Reproducible Research for Bioinformatics: Best Practices

Bioinformatics and Research Computing (BaRC)

http://jura.wi.mit.edu

Outline

- Examples highlighting issues related to reproducibility
- Defining reproducibility
  - P-hacking
- Best practices and recommendations
- Discussion

National Research Council Statement on Reproducible Research

The general norms of science emphasize the principle openness. Scientists are generally expected to exchange research data as well as unique research materials that are essential to the replication or extension of reported findings.

Responsible Science, Vol I: Ensuring the Integrity of the Research Process (1992), The National Academies Press
https://www.nap.edu/read/1864/chapter/4

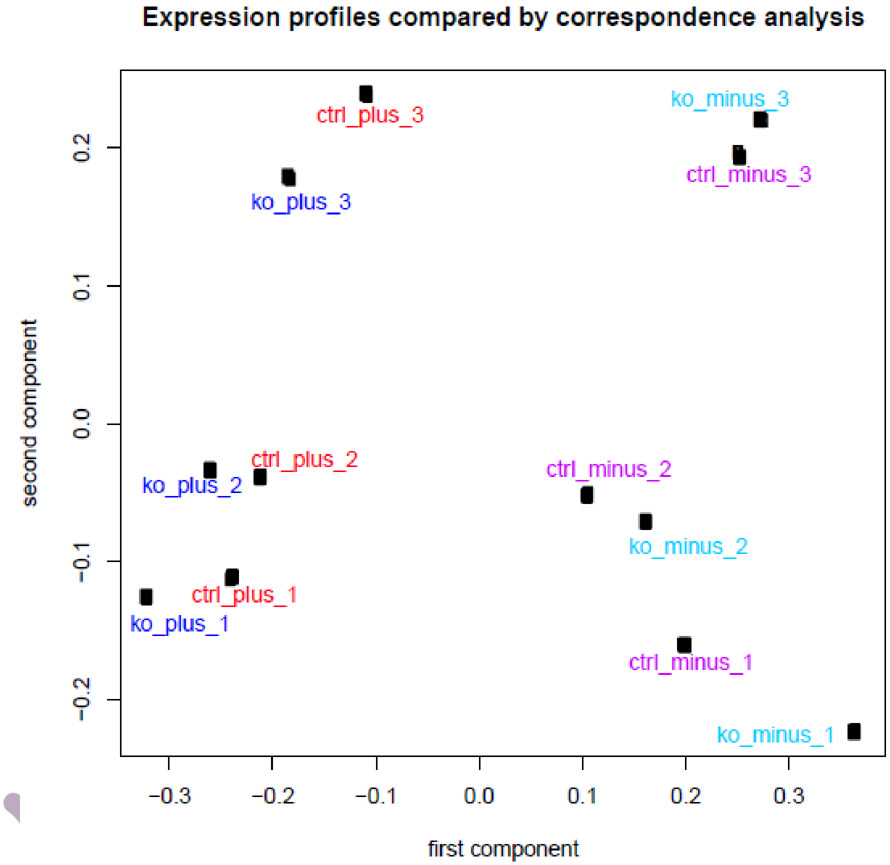
Different Results Due to Version of Software/Package

- DESeq2 v1.16 or 1.18 (compared to older versions)
- Limma
  - Defined DE genes as adj. p-value < 0.05 and FC of at least 2; gene LACTB:

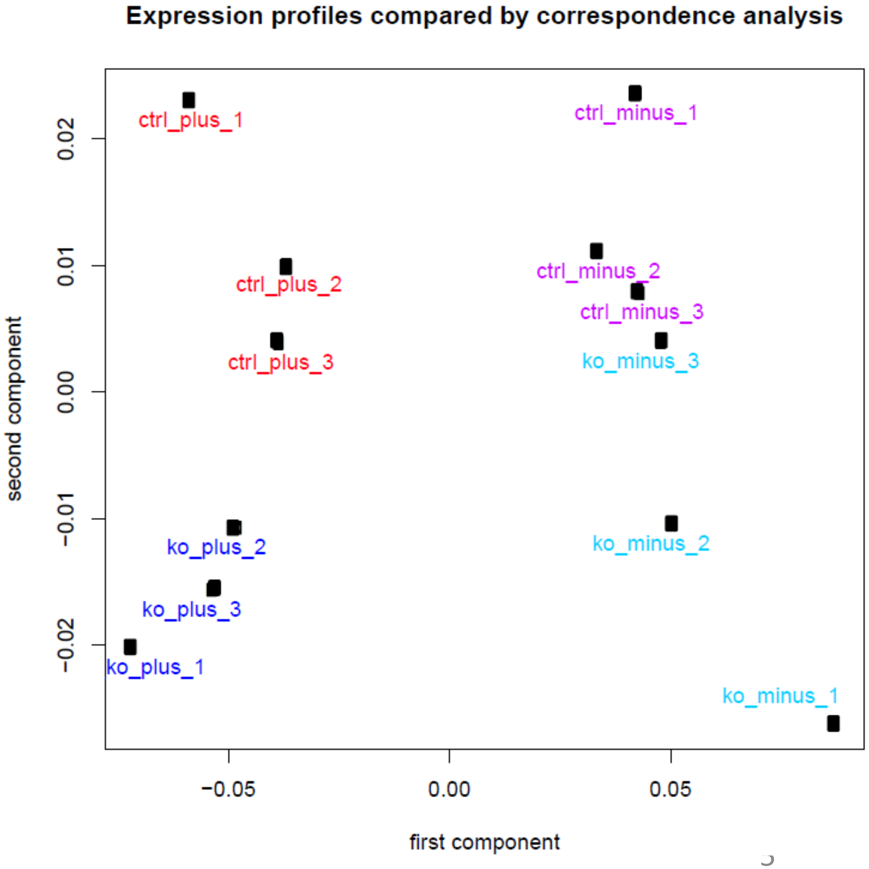
Version	log2 ratio
v3.4.5 (~2010)	1.0038
v? (~2015)	0.9454

QC to check for batch-effects

- Before and after batch-correcting samples from three different batches



Expression profiles compared by correspondence analysis



Expression profiles compared by correspondence analysis

Cell Line Contamination

- Most common: HeLa*

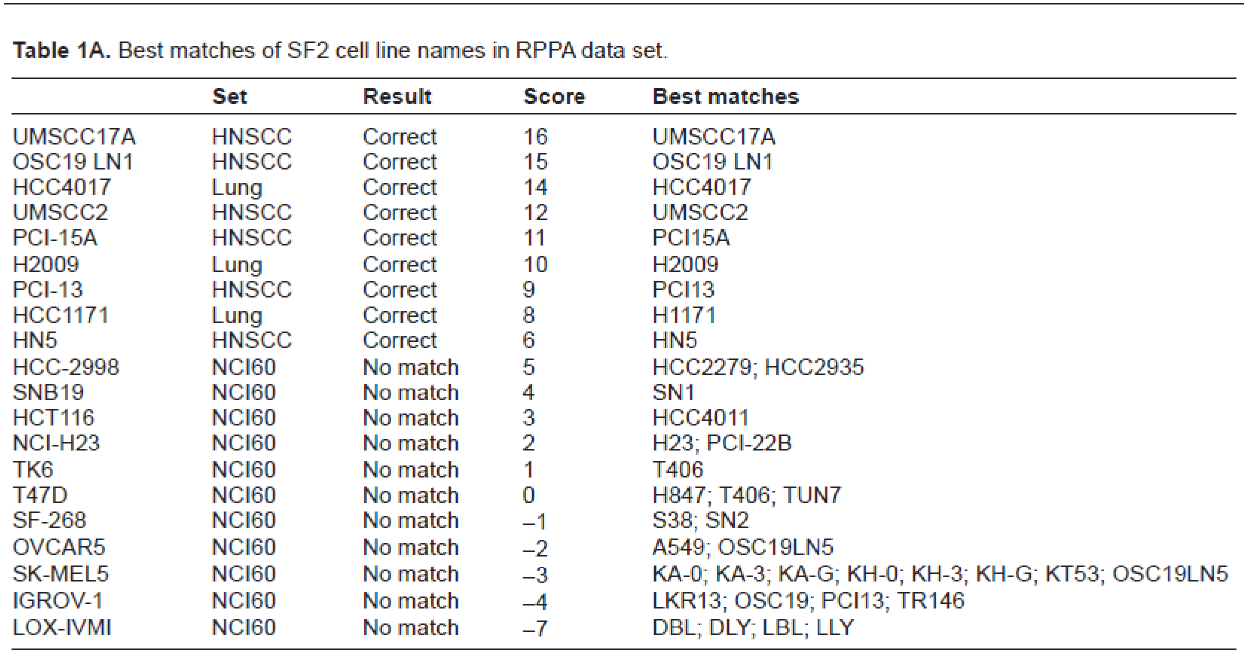


Table 1A. Best matches of SF2 cell line names in RPPA data set.

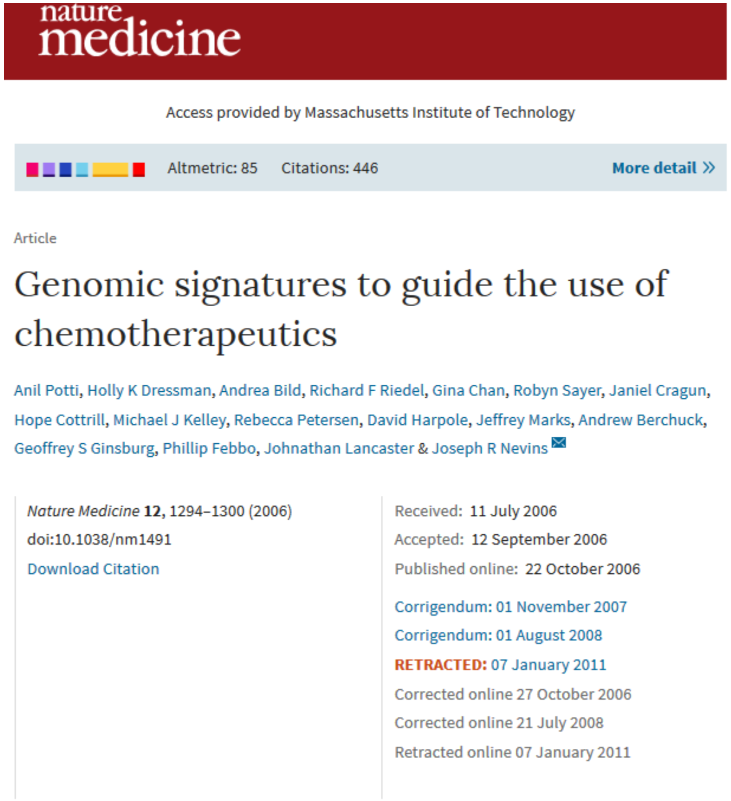
	Set	Result	Score	Best matches
UMSCC17A	HNSCC	Correct	16	UMSCC17A
OSC19 LN1	HNSCC	Correct	15	OSC19 LN1
HCC4017	Lung	Correct	14	HCC4017
UMSCC2	HNSCC	Correct	12	UMSCC2
PCI-15A	HNSCC	Correct	11	PCI15A
H2009	Lung	Correct	10	H2009
PCI-13	HNSCC	Correct	9	PCI13
HCC1171	Lung	Correct	8	H1171
HN5	HNSCC	Correct	6	HN5
HCC-2998	NCI60	No match	5	HCC2279; HCC2935
SNB19	NCI60	No match	4	SN1
HCT116	NCI60	No match	3	HCC4011
NCI-H23	NCI60	No match	2	H23; PCI-22B
TK6	NCI60	No match	1	T406
T47D	NCI60	No match	0	H847; T406; TUN7
SF-268	NCI60	No match	–1	S38; SN2
OVCAR5	NCI60	No match	–2	A549; OSC19LN5
SK-MEL5	NCI60	No match	–3	KA-0; KA-3; KA-G; KH-0; KH-3; KH-G; KT53; OSC19LN5
IGROV-1	NCI60	No match	–4	LKR13; OSC19; PCI13; TR146
LOX-IVMI	NCI60	No match	–7	DBL; DLY; LBL; LLY

Weng, J., et al. *Blasted Cell Line Names* Cancer Informatics (2010)

*http://iclac.org/databases/cross-contaminations/

Forensic Bioinformatics: Case of Potti et al. (2006)

- Aspects of raw data and results are used to infer what methods must have been employed
- Initial claim (in Potti et al.), expression profile in NCI60 to derive signatures of sensitivity to specific drugs, and predict patient response



nature medicine

Access provided by Massachusetts Institute of Technology

Altmetric: 85 Citations: 446 More detail »

Article

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)
doi:10.1038/nm1491
Download Citation

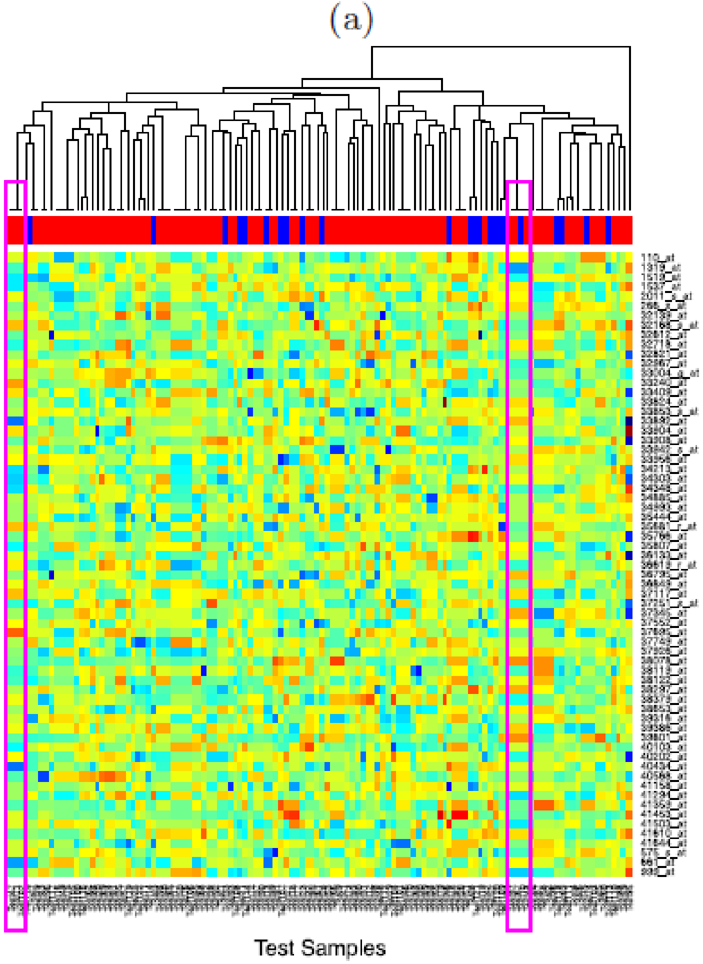
Received: 11 July 2006
Accepted: 12 September 2006
Published online: 22 October 2006

Corrigendum: 01 November 2007
Corrigendum: 01 August 2008
RETRACTED: 07 January 2011
Corrected online 27 October 2006
Corrected online 21 July 2008
Retracted online 07 January 2011

Baggerly, KA., & Coombes, KR *Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in HTP Biology* Annals of Applied Statistics (2009)

Forensic Bioinformatics

- Used an "independent dataset" (GEO), to show patients who were sensitive or resistant but the numbers seemed to be reversed/mixed.
- Heatmap showed sample duplications



Baggerly, KA., & Coombes, KR *Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in HTP Biology* Annals of Applied Statistics (2009)

Forensic Bioinformatics: Conclusion

- Poor documentation hid both in/sensitive label reversal, and the incorrect use of duplicate samples
- Common problems are *simple*
  - Confounding experimental design
  - Mixing up gene labels (off-by-one errors) and group/sample labels
- Incomplete/poor documentation hides the simplicity

Baggerly, KA., & Coombes, KR *Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in HTP Biology* Annals of Applied Statistics (2009)

Forensic Bioinformatics: Recommendations

- Reports to be written using Sweave
- sessionInfo (list libraries and versions)
- Working dir and location of raw data specified
- Check for common types of errors
  - Mostly introduced by separating data and annotation
  - Use numbers/binary and not names (eg. 0 or 1 instead of sensitive or insensitive)

Baggerly, KA., & Coombes, KR *Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in HTP Biology* Annals of Applied Statistics (2009)

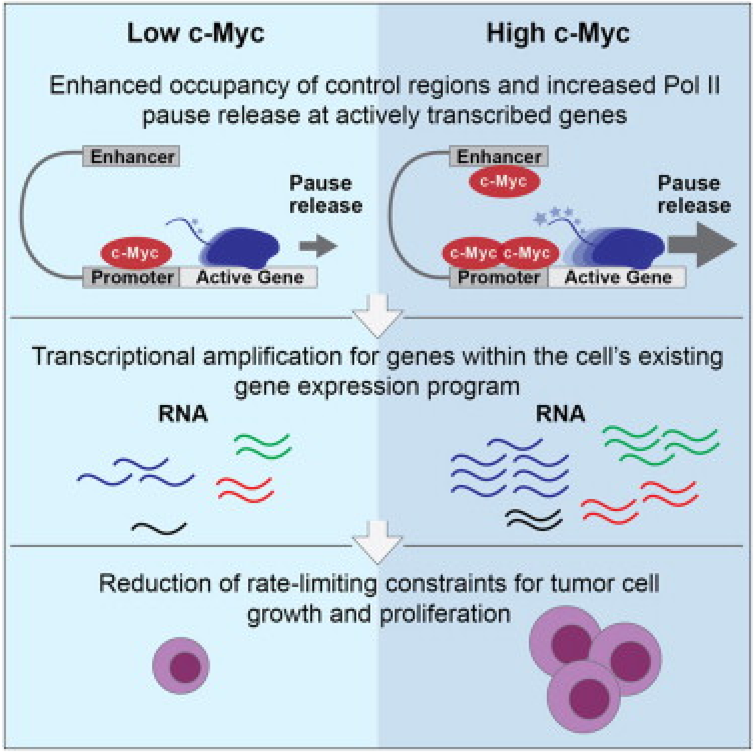
Reproducibility Project: Cancer Biology

- Independently replicating a subset of experiment results from a number of high-profile papers* in cancer biology (2010-2012)
- Registered Reports (Center for Open Science)
  – Specifying in advance which experiments will be done (not mid-course), and number of replicates
- Results from first five Replication Studies were published in Jan 2017
  – Two reproduced important parts of the original papers, one did not. Remaining two were uninterpretable: control tumors grew too quickly or slowly to reliably measure experimental intervention.

https://cos.io/rpcb/
https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology
*50 of the 400 most cited papers from both Scopus and Web of Science

Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc

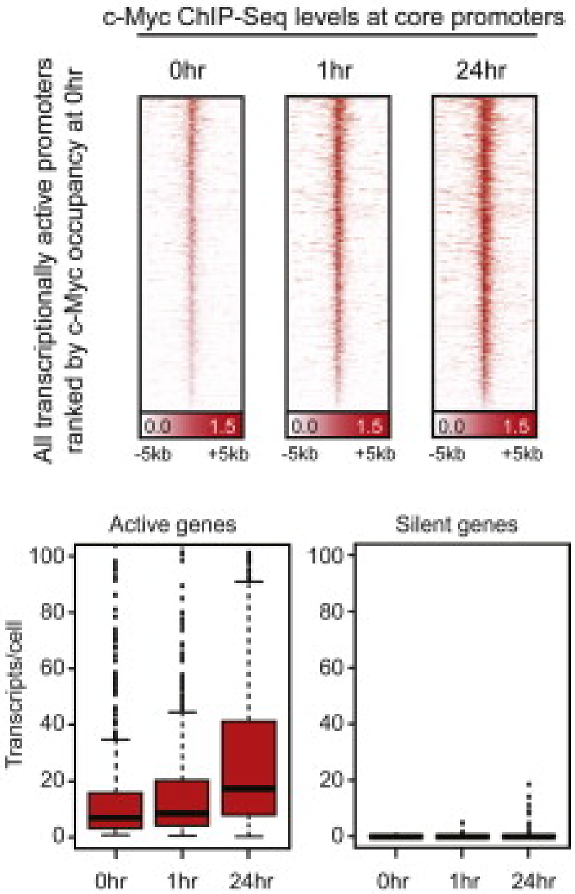
- Lin et al. Cell (2012)



Enhanced occupancy of control regions and increased Pol II pause release at actively transcribed genes

Transcriptional amplification for genes within the cell's existing gene expression program

Reduction of rate-limiting constraints for tumor cell growth and proliferation



Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc

We found overexpression of c-Myc increased total levels of RNA in P493-6 Burkitt's lymphoma cells; **however, while the effect was in the same direction as the original study (Figure 3E; Lin et al., 2012), statistical significance and the size of the effect varied between the original study and the two different lots of serum tested in this replication.** Digital gene expression analysis for a set of genes was also performed on P493-6 cells before and after c-Myc overexpression. Transcripts from genes that were active before c-Myc induction increased in expression following c-Myc overexpression, similar to the original study (Figure 3F; Lin et al., 2012). **Transcripts from genes that were silent before c-Myc induction also increased in expression following c-Myc overexpression, while the original study concluded elevated c-Myc had no effect on silent genes (Figure 3F; Lin et al., 2012)**

https://elifesciences.org/articles/30274 (Jan 2018)

Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc

- Total levels following c-Myc overexpression:
  - Variation in RNA expression during cell culture passage, low yield of the RNA isolation procedure
- Defining silent/active genes
- Statistical analysis (performed with R)
  - Analysis of gene expression using Wilcoxon signed-rank test
  - Statement in Registered Report (2015): Data were checked to ensure assumptions of statistical tests were met (ANOVA)
    - Normality: Shapiro-Wilk test and Q-Q
    - Levene's test to assess homoscedasticity

https://elifesciences.org/articles/30274 (Jan 2018)

Outdated Software Widespread

- GTEx study (Oct 2017) used TopHat v1.4
  - *"The original TopHat program is very far out of date, not just in time, but in performance—it's really been superseded"* – Pacter
  - *"The original analyses of that would have been performed months before that time"* – Kristin Ardlie (GTEx)
    - finalized data in 2014 and made public in 2015
  - In 2017 TopHat still had ~6500 citations
- Original GTEx data used Flux Capacitor for quantification

Scientists Continue to Use Outdated Methods, *The Scientist (Jan 2018)*
https://www.the-scientist.com/?articles.view/articleNo/51260/title/Scientists-Continue-to-Use-Outdated-Methods/

Outdated Annotation/Database used by Popular Tools

- Enrichment analysis using outdated DAVID
  - In 2015, 67% of ~3900 publications (on genomic analysis) cited DAVID
    - last revised in 2010, and only captured ~26% of BP using current resources
  - DAVID was updated shortly after the publication (in 2016)

Wadi L., et al. Impact of outdated gene annotations on pathway enrichment analysis *Nat Methods (2016)*

Which Assembly/Annotation Used in the Study?

- Human: hg18, hg19, hg38
- Mouse: mm8, mm9, mm10
- Ensembl vs RefSeq annotation
  - Version
  - Identifier

Requirements for Publications: STAR methods (Cell Press)

- Structured, Transparent, Accessible Reporting
  - Introduced in 2016
- Quantification and Statistical Analysis
  - statistical test used
  - Exact value of n, and what n represents
  - Definition of center, and dispersion/precision measures
  - How significance was defined
  - Strategies for randomization
  - Inclusion/exclusion of any data or subjects
- Data and Software Availability
  - Datasets must be made free available from the date of publication
  - Submission of data to a public repository
  - Software and data resources should be reported by providing a short description of the software or custom script and URL to obtain them (unless in supp.)

Requirements for Publications: Research Summary (Nature)

- Comparisons of interest are clearly defined
- All statistical methods identified unambiguously
- Data meets all assumptions of tests applied
- Adjustments made for multiple testing is explained
- Any data transformations are clearly described and justified

nature.com/authors/policies/availability.html
nature.com/authors/policies/ReportingSummary.pdf

Ideal Journal for Reproducibility? Biostatistics

Our reproducible research policy is for papers in the journal to be kite-marked “D” if the data on which they are based are freely available, “C” if the authors’ code is freely available, and R if both data and code are available***, and our Associate Editor for Reproducibility is able to use these to reproduce the results in the paper***. Data and code are published electronically on the journal’s website as Supplementary Materials.

Code Availability

Authors are strongly encouraged to submit code supporting their publications. Authors should submit a link to a Github repository and to a specific example of the code on a code archiving service such as Figshare or Zenodo.

Defining Reproducible Research

- *Ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator* - NSF
- Reproducibility, at minimum, requires:
  - analytical data sets (original/raw or processed)
  - relevant metadata
  - analytical code
  - related software
- To what extent deviations are acceptable in reproducibility?

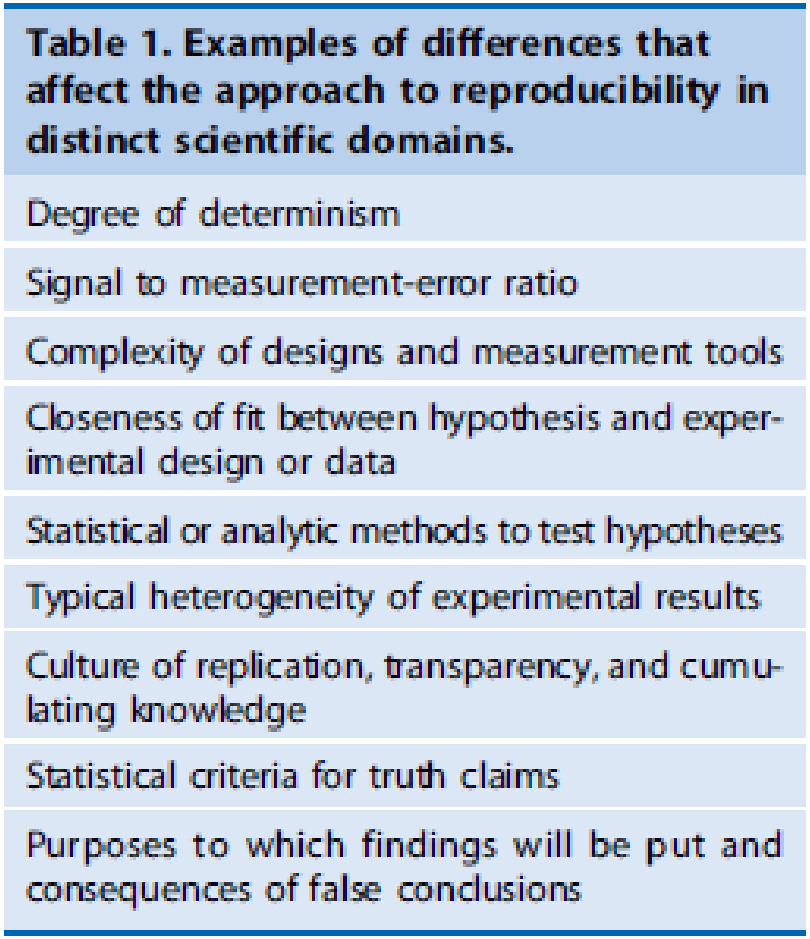


Table 1. Examples of differences that affect the approach to reproducibility in distinct scientific domains.
Degree of determinism
Signal to measurement-error ratio
Complexity of designs and measurement tools
Closeness of fit between hypothesis and experimental design or data
Statistical or analytic methods to test hypotheses
Typical heterogeneity of experimental results
Culture of replication, transparency, and cumulating knowledge
Statistical criteria for truth claims
Purposes to which findings will be put and consequences of false conclusions

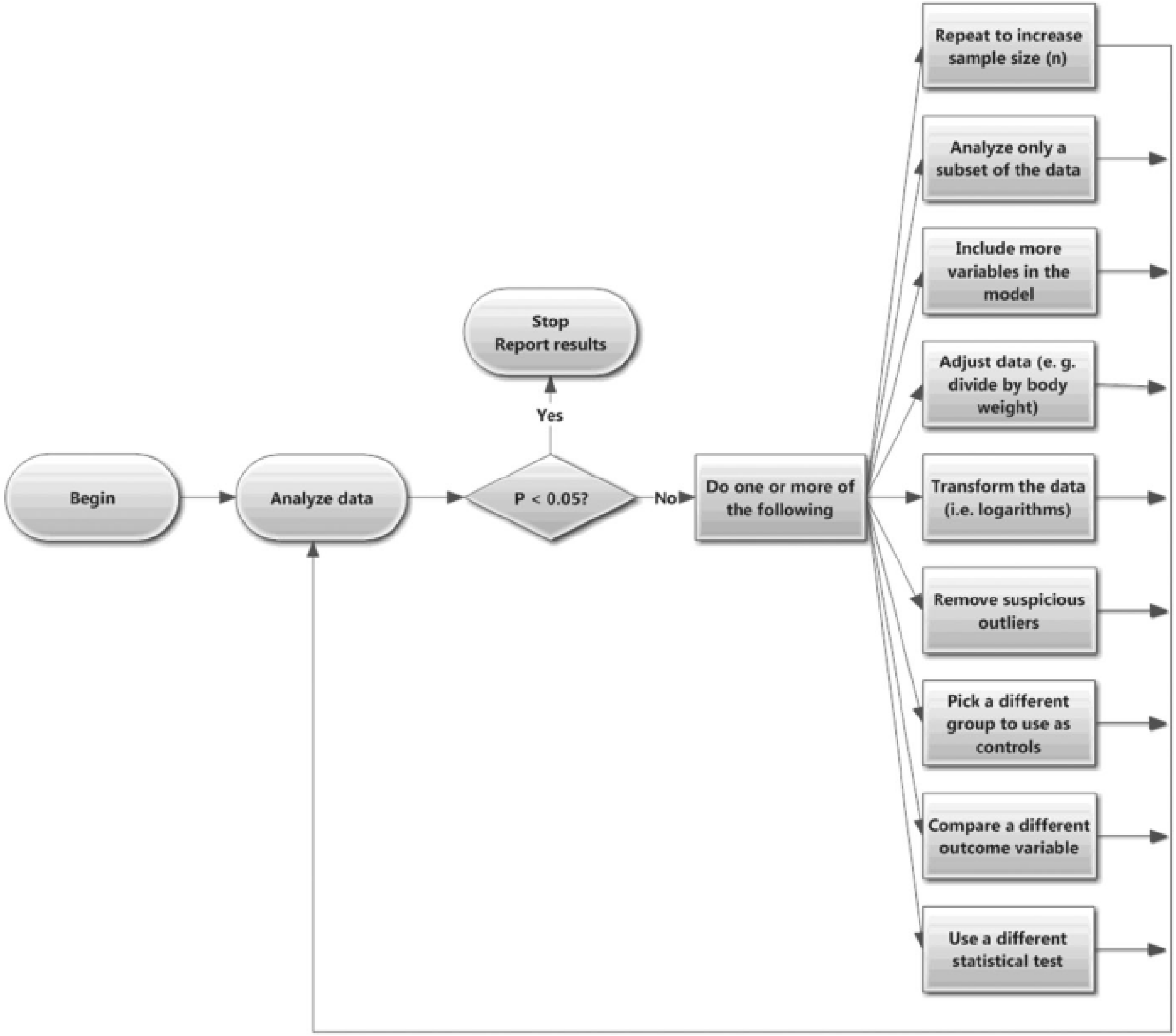
Goodman, S.N., et al. *What does research reproducibility mean?* Sci Translational Med (2016)

Practices That Leads to Irreproducible Research

- Multiple comparisons
  - Silent multiplicity
- File-drawer problem
  - Selective outcome reporting
- Pseudoreplication
- Data mining/dredging/torturing
  - Data snooping
- Hypothesizing after the results are known (HARKing)
- Significance questing
  - Specification searching
  - P-hacking

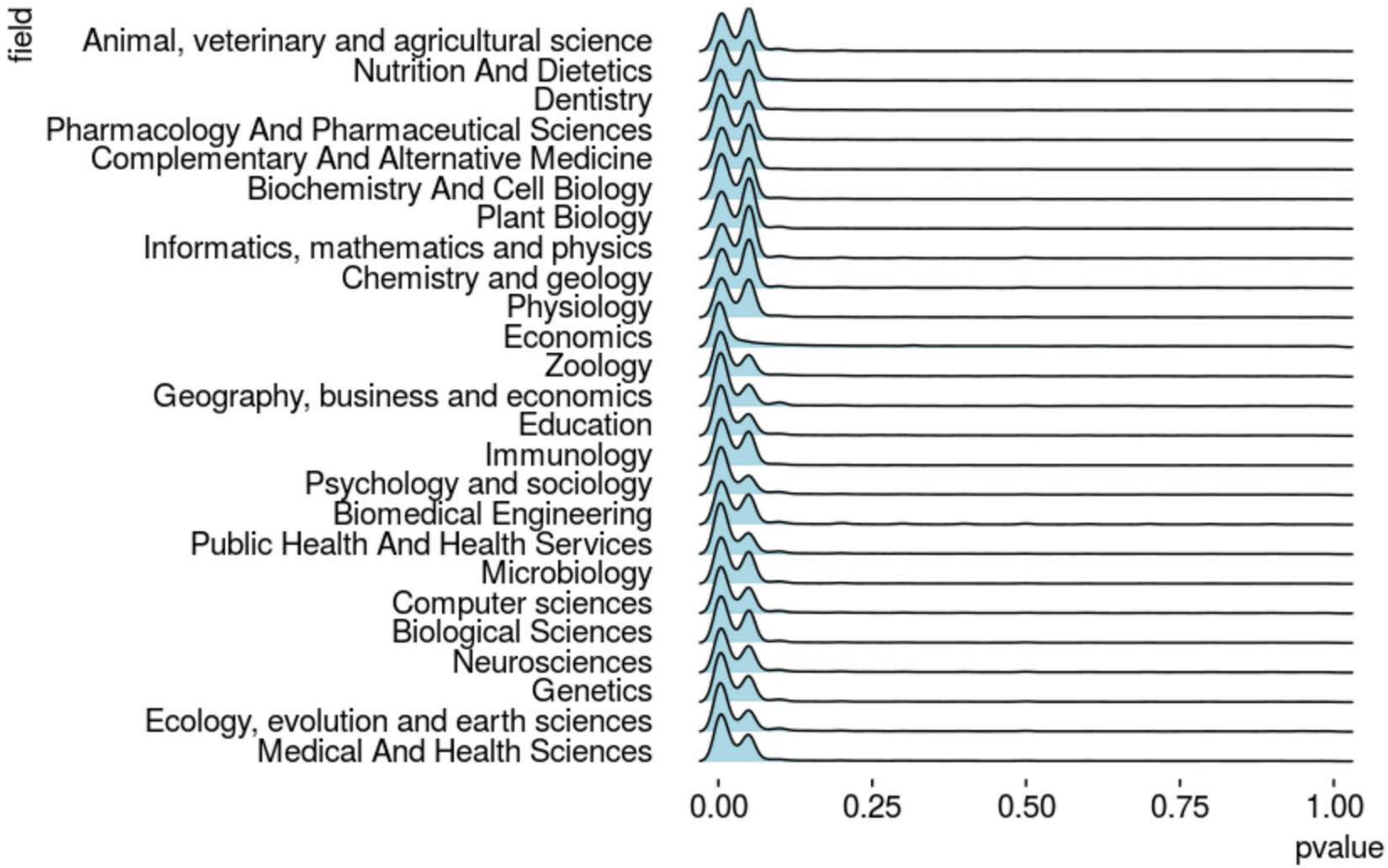
Goodman, S.N., et al *What does research reproducibility mean?* Sci Translational Med (2016)

P-Hacking Flow Chart


Begin
Analyze data
P < 0.05?
Yes
Stop
Report results
No
Do one or more of the following
Repeat to increase sample size (n)
Analyze only a subset of the data
Include more variables in the model
Adjust data (e. g. divide by body weight)
Transform the data (i.e. logarithms)
Remove suspicious outliers
Pick a different group to use as controls
Compare a different outcome variable
Use a different statistical test

Motulsky, H.J. *Common misconceptions about data analysis and statistics* BJP (2015)

P-Values Across Fields

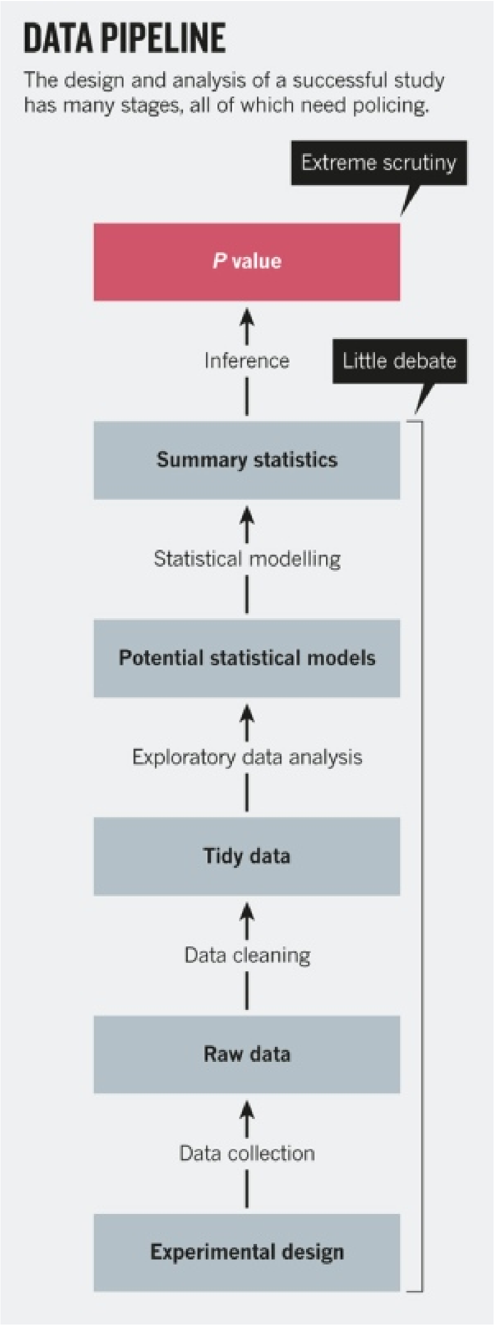


https://simplystatistics.org/2017/07/26/announcing-the-tidypvals-package/

AMSTAT Statement on Statistical Significance and P-Values

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

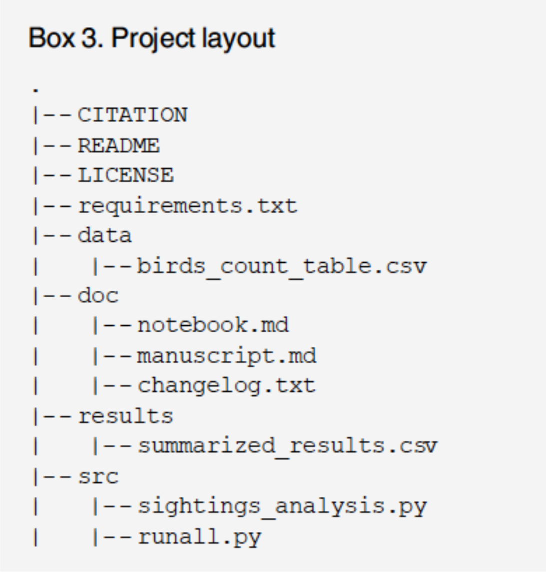
P-value is not the only issue!



Leek, J.T., and Peng, R.D. *P-values are just the tip of the iceberg* Nat. (2015)

Best Practices

- Data Management
  - saving both raw and intermediate forms
  - documenting all steps
  - creating analysis-friendly “tidy” data, no PDFs, etc.
- Software
  - modular and more functions
  - use well-maintained libraries/packages, test them before using
- Collaboration
  - create shared “to-do” list
  - communication strategies
- Project Organization
  - naming of project and directory
    - *src* or *bin* for scripts
    - *results* dir for output



Box 3. Project layout

```
.
|--CITATION
|--README
|--LICENSE
|--requirements.txt
|--data
|    |--birds_count_table.csv
|--doc
|    |--notebook.md
|    |--manuscript.md
|    |--changelog.txt
|--results
|    |--summarized_results.csv
|--src
|    |--sightings_analysis.py
|    |--runall.py
```

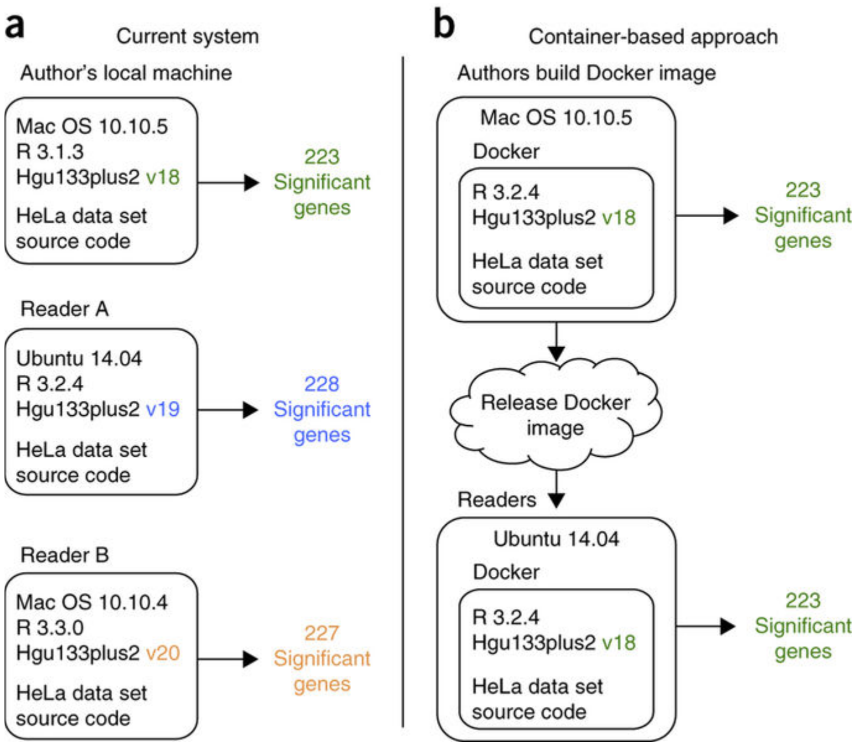
Wilson, G., et al. *Good Enough Practices in Scientific Computing* PLOS Comp Bio (2017)

10 Simples Rules For Reproducible Research

- For Every Result, Keep Track of How It Was Produced
- Avoid Manual Data Manipulation Steps
- Archive the Exact Versions of All External Programs Used
- Version Control All Custom Scripts
- Record All Intermediate Results, When Possible in Standardized Formats
- For Analyses That Include Randomness, Note Underlying Random Seeds
- Always Store Raw Data behind Plots
- Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Connect Textual Statements to Underlying Results
- Provide Public Access to Scripts, Runs, and Results

Sandve, G.K., et al. *Ten Simple Rules for Reproducible Computational Research* PLOS Comp Bio (2013)

Docker and Workflows



a Current system

Author's local machine

Mac OS 10.10.5
R 3.1.3
Hgu133plus2 v18
HeLa data set
source code

223 Significant genes

Reader A

Ubuntu 14.04
R 3.2.4
Hgu133plus2 v19
HeLa data set
source code

228 Significant genes

Reader B

Mac OS 10.10.4
R 3.3.0
Hgu133plus2 v20
HeLa data set
source code

227 Significant genes

b Container-based approach

Authors build Docker image

Mac OS 10.10.5
Docker
R 3.2.4
Hgu133plus2 v18
HeLa data set
source code

223 Significant genes

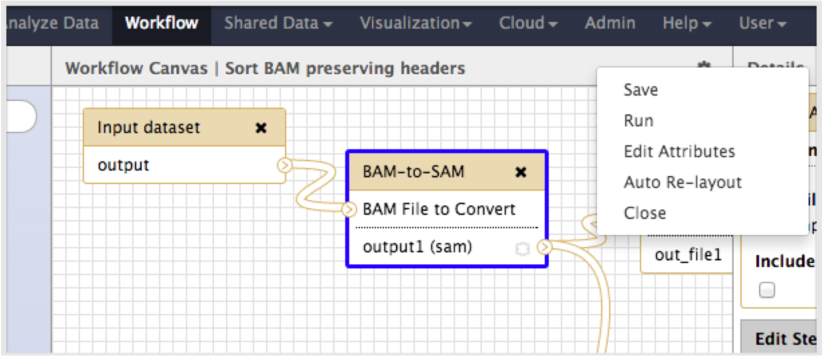
Release Docker image

Readers

Ubuntu 14.04
Docker
R 3.2.4
Hgu133plus2 v18
HeLa data set
source code

223 Significant genes

Beaulieu-Jones, B.K and Greene C.S., *Reproducibility of computational workflows is automated using continuous analysis* Nat Biotech (2017)



galaxyproject.org

Questions

- Have you run into issues replicating a method? Encountered p-hacking?
  - What happens when the scientist didn't do replicates?
- How do you organize your projects?
  - Do you make an "old" folder when you need to re-run?
  - Do you always keep a README or similar file?
- Which version of software do you use, the most recent/latest?
  - How do you keep track? eg. module load star/2.5.2a
  - Do you keep updated with releases/changes made?
- Do you always check for (potential) batch-effects?
- How often do you update annotations/databases?
- Do you use R Markdown or Jupyter Notebook? Workflows?
- How do you share scripts/data with i) scientists? ii) community/public?
  - Who maintains the software once the scientist leaves?
- How to do you ensure consistency in the analysis within your group? SOPs?