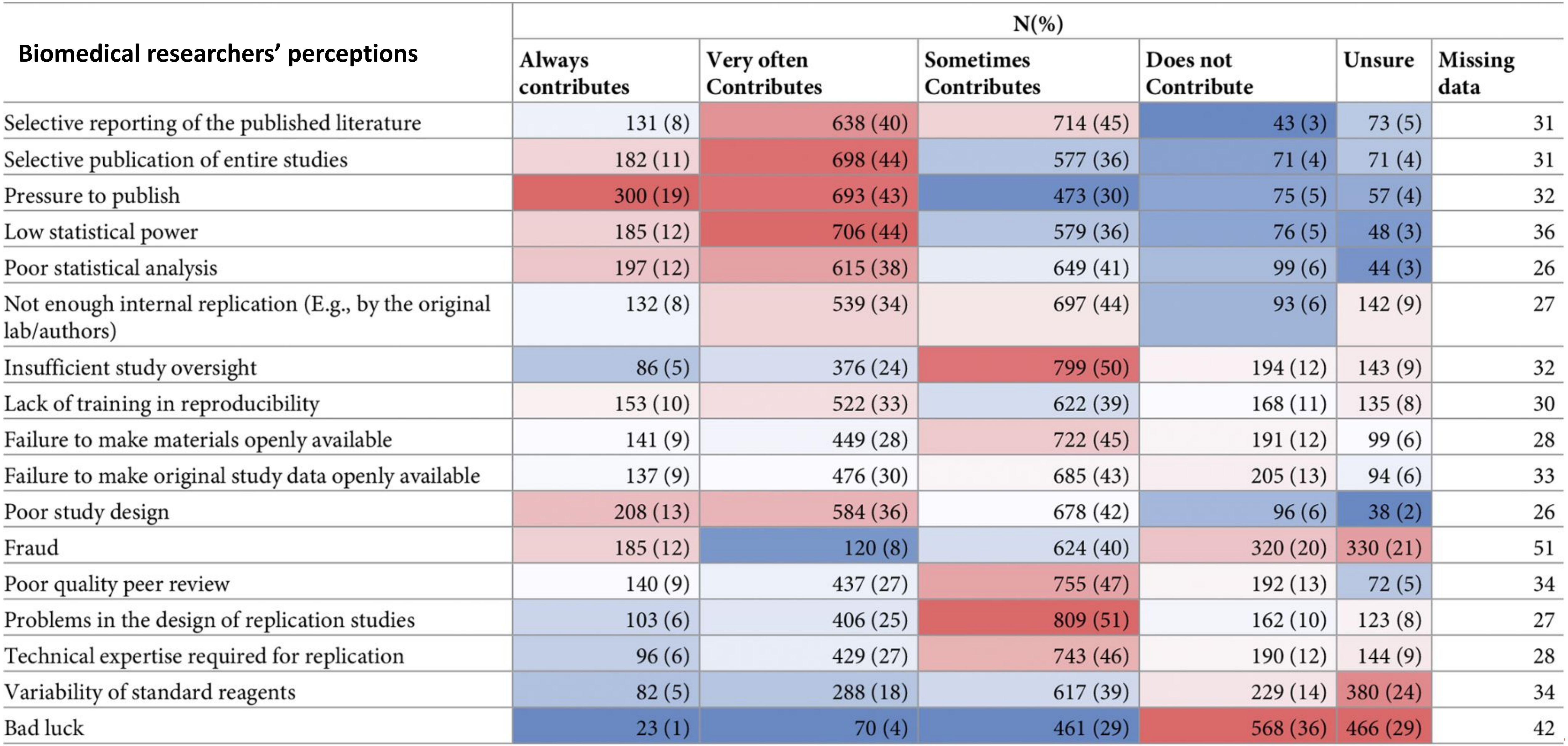
4. Baker 2015

AMGEN preclinical reproducibility survey



5. Begley & Ellis 2012.

The forces at play



Biomedical researchers' perceptions	N(%)					
	Always contributes	Very often Contributes	Sometimes Contributes	Does not Contribute	Unsure	Missing data
Selective reporting of the published literature	131 (8)	638 (40)	714 (45)	43 (3)	73 (5)	31
Selective publication of entire studies	182 (11)	698 (44)	577 (36)	71 (4)	71 (4)	31
Pressure to publish	300 (19)	693 (43)	473 (30)	75 (5)	57 (4)	32
Low statistical power	185 (12)	706 (44)	579 (36)	76 (5)	48 (3)	36
Poor statistical analysis	197 (12)	615 (38)	649 (41)	99 (6)	44 (3)	26
Not enough internal replication (E.g., by the original lab/authors)	132 (8)	539 (34)	697 (44)	93 (6)	142 (9)	27
Insufficient study oversight	86 (5)	376 (24)	799 (50)	194 (12)	143 (9)	32
Lack of training in reproducibility	153 (10)	522 (33)	622 (39)	168 (11)	135 (8)	30
Failure to make materials openly available	141 (9)	449 (28)	722 (45)	191 (12)	99 (6)	28
Failure to make original study data openly available	137 (9)	476 (30)	685 (43)	205 (13)	94 (6)	33
Poor study design	208 (13)	584 (36)	678 (42)	96 (6)	38 (2)	26
Fraud	185 (12)	120 (8)	624 (40)	320 (20)	330 (21)	51
Poor quality peer review	140 (9)	437 (27)	755 (47)	192 (13)	72 (5)	34
Problems in the design of replication studies	103 (6)	406 (25)	809 (51)	162 (10)	123 (8)	27
Technical expertise required for replication	96 (6)	429 (27)	743 (46)	190 (12)	144 (9)	28
Variability of standard reagents	82 (5)	288 (18)	617 (39)	229 (14)	380 (24)	34
Bad luck	23 (1)	70 (4)	461 (29)	568 (36)	466 (29)	42

https://doi.org/10.1371/journal.pbio.3002870.t002

6. Cobey et al, 2024

The struggle for reliable science

Negative forces/outcomes

- Poor training/supervision/culture
- Sloppy methodology/record keeping
- Bibliometric misuse
- Pressure to publish
- Chasing high impact
- Competition against peers
- Lack of resource sharing
- Corporate exploitation
- Collapse of peer review system
- Predatory journals
- Failure of science funding
- Research misconduct

Positive forces/outcomes

- Quality over quantity; Slow science
- Eschew bibliometrics
- Co-operation instead of competition
- Sharing resources like data and code
- Mentoring
- Participation in society-led and non-profit journals
- Preprinting and retaining copyright
- Meta-research*
- Advocacy for best practices*

Individual researchers, research teams, institutions, journals and funding bodies all play a role in promoting quality science

The state of play in bioinformatics

- A 2009 systematic evaluation showing only 2 of 18 articles could be reproduced (11%) [7]
- In 2020 an NIH pilot study tried to replicate 5 bioinformatics projects but couldn't reproduce *any* [8]
- In 2024, a systematic analysis of Jupyter notebooks in biomedical articles showed only 879/22578 notebooks (2.9%) gave similar results [9]

Less than 10% of bioinformatics papers are reproducible, due to lack of data and code sharing, poor documentation and broken code.

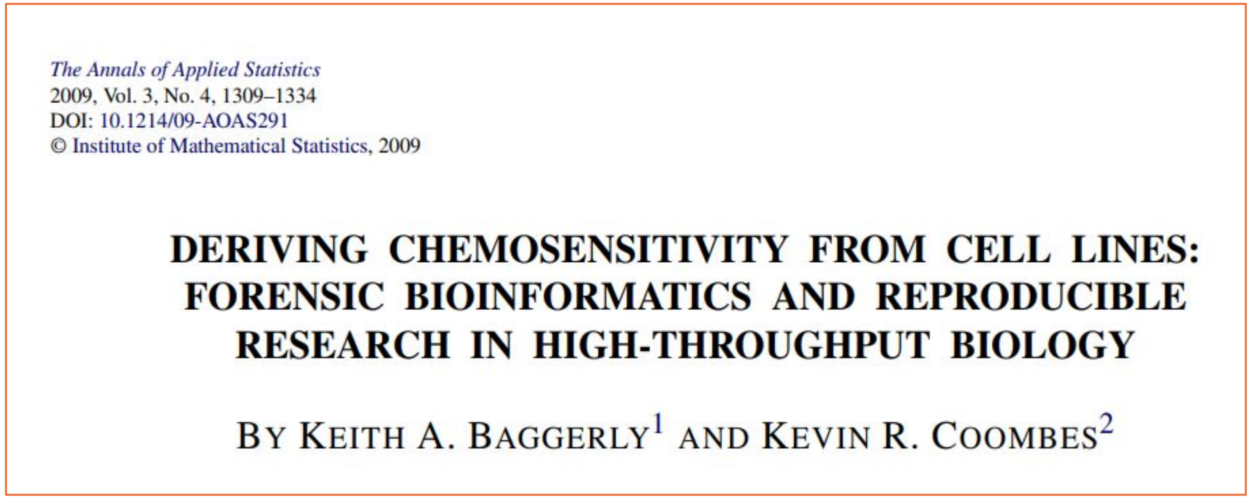
No one is checking

7. Ioannidis et al, 2009 ; 8. Zaringhalam and Federer 2020; 9. Samuel and Mietchen 2024.

Case study

Potti et al (2006) had a number of problems:

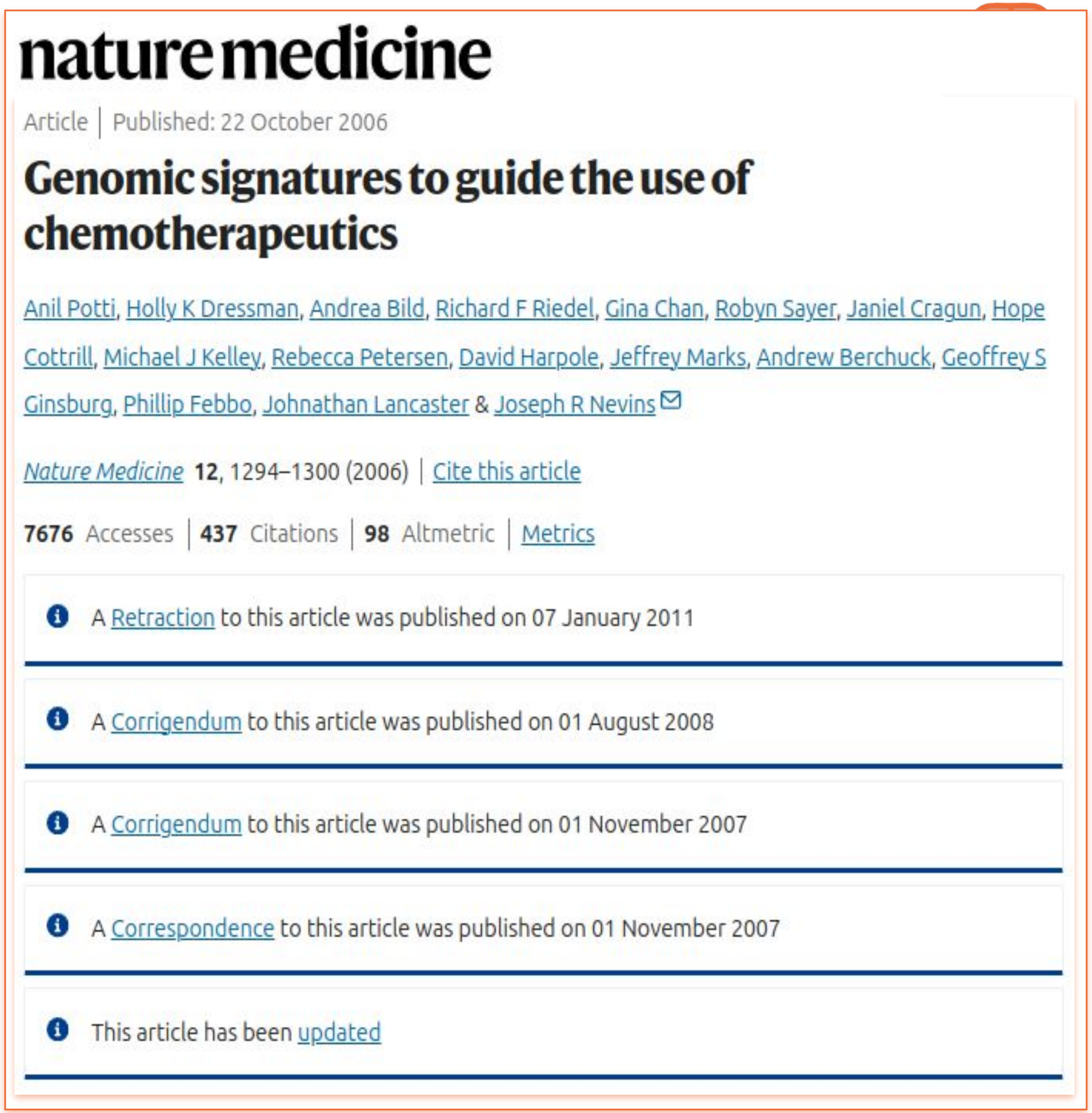
- Swapped “case” and “control” labels
- Some patients duplicated
- Some results ascribed to wrong drug
- Lack of documentation and code
- Likely analysed data with Excel, MatLab and other tools



The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1309–1334
DOI: 10.1214/09-AOAS291
© Institute of Mathematical Statistics, 2009

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY[1] AND KEVIN R. COOMBES[2]



nature medicine

Article | Published: 22 October 2006

Genomic signatures to guide the use of chemotherapeutics

Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster & Joseph R Nevins

Nature Medicine **12**, 1294–1300 (2006) | Cite this article

7676 Accesses | **437** Citations | **98** Altmetric | Metrics

A Retraction to this article was published on 07 January 2011

A Corrigendum to this article was published on 01 August 2008

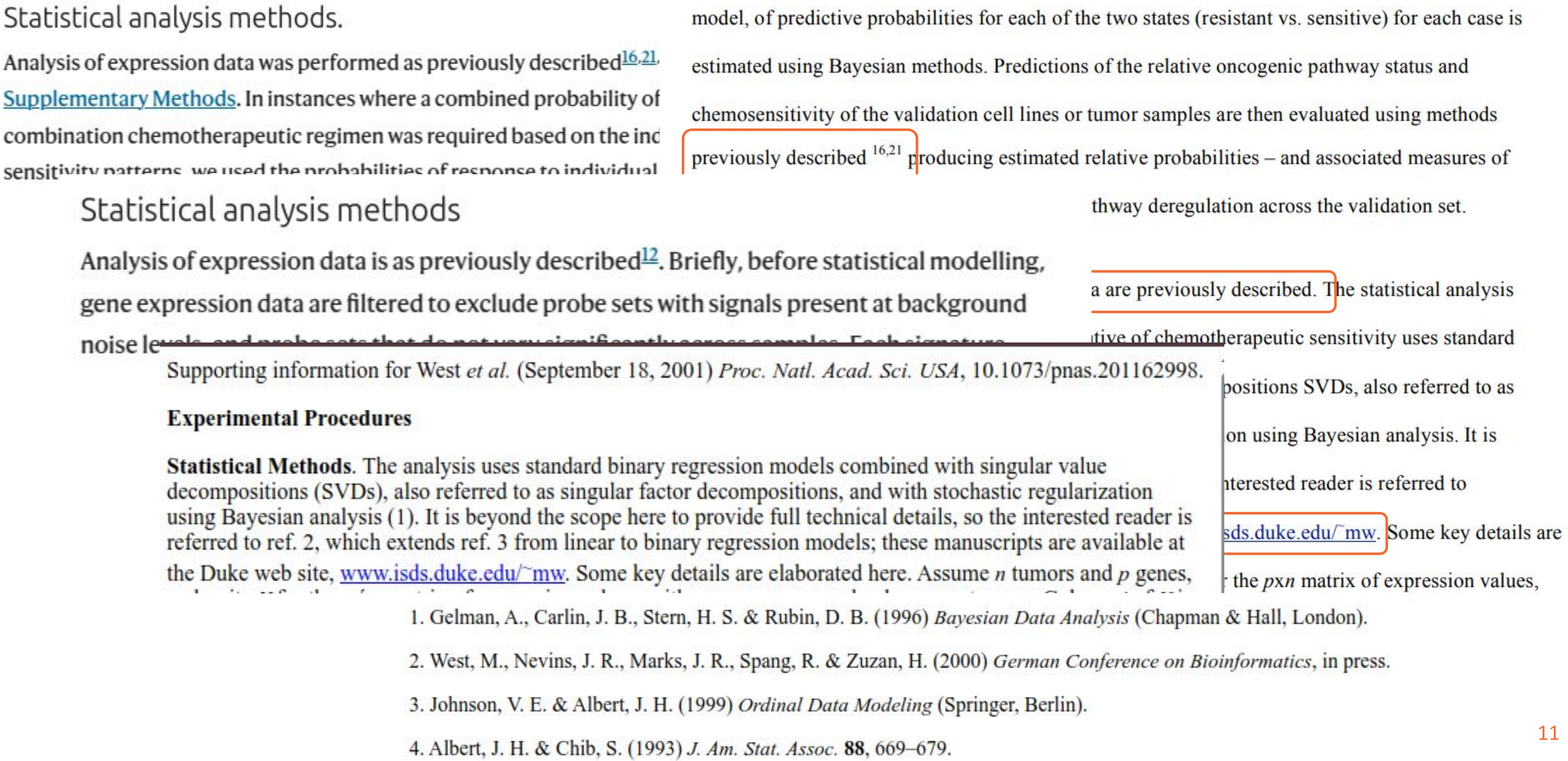
A Corrigendum to this article was published on 01 November 2007

A Correspondence to this article was published on 01 November 2007

This article has been updated

10. Potti et al, 2006; 11. Baggerly & Coombes 2010.

Case study - Potti et al



Statistical analysis methods.

Analysis of expression data was performed as previously described[16,21]. Supplementary Methods. In instances where a combined probability of combination chemotherapeutic regimen was required based on the ind sensitivity patterns, we used the probabilities of response to individual

model, of predictive probabilities for each of the two states (resistant vs. sensitive) for each case is estimated using Bayesian methods. Predictions of the relative oncogenic pathway status and chemosensitivity of the validation cell lines or tumor samples are then evaluated using methods previously described [16,21] producing estimated relative probabilities – and associated measures of

Statistical analysis methods

Analysis of expression data is as previously described[12]. Briefly, before statistical modelling, gene expression data are filtered to exclude probe sets with signals present at background noise le

thway deregulation across the validation set.

a are previously described. The statistical analysis

itive of chemotherapeutic sensitivity uses standard

positions SVDs, also referred to as

on using Bayesian analysis. It is

nterested reader is referred to

sds.duke.edu/~mw. Some key details are

the *pxn* matrix of expression values,

Supporting information for West *et al.* (September 18, 2001) *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.201162998.

Experimental Procedures

Statistical Methods. The analysis uses standard binary regression models combined with singular value decompositions (SVDs), also referred to as singular factor decompositions, and with stochastic regularization using Bayesian analysis (1). It is beyond the scope here to provide full technical details, so the interested reader is referred to ref. 2, which extends ref. 3 from linear to binary regression models; these manuscripts are available at the Duke web site, www.isds.duke.edu/~mw. Some key details are elaborated here. Assume *n* tumors and *p* genes,

1. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1996) *Bayesian Data Analysis* (Chapman & Hall, London).

2. West, M., Nevins, J. R., Marks, J. R., Spang, R. & Zuzan, H. (2000) *German Conference on Bioinformatics*, in press.

3. Johnson, V. E. & Albert, J. H. (1999) *Ordinal Data Modeling* (Springer, Berlin).

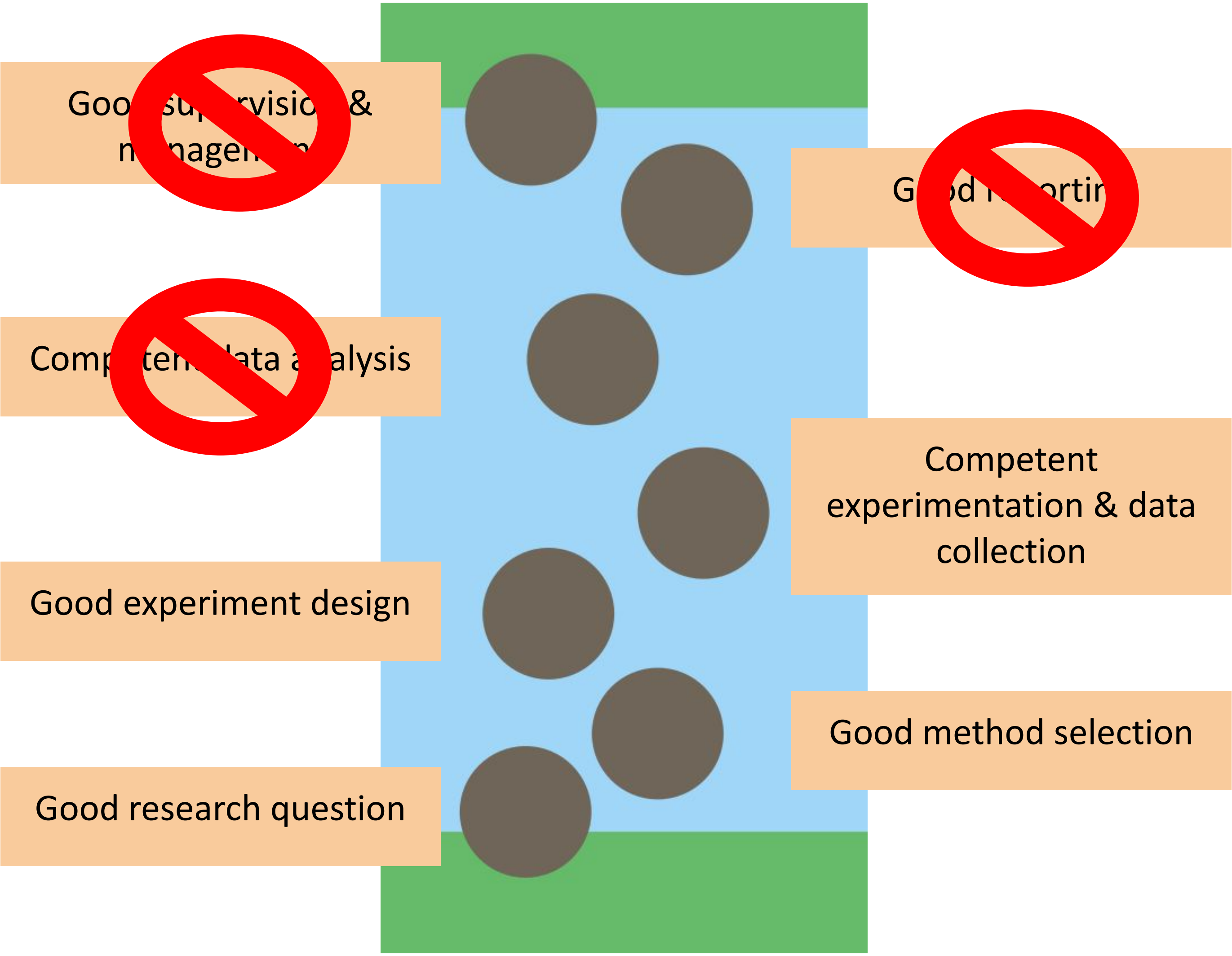
4. Albert, J. H. & Chib, S. (1993) *J. Am. Stat. Assoc.* **88**, 669–679.

Case study outcome

- Retraction of at least 9 research papers
- Three clinical trial ran from 2007 to 2010 involving 117 patients [11]
- Potti was suspended and he later resigned after investigations found fraudulent claims in other internal documents including grant applications
- CancerGuide Diagnostics company collapsed
- Duke was served eight lawsuits from families of deceased trial participants seeking compensation
- Reputational loss

12. The Cancer Letter, 2015; 13. Kaiser 2015.

Case study



Goo su rvisio & n nagen n

Comp ten lata a alysis

Good experiment design

Good research question

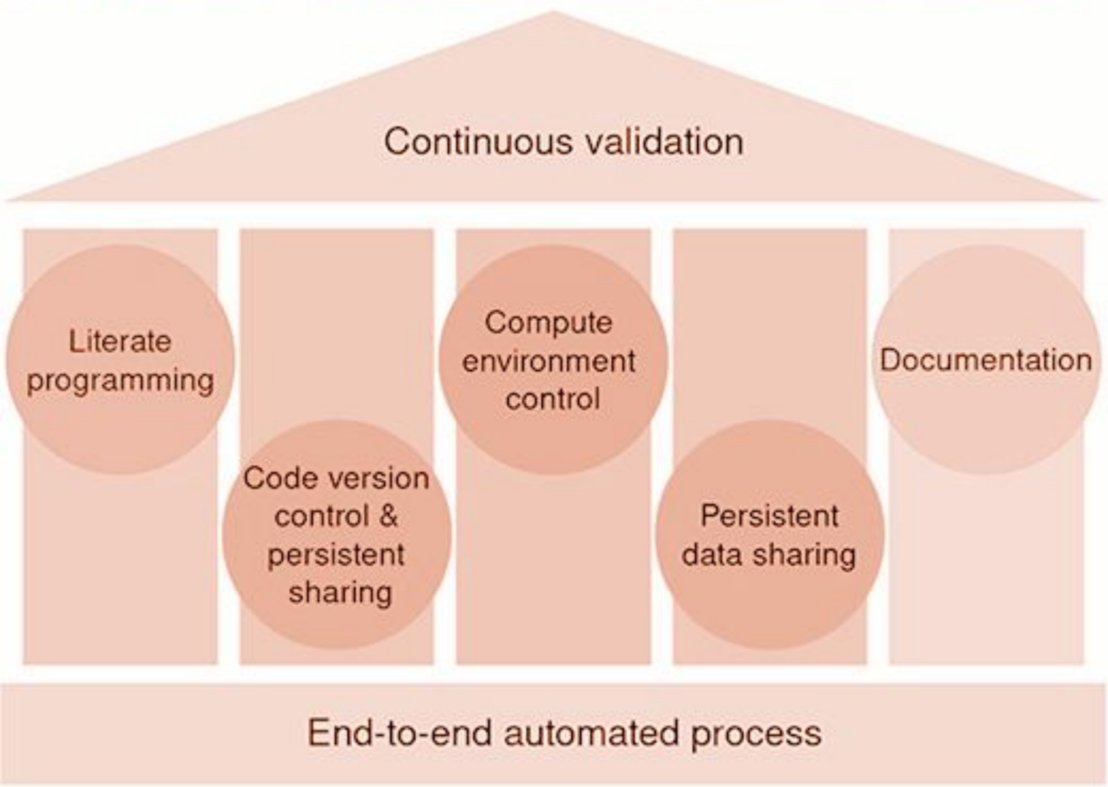
G od r ortir

Competent experimentation & data collection

Good method selection

The five pillars

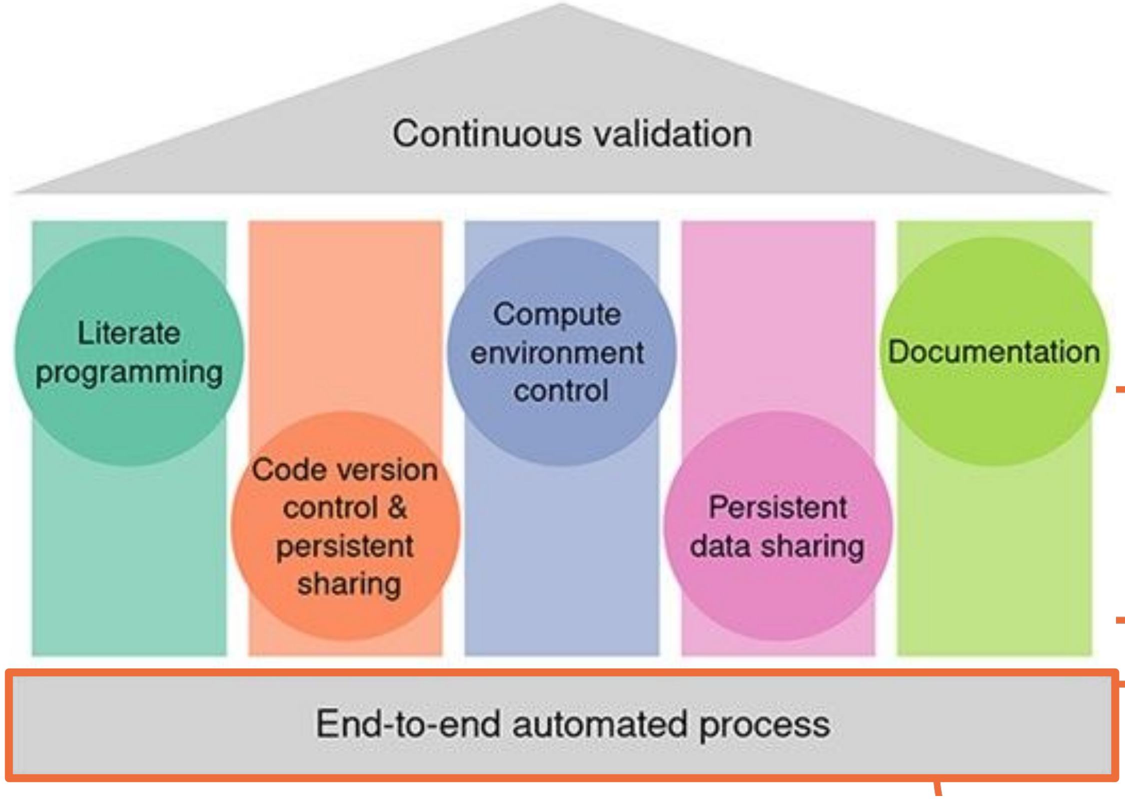
A framework for reproducibility and auditability


Continuous validation
Literate programming
Code version control & persistent sharing
Compute environment control
Persistent data sharing
Documentation
End-to-end automated process

14. *Ziemann et al, 2023.*

Foundation: Automated process

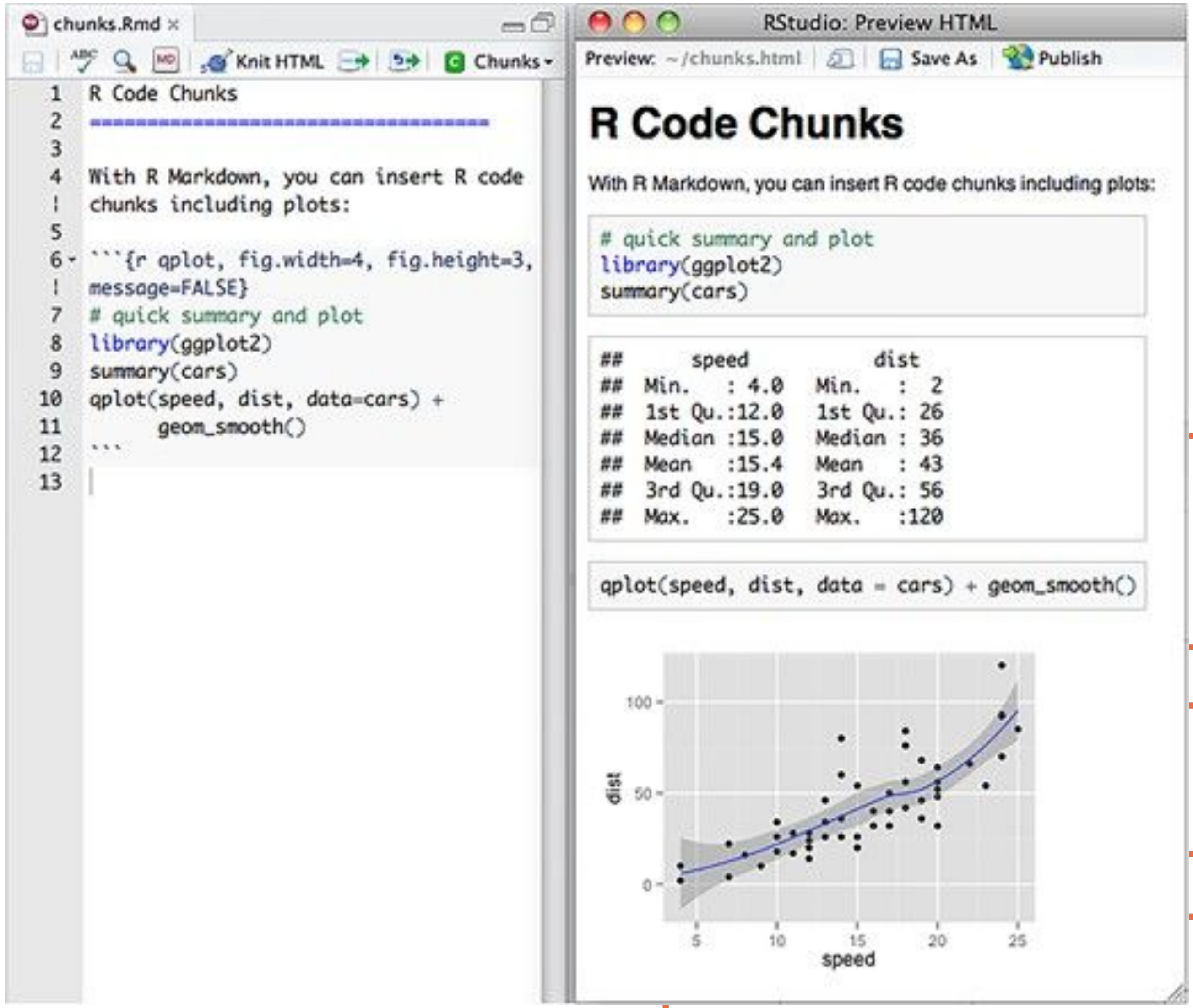
- Manual processes incl spreadsheets and web tools cannot reach high degree of reproducibility
- Methodological descriptions often omit key details, which is why code is better
- End-to-end: from fetching data to generating charts, tables and facts


Continuous validation
Literate programming
Code version control & persistent sharing
Compute environment control
Persistent data sharing
Documentation
End-to-end automated process

14. Ziemann et al 2023

Pillar 1: Literate programming

- Literate programming combines ‘chunks’ of analytical code with human-readable text
- Rendered report contains key figures, tables and data - in context and in order
- Demonstrates provenance
- Options: R Markdown, Jupyter, Quarto



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

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

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

R Code Chunks

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

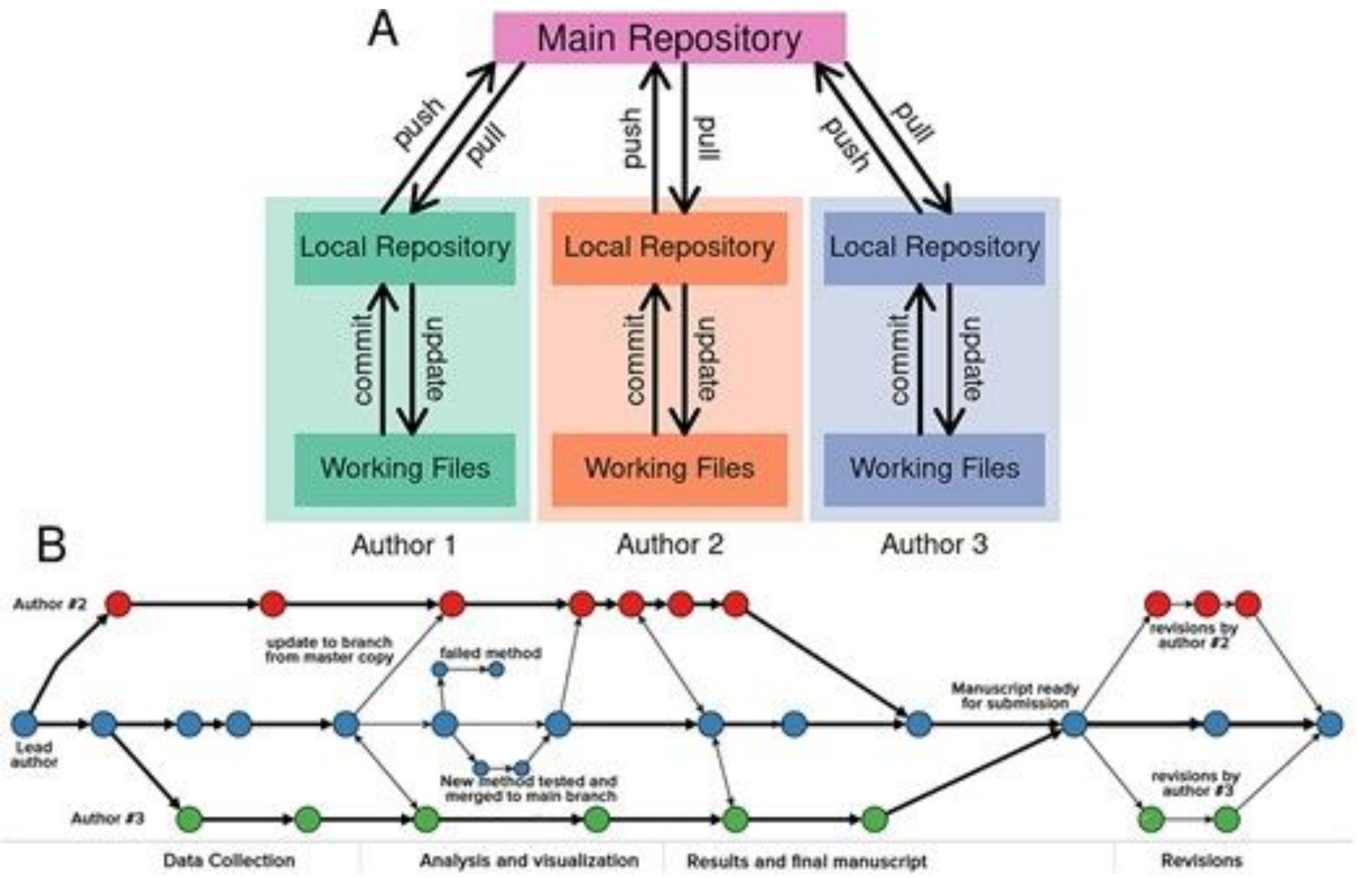
```
# quick summary and plot
library(ggplot2)
summary(cars)
```

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

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

15. *Grolemund & Wickham 2017*

Pillar 2: Code management

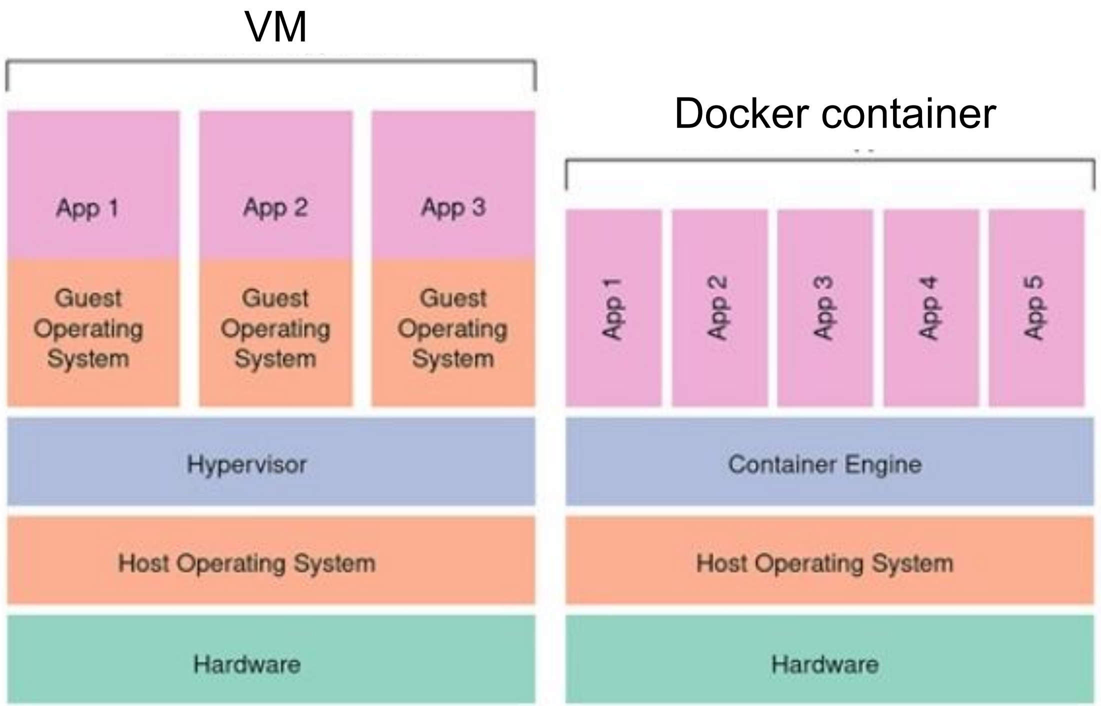


16. Ram 2013

- “Track changes” for large and complex workflow scripts and documentation
- Assists with project management (milestones, issue tracking, task allocation, etc)
- Easy distribution to consumers
- Not a solution to long-term code preservation. Software Heritage and Zenodo are good for that

Pillar 3: Compute environment control

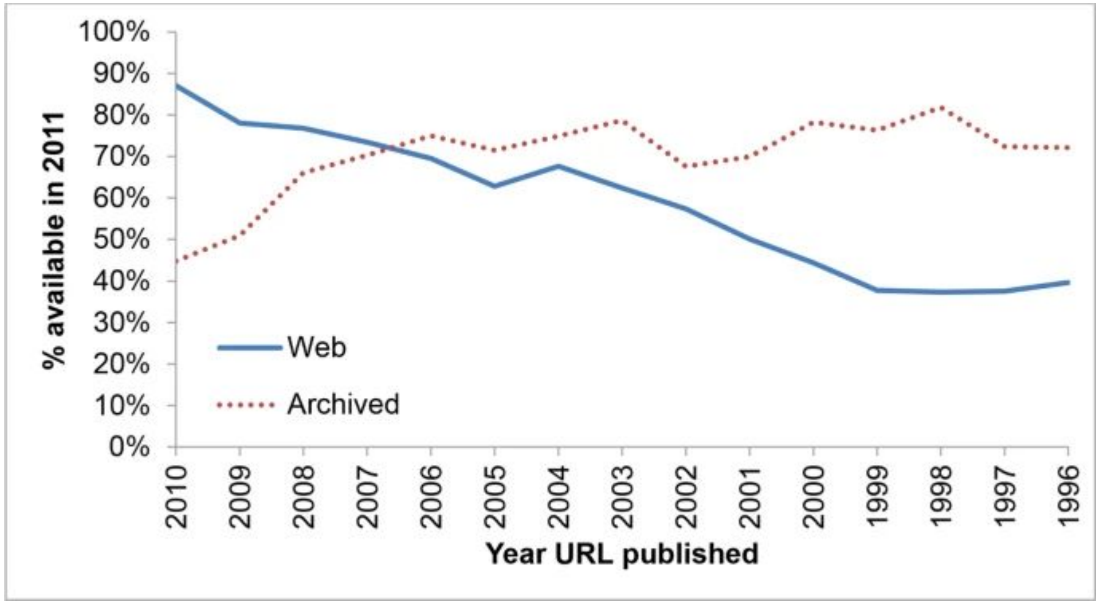
- Code and data are insufficient to reproduce computational research, we also need the “environment” - the set of software dependencies
- To simplify reproducibility, we should be providing “virtual machines” or “containers” loaded with the software and configuration needed to accurately execute the analysis according to the publication
- Dockerhub is a convenient way to share container images, but isn’t a solution for long term preservation
- Docker, Apptainer and GNU Guix are good options



```
sudo apt update && sudo apt install docker.io -y # install docker

sudo docker run -it --entrypoint /bin/bash mziemann/enrichment_recipe # enter container

Rscript -e 'rmarkdown::render("example.Rmd")' # execute workflow

exit # exit container

docker cp $(docker ps -aql):/enrichment_recipe/example.html . # copy report to host system

firefox example.html # inspect results
```

Pillar 4: Persistent data sharing

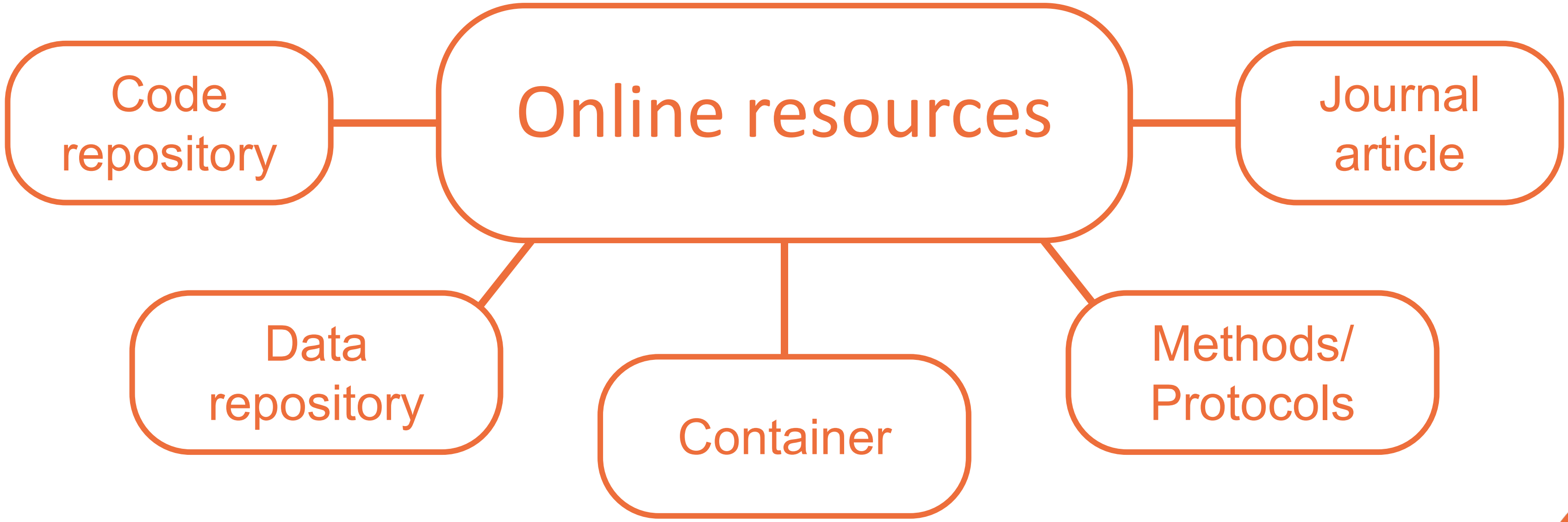
The accessibility of URLs in journal articles declines with age



17. Hennessey & Ge 2013

- Genomics has a culture of data sharing, but this is not universal in medicine or other aspects of life science
- Use a dedicated data repository for the specific type of data, or Zenodo for other types
- Avoid DropBox, Google Drive and other ephemeral cloud providers
- Avoid large supplementary files, these are not findable
- Ensure the data labels are consistent with the vocabulary of the journal article

Pillar 5: Documentation

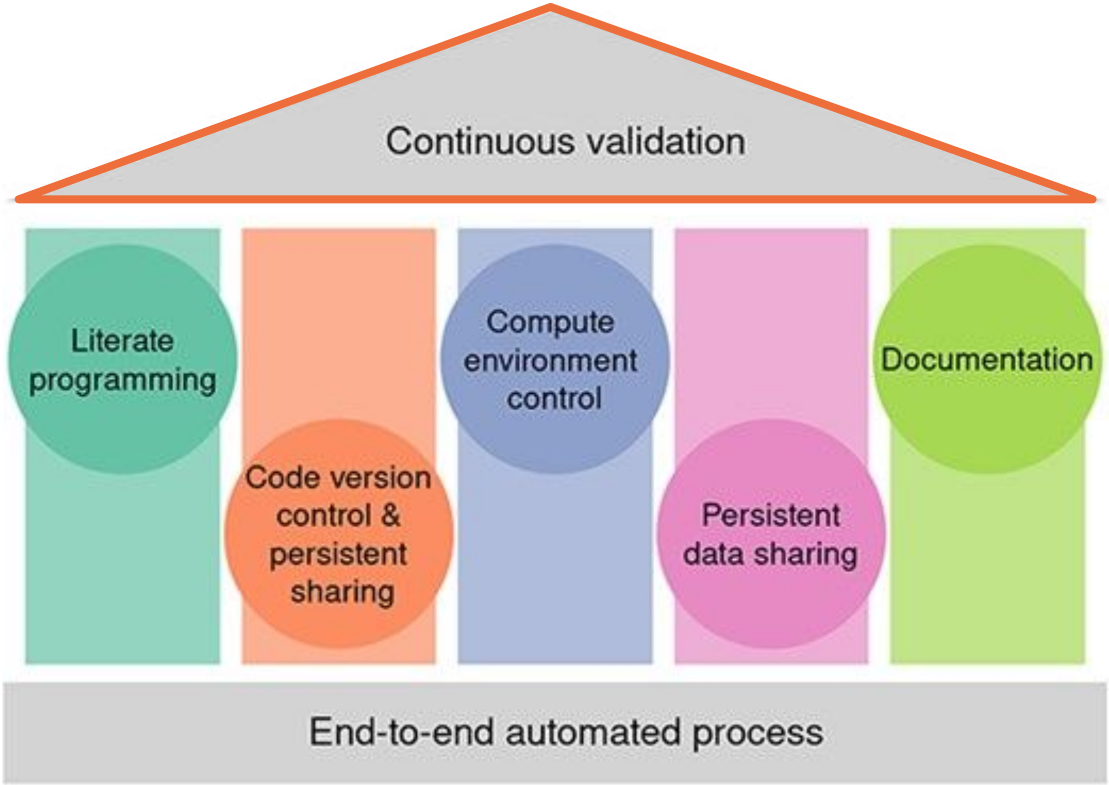


Documentation is “glue”

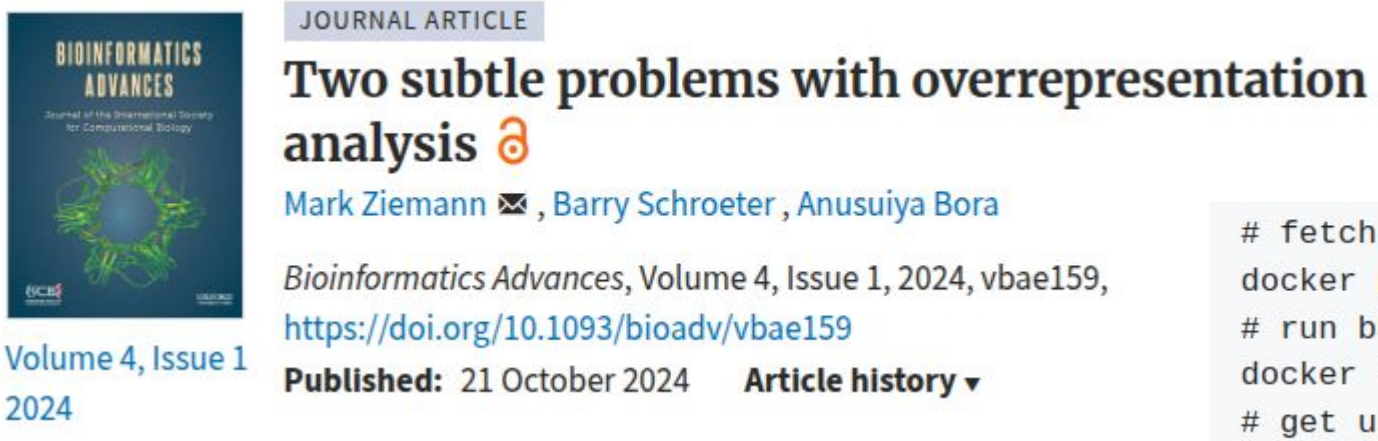
- Where to find all the necessary resources?
- How to reproduce it?
- What computational resources are needed?
- How to raise issues and contribute?

Pediment: Continuous validation

- A lot can go wrong in a research workflow, so "sanity checks" are essential; small tests to spot irregularities in data or results
- Human readable sanity checks to be saved in the compiled report
- Checks are run each time the code or data set undergoes changes


Continuous validation
Literate programming
Code version control & persistent sharing
Compute environment control
Persistent data sharing
Documentation
End-to-end automated process

Practicing what we preach



JOURNAL ARTICLE

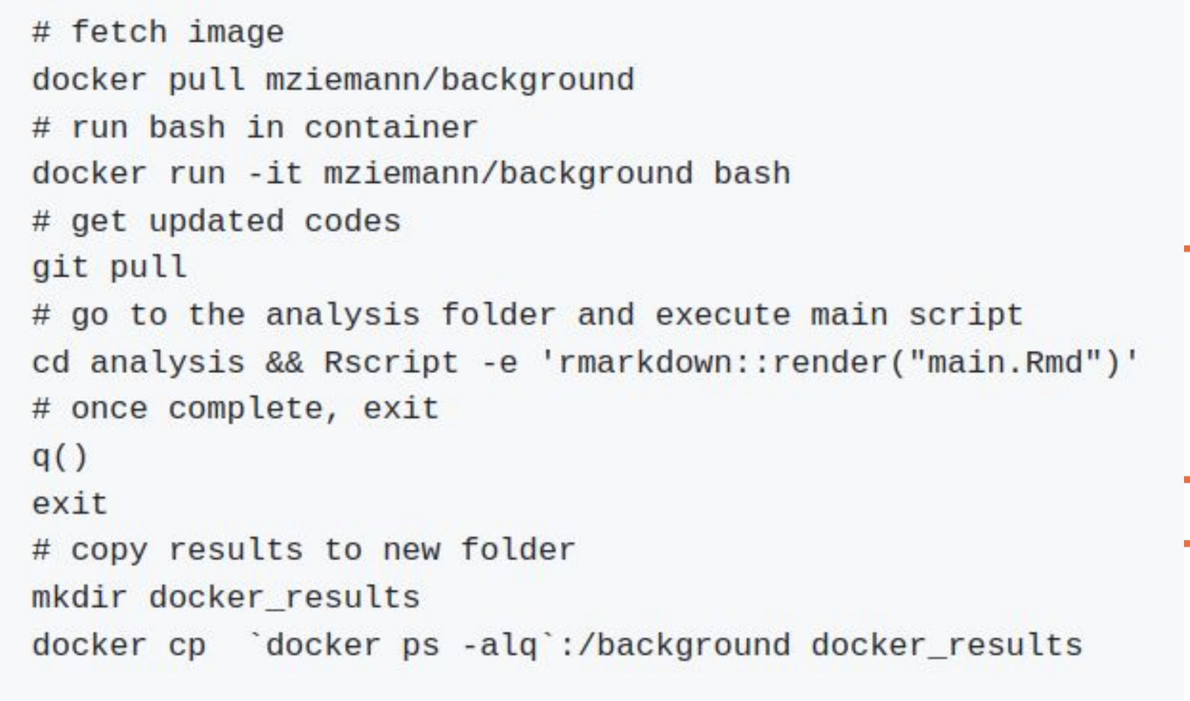
Two subtle problems with overrepresentation analysis

Mark Ziemann ✉ , Barry Schroeter , Anusuiya Bora

Bioinformatics Advances, Volume 4, Issue 1, 2024, vbae159, https://doi.org/10.1093/bioadv/vbae159

Published: 21 October 2024 **Article history ▾**

Volume 4, Issue 1 2024

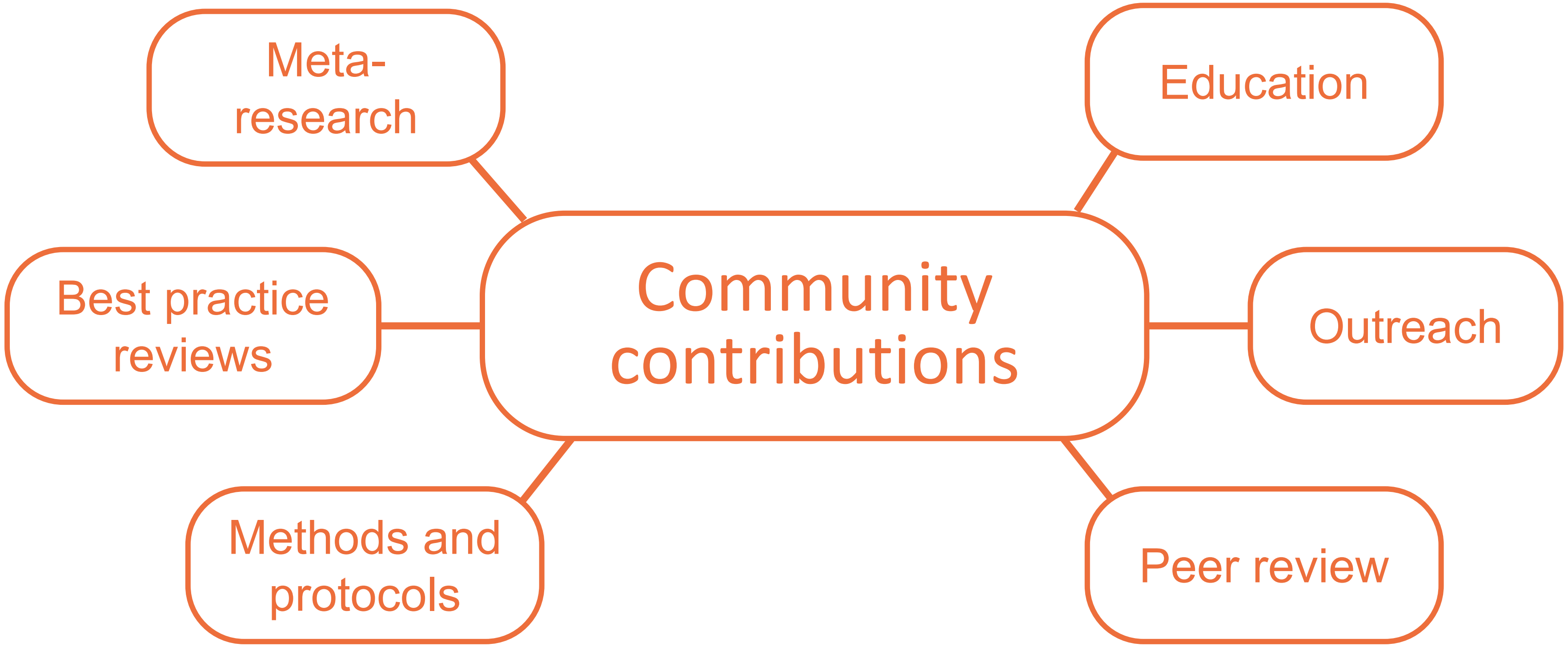


```
# fetch image
docker pull mziemann/background
# run bash in container
docker run -it mziemann/background bash
# get updated codes
git pull
# go to the analysis folder and execute main script
cd analysis && Rscript -e 'rmarkdown::render("main.Rmd")'
# once complete, exit
q()
exit
# copy results to new folder
mkdir docker_results
docker cp  `docker ps -alq`:/background docker_results
```

- Publicly available data
- Code on GitHub and Zenodo
- Docker image on Zenodo
- R/Shiny tool for interacting
- Validated data-to-manuscript script

18. Ziemann et al, 2024

Towards a reliable research culture



Meta-research

Education

Best practice reviews

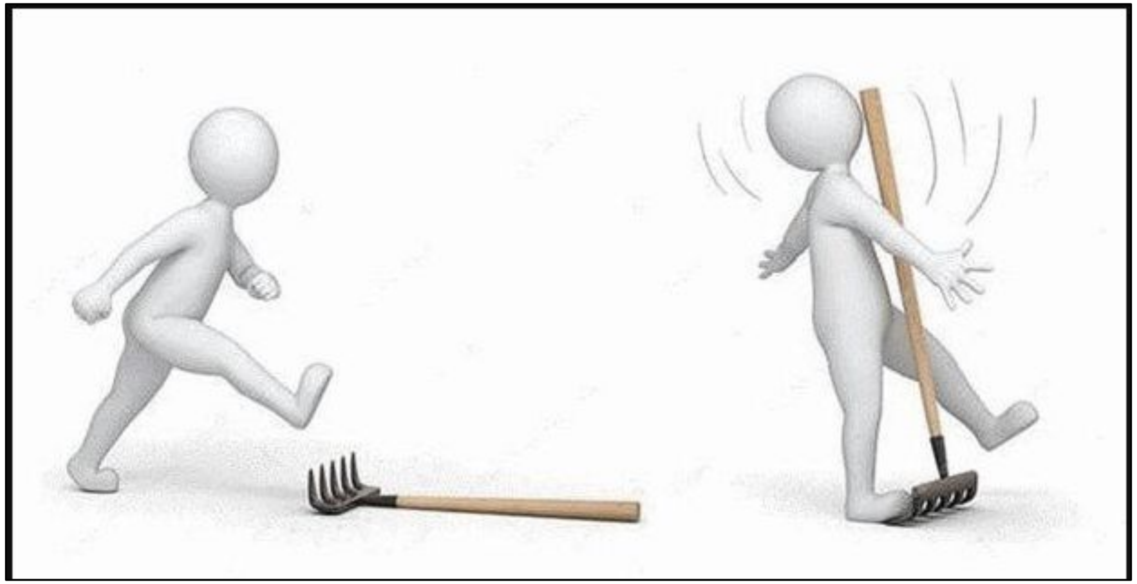
Community contributions

Outreach

Methods and protocols

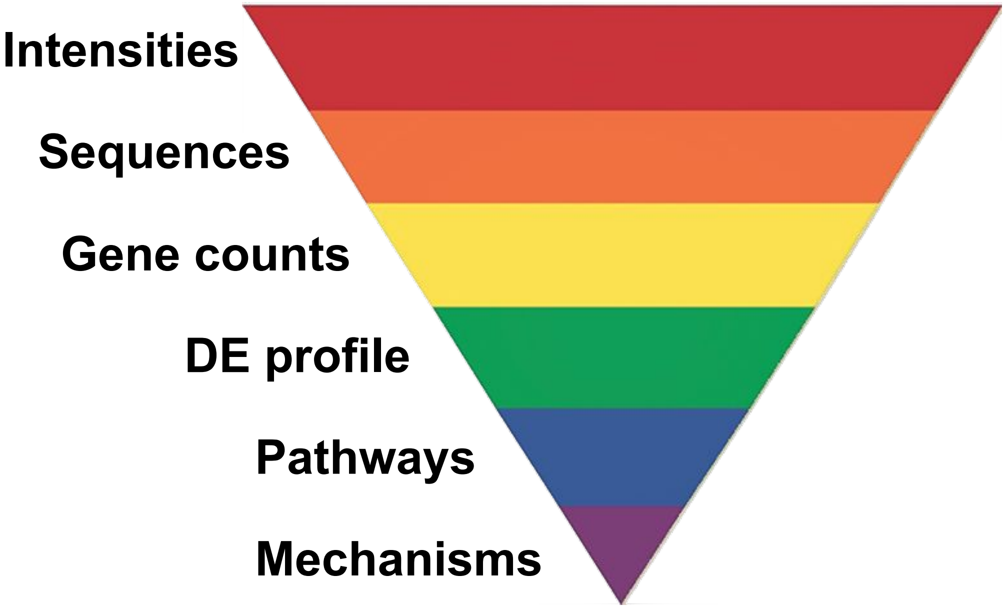
Peer review

Meta-research on pathway enrichment analysis methodology



Pathway enrichment analysis

- Also known as “functional enrichment analysis”, “gene set analysis” or “ontology analysis”
- A class of tools used to summarise omics data to examine the differential regulation of known biological pathways
- Contains clues about “mechanisms” critical to conclusions of biological studies
- Applicable to diverse data sets
- Highly cited, 67k abstract mentions in PubMed



Intensities

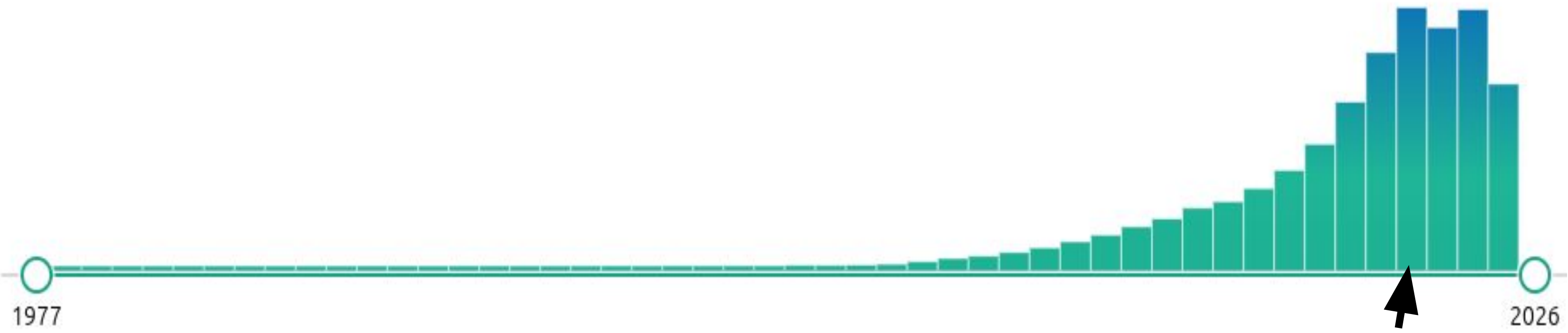
Sequences

Gene counts

DE profile

Pathways

Mechanisms

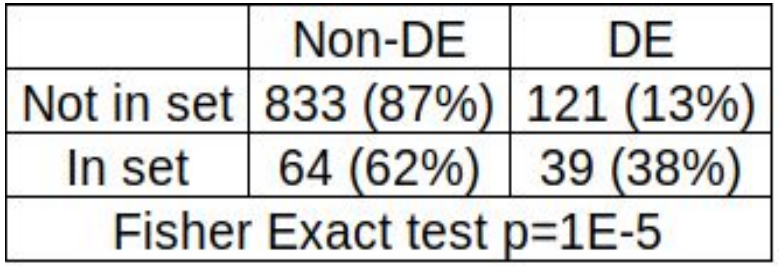


2023: 9,313

Two approaches to pathway analysis

Over-representation analysis (ORA)

Selected genes meeting an arbitrary significance threshold are tested for enrichment in different “pathways” (gene sets) as compared to a background list. Typically uses hypergeometric test.

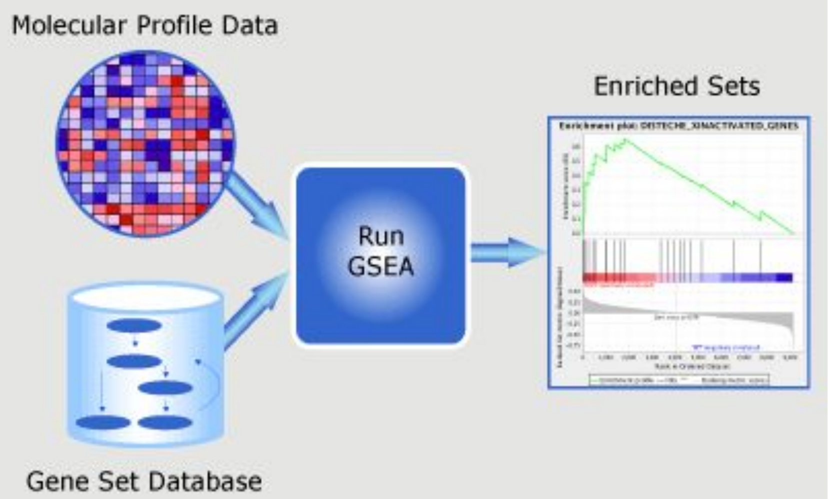


	Non-DE	DE
Not in set	833 (87%)	121 (13%)
In set	64 (62%)	39 (38%)
Fisher Exact test p=1E-5		

- Easy & fast
- Dependent on threshold selection
- Less sensitive

Functional class scoring (FCS)

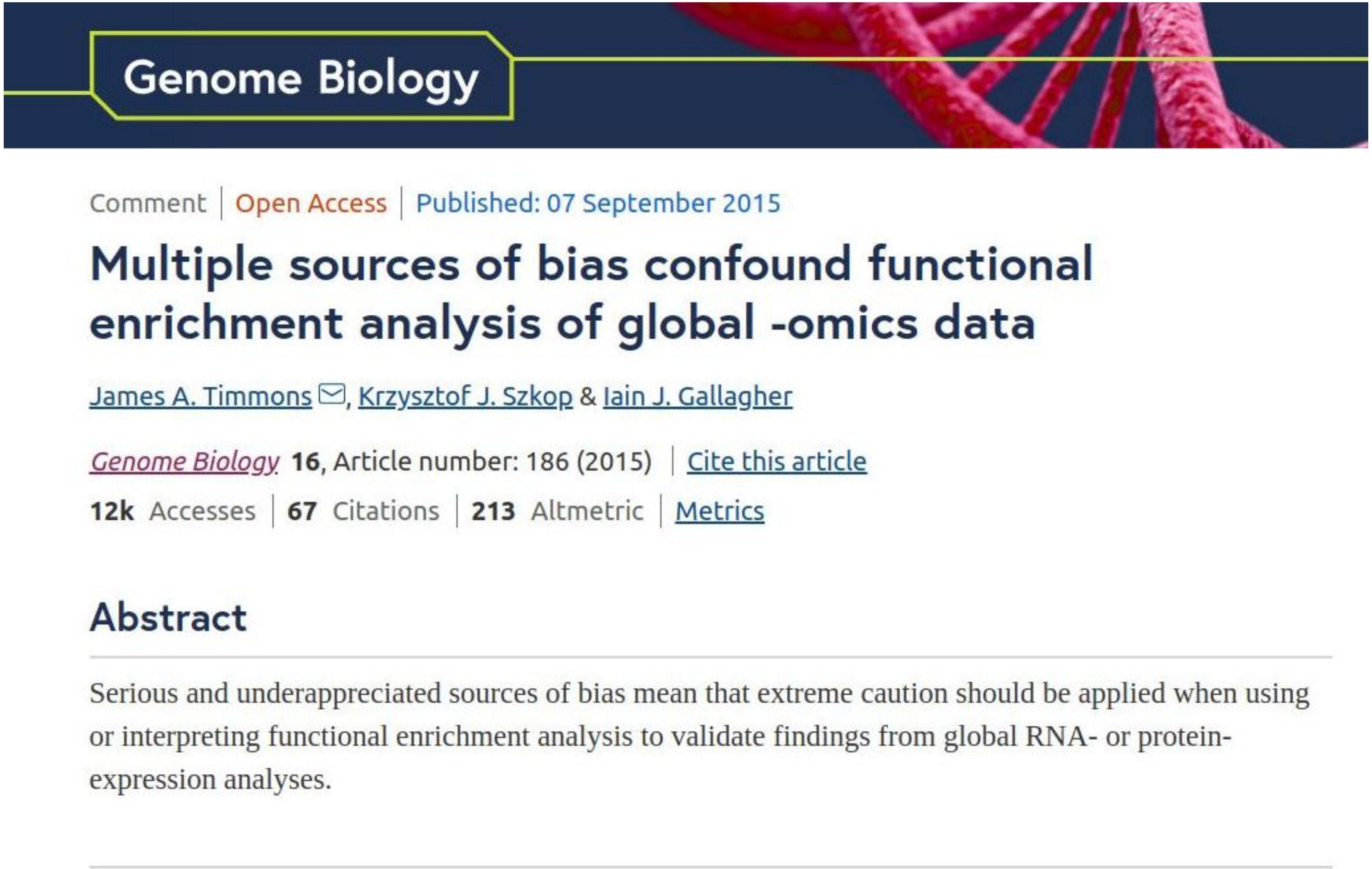
All detected genes are ranked by a differential abundance score (eg: fold change, t-stat) followed by a test to examine whether genes belonging to a set have a non-random distribution.



- More sensitive
- More complicated

19. Khatri et al, 2012

Methodological issues



Comment | Open Access | Published: 07 September 2015

Multiple sources of bias confound functional enrichment analysis of global -omics data

James A. Timmons, Krzysztof J. Szkop & Iain J. Gallagher

Genome Biology **16**, Article number: 186 (2015) | Cite this article

12k Accesses | **67** Citations | **213** Altmetric | Metrics

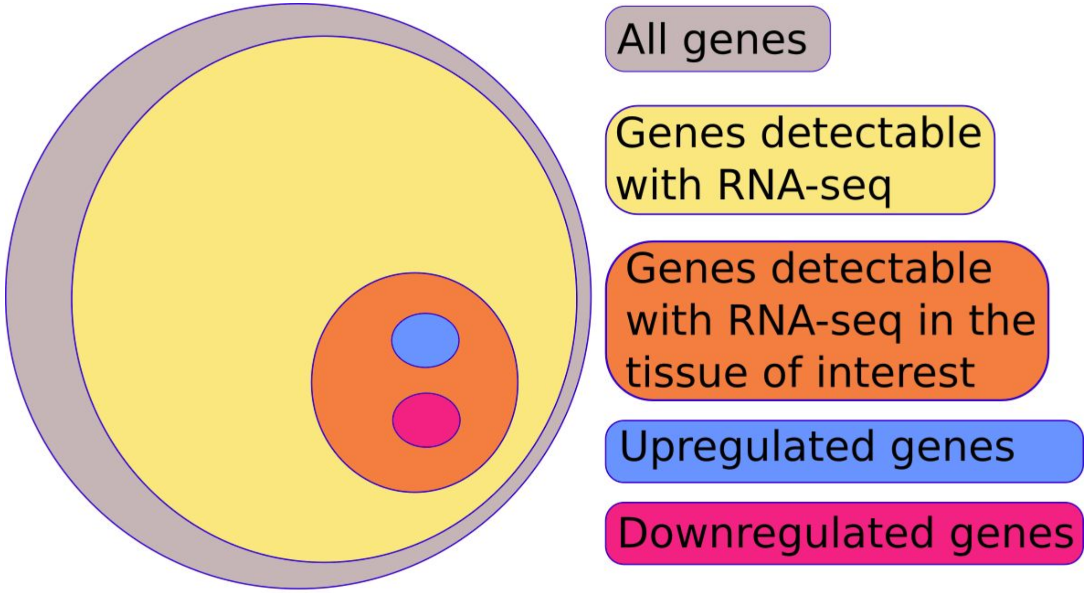
Abstract

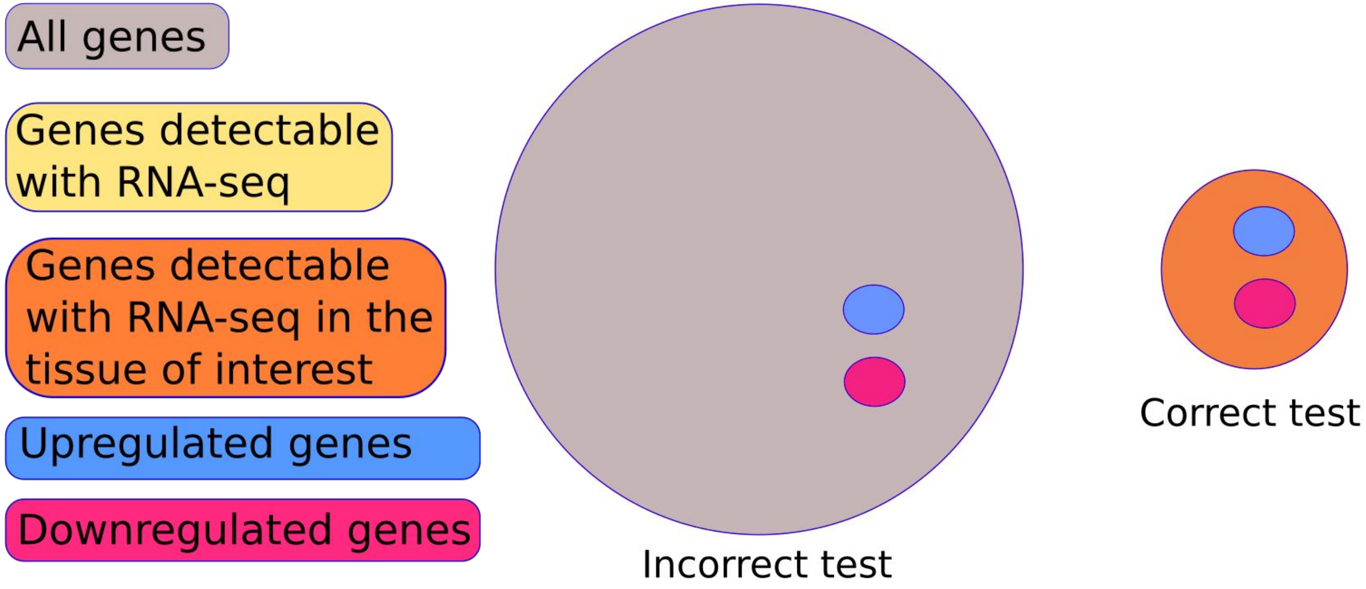
Serious and underappreciated sources of bias mean that extreme caution should be applied when using or interpreting functional enrichment analysis to validate findings from global RNA- or protein-expression analyses.

20. Timmons et al, 2015

ORA methodological issue: sampling bias

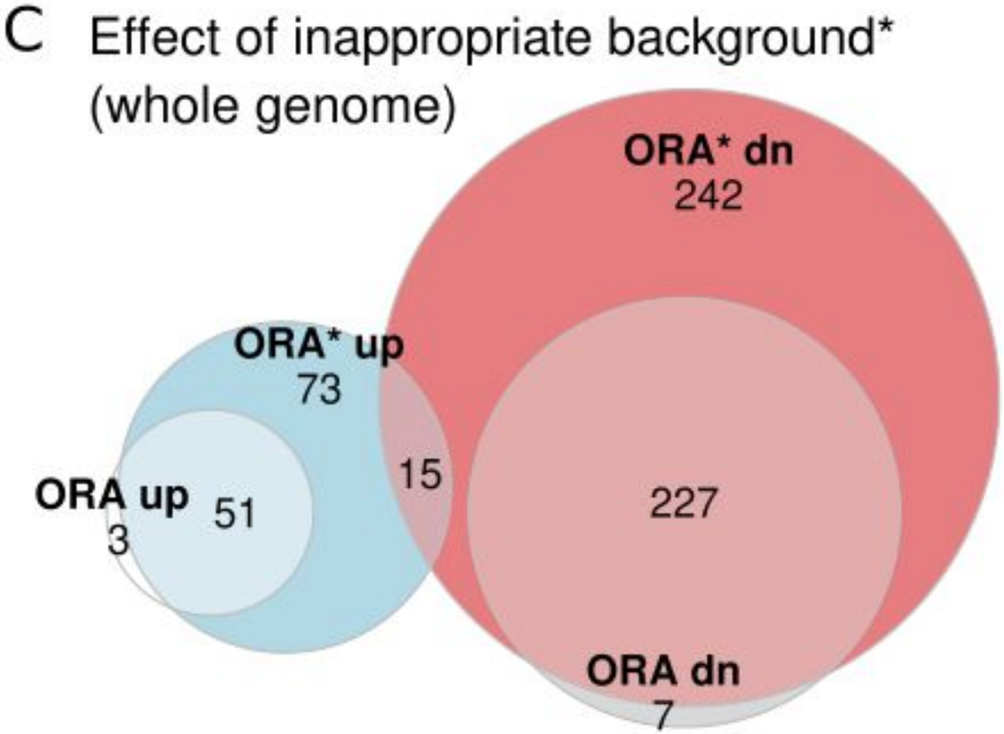
- In any cell or tissue, most genes are silent
- Dysregulated genes are a subset of expressed genes
- Therefore enrichment should be determined by comparison to other expressed genes (~15k), not the whole set of annotated genes (~60k)


All genes
Genes detectable with RNA-seq
Genes detectable with RNA-seq in the tissue of interest
Upregulated genes
Downregulated genes


All genes
Genes detectable with RNA-seq
Genes detectable with RNA-seq in the tissue of interest
Upregulated genes
Downregulated genes
Incorrect test
Correct test

Consequences

- Example shows wrong background(*) caused 330 type-I errors and 10 type-II errors (Jaccard=0.44).
- Impact is worse than omitting false discovery rate correction for multiple testing (Jaccard=0.56)

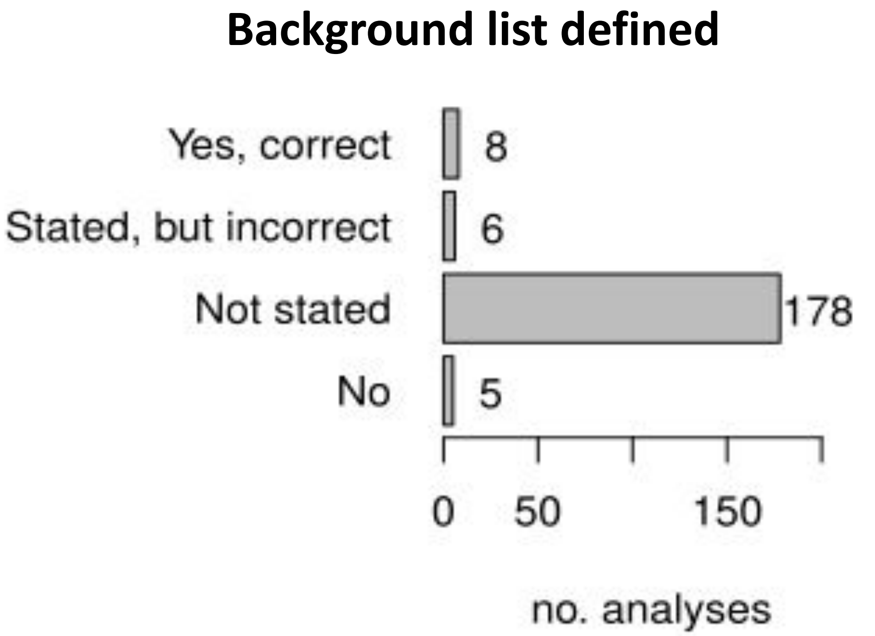


Significant pathways (FDR<0.05)

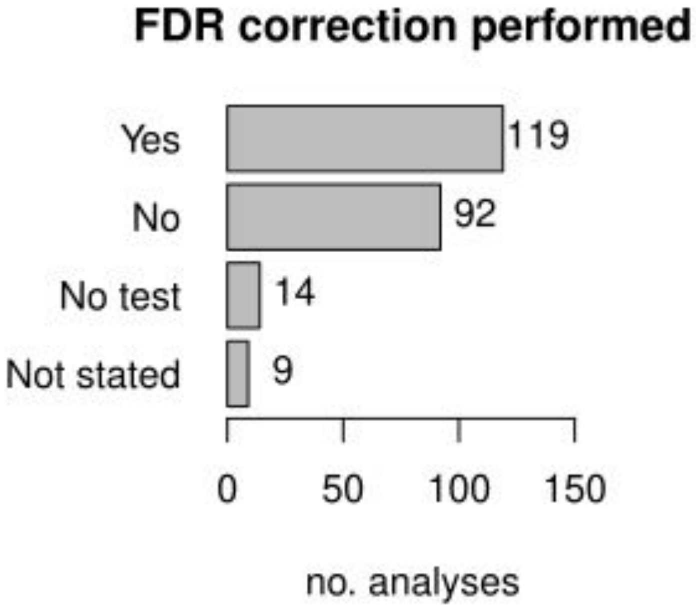
21. Wijesooriya et al. 2022.

How common are errors in the literature?

- Background list correctly reported in only 4% of 197 studies using ORA [20]
- Only 50% of studies conducted FDR correction [20]
- A preliminary study of 147 high impact articles (SJR>5) shows slightly better results[21]:
  - Correct background: 4% -> 16%
  - Correct FDR: 50% -> 59%
- Are researchers using poor methodology because it gives them more “significant” results? [22]



Background list defined

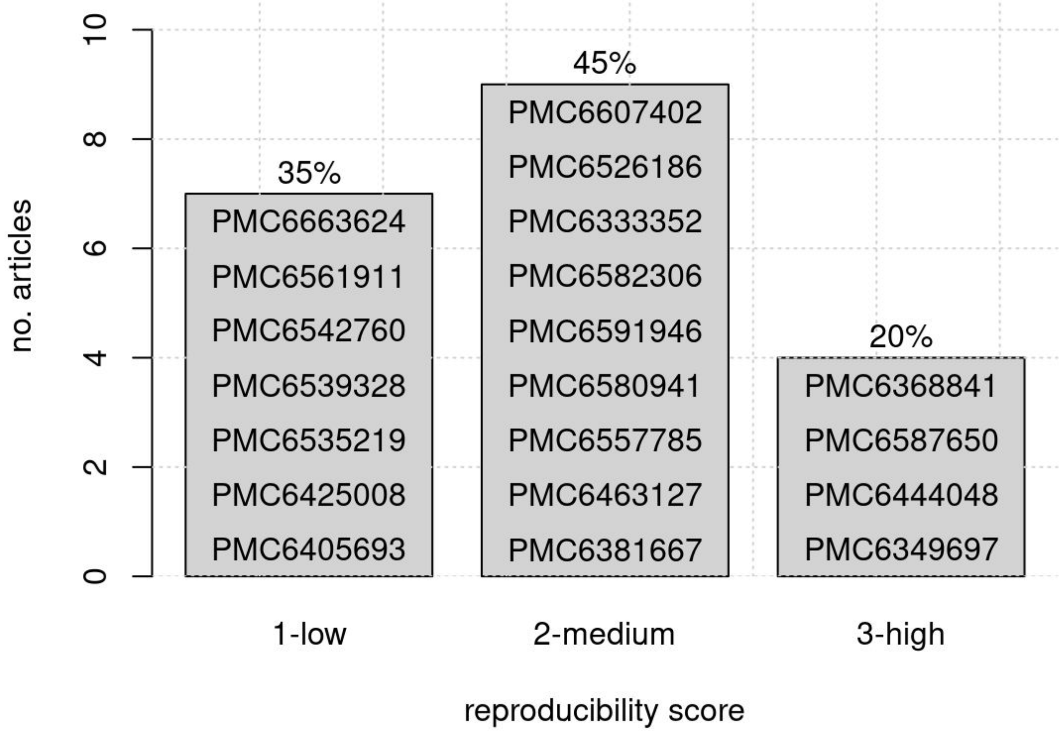


FDR correction performed

21. Wijesooriya et al. 2022.
22. Unpublished results
23. Smaldino & McElreath 2016.

Errors result in poor reproducibility

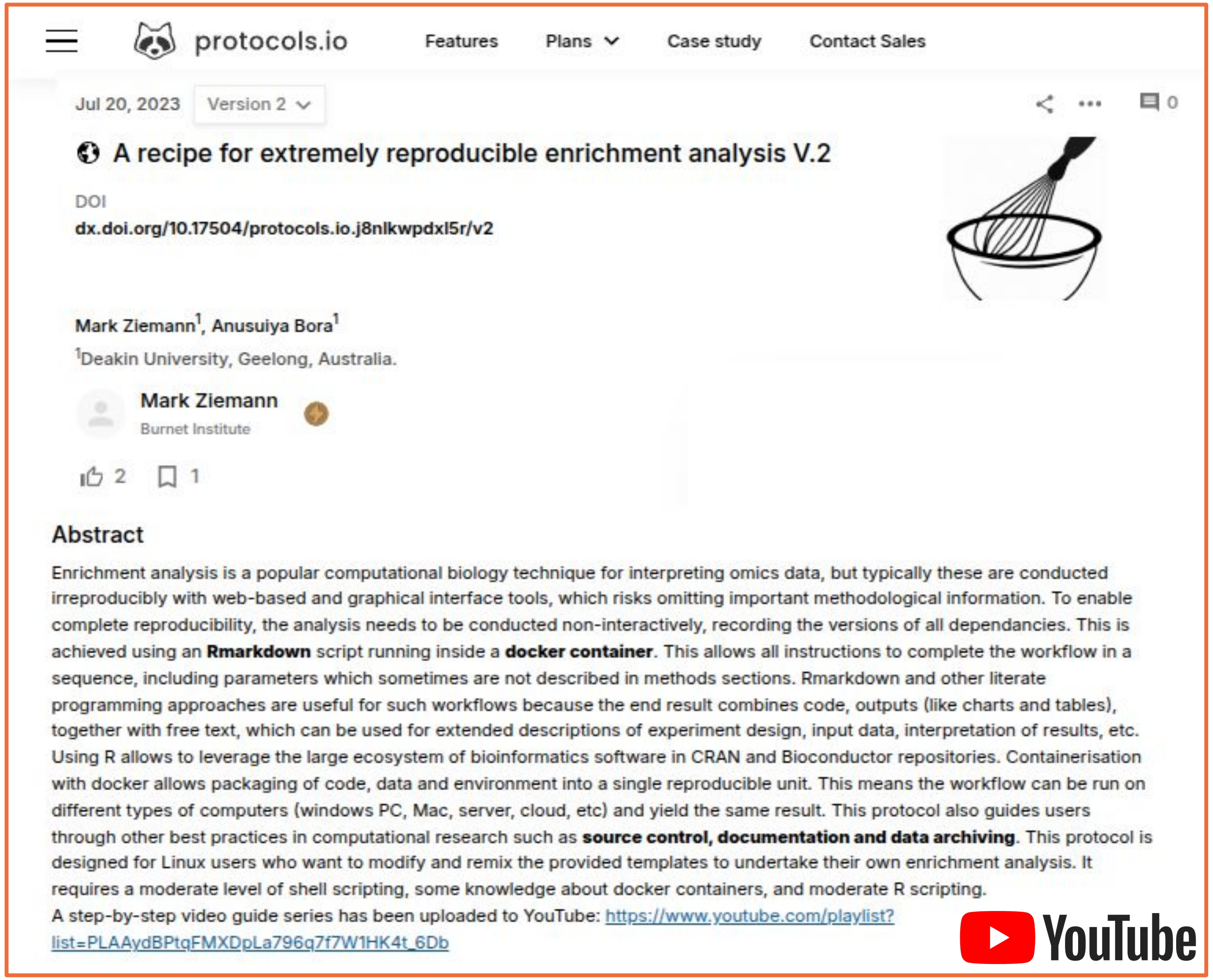
A pilot reproducibility study of 20 enrichment studies from 2019 shows only 4 were highly reproducible while 7 had severe problems that compromised conclusions



DAVID 6.8 and earlier are no longer available (~20,000 pubmed articles)

24. Bora & Ziemann 2023.

Protocol



Jul 20, 2023

A recipe for extremely reproducible enrichment analysis V.2

DOI

dx.doi.org/10.17504/protocols.io.j8nlkwpdxl5r/v2

Mark Ziemann[1], Anusuiya Bora[1]

[1]Deakin University, Geelong, Australia.

Mark Ziemann

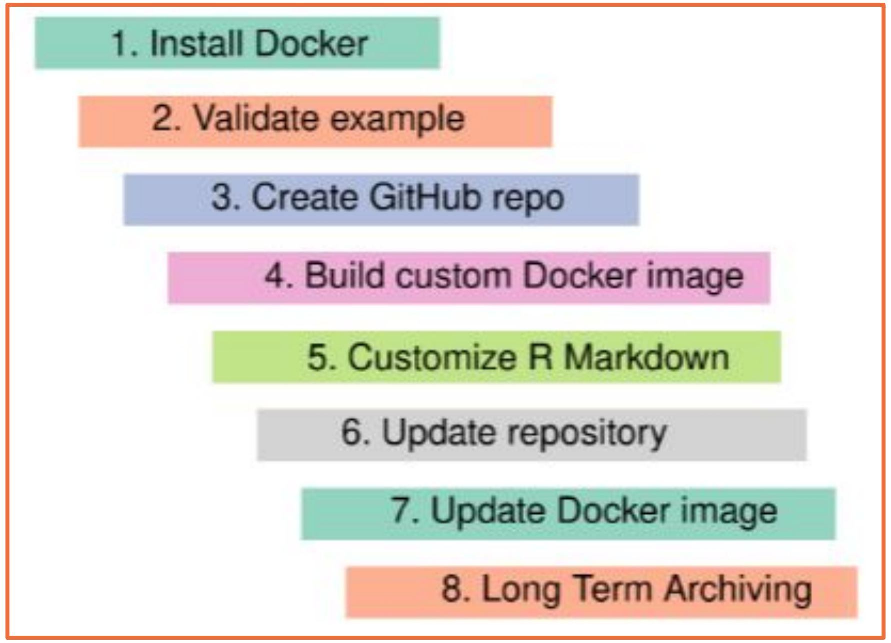
Burnet Institute

2 1

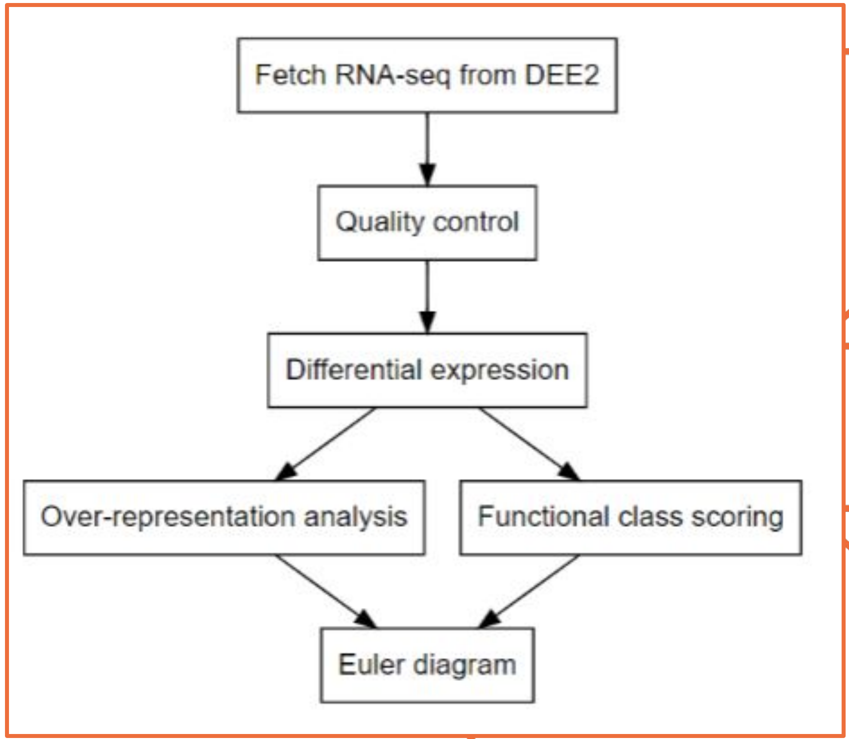
Abstract

Enrichment analysis is a popular computational biology technique for interpreting omics data, but typically these are conducted irreproducibly with web-based and graphical interface tools, which risks omitting important methodological information. To enable complete reproducibility, the analysis needs to be conducted non-interactively, recording the versions of all dependancies. This is achieved using an **Rmarkdown** script running inside a **docker container**. This allows all instructions to complete the workflow in a sequence, including parameters which sometimes are not described in methods sections. Rmarkdown and other literate programming approaches are useful for such workflows because the end result combines code, outputs (like charts and tables), together with free text, which can be used for extended descriptions of experiment design, input data, interpretation of results, etc. Using R allows to leverage the large ecosystem of bioinformatics software in CRAN and Bioconductor repositories. Containerisation with docker allows packaging of code, data and environment into a single reproducible unit. This means the workflow can be run on different types of computers (windows PC, Mac, server, cloud, etc) and yield the same result. This protocol also guides users through other best practices in computational research such as **source control, documentation and data archiving**. This protocol is designed for Linux users who want to modify and remix the provided templates to undertake their own enrichment analysis. It requires a moderate level of shell scripting, some knowledge about docker containers, and moderate R scripting.

A step-by-step video guide series has been uploaded to YouTube: https://www.youtube.com/playlist?list=PLAAydBPtqFMXDpLa796q7f7W1HK4t_6Db



1. Install Docker
2. Validate example
3. Create GitHub repo
4. Build custom Docker image
5. Customize R Markdown
6. Update repository
7. Update Docker image
8. Long Term Archiving



24. Bora & Ziemann 2023.

Future directions

- Systematic reproducibility analysis of enrichment analyses of single cell transcriptome studies
- Human factors (questionnaire)
- Using LLMs to checklist methodology
- Development of a user-friendly AND reproducible tool

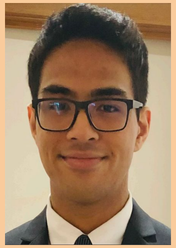
Deakin-Burnet Bioinformatics group members

Anusuiya Bora, PhD Candidate

Towards reliable and reproducible enrichment analysis

Jonathan Salazar, Biomedical Science Hons

Is pathway analysis of single cell transcriptome data reliable?



Kaumadi Wijesooriya, Master of Biotechnology Graduate, Casual Research Assistant.

Past members

Sia Mehta
Dr. Sameer A Jadaan
Tanuveer Kaur

Sehansi Karunaratne
Kaushalya Perera

mark.ziemann@burnet.edu.au

REFERENCES

1. Goodman et al, 2016 DOI:10.1126/scitranslmed.aaf5027
2. Gundersen 2021 DOI:10.1098/rsta.2020.0210
3. Baker 2016, DOI:10.1038/533452a
4. Baker 2015, DOI:10.1038/nature.2015.17433
5. Begley & Ellis 2012 DOI:10.1038/483531a
6 .Cobey et al, 2024, DOI:10.1371/journal.pbio.3002870
7. Ioannidis et al, 2009, DOI:10.1038/ng.295
8. Zaringhalam & Federer 2020, DOI:10.5281/zenodo.3818329
9. Samuel & Mietchen 2024, DOI:10.1093/gigascience/giad113
10. Potti et al, 2006, DOI:10.1038/nm1491
11. Baggerly & Coombes 2010, DOI:10.1214/09-AOAS291
12. The Cancer Letter, https://cancerletter.com/the-cancer-letter/20150123_2/
13. Kaiser 2015 DOI:10.1126/science.aad7410.
14. Ziemann et al, 2023, DOI:10.1093/bib/bbad375
15. Grolemund & Wickham 2017, http://r4ds.hadley.nz
16. Ram 2013 DOI:10.1186/1751-0473-8-7
17. Hennessey & Ge 2013, DOI:10.1186/1471-2105-14-S14-S5
18. Ziemann et al, 2024, DOI:10.1093/bioadv/vbae159
19. Khatri et al, 2012, DOI:10.1371/journal.pcbi.1002375
20. Timmons et al, 2015, DOI:10.1186/s13059-015-0761-7
21. Wijesooriya et al. 2022, DOI:10.1371/journal.pcbi.1009935
22. Unpublished results
23. Smaldino & McElreath 2016, DOI:10.1098/rsos.160384
24. Bora & Ziemann 2023, DOI:10.31219/osf.io/r6kxg