Reproducible research in data science - a bioinformatics primer

Mikhail Dozmorov, Ph.D.
Department of Biostatistics, VCU
mikhail.dozmorov@vcuhealth.org
01/29/2016

Reproducible research in science

- Science is the systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into testable laws and theories.

- The success and credibility of science are anchored in the willingness of scientists to expose their ideas and results to independent testing and replication by other scientists.

http://www.aps.org/policy/statements/99_6.cfm

What is reproducible research?

- Reproducibility
- Replicability
- Repeatability
- Reliability
- Robustness
- Generalizability

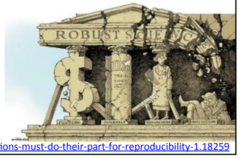
Steve Goodman, Stanford, March 18, 2015

What is reproducible research?

Reproducible research is the ultimate standard for strengthening scientific evidence by independent:

- Investigators
- Data
- Analytical methods
- Laboratories
- Instruments

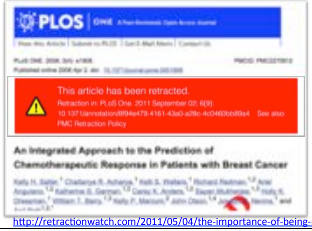
Replication is particularly important in studies that can impact broad policy or regulatory decisions



http://www.nature.com/news/robust-research-institutions-must-do-their-part-for-reproducibility-1.18259

Why do we care?

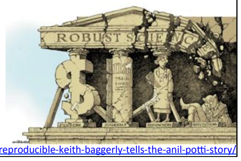
- High-throughput data-generating technologies are increasingly used to make clinical recommendations and treatment decisions
- A problem may be overlooked .. Published .. Get in clinical trials


PLOS ONE

This article has been retracted.

Retraction in: PLoS One. 2011 September 02; 6(9): 10.1371/annotation/8f94e479-4161-43a0-a28c-4c0460bb89a4 See also: PMC Retraction Policy

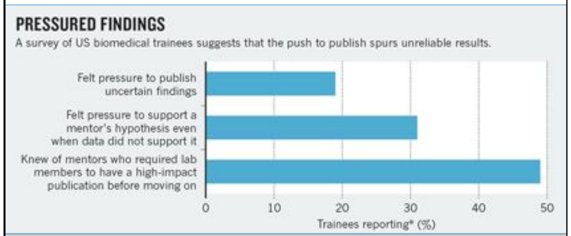
An Integrated Approach to the Prediction of Chemotherapeutic Response in Patients with Breast Cancer



http://retractionwatch.com/2011/05/04/the-importance-of-being-reproducible-keith-baggerly-tells-the-anil-potti-story/

Why reproducible research is questioned now?

- Publish or perish


PRESSURED FINDINGS

A survey of US biomedical trainees suggests that the push to publish spurs unreliable results.

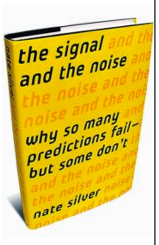
http://www.nature.com/news/robust-research-institutions-must-do-their-part-for-reproducibility-1.18259

Why irreproducibility problems arise?

- Humans are good at recognizing patterns

Human beings do not have very many natural defenses. We are not all that fast, and we are not all that strong. We do not have claws or fangs or body armor. We cannot spit venom. We cannot camouflage ourselves. And we cannot fly. Instead, we survive by means of our wits. Our minds are quick. **We are wired to detect patterns** and respond to opportunities and threats without much hesitation.

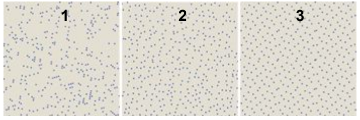
- Nate Silver

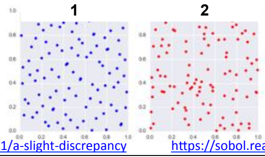


http://www.amazon.com/gp/product/0143125087?redirect=true&ref_=s9_simh_gw_g14_i1_r

Why irreproducibility problems arise?

- Humans are good at recognizing patterns

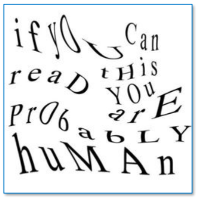




http://bit-player.org/2011/a-slight-discrepancy https://sobol.readthedocs.org/en/latest/

Why irreproducibility problems arise?

- Humans are good at recognizing patterns



http://neuralnetworksanddeeplearning.com/chap1.html
http://www.npr.org/templates/story/story.php?storyId=130594039

Why irreproducibility problems arise?

- Humans are good at recognizing patterns

... improved technology did not cover for the lack of theoretical understanding about the economy; it only gave economists faster and more elaborate ways to mistake noise for a signal.

- Nate Silver

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

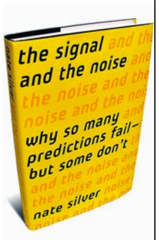
There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have

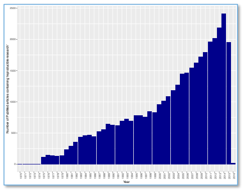
is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may

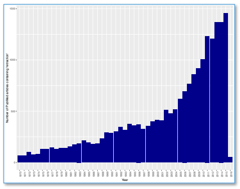


http://www.amazon.com/gp/product/0143125087?redirect=true&ref_=s9_simh_gw_g14_i1_r
http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

PubMed on "Reproducible research" vs. "Retraction"

"Reproducible research" "Retraction"



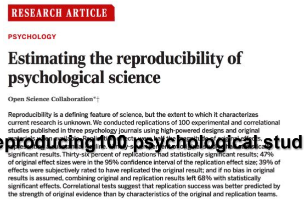


Number of publications per year, from 1970 to January 2016

Retraction Watch Tracking retractions as a window into the scientific process

http://retractionwatch.com/

Reproducibility initiatives



RESEARCH ARTICLE

PSYCHOLOGY

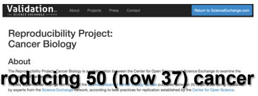
Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Reproducing100 psychological studies

http://science.sciencemag.org/content/sci/349/6251/aac4716.full.pdf



Validation About Projects Press Contact Return to ScienceExchange.com

Reproducibility Project: Cancer Biology

About

The Reproducibility Project: Cancer Biology is a collaboration between the Center for Open Science and Science Exchange to examine the [illegible] The key experimental findings [illegible] the [illegible] cancer biology studies published between 2010-2012 are being replicated by experts from the Science Exchange network, according to best practices for replication established by the Center for Open Science.

Reproducing 50 (now 37) cancer studies

http://validation.scienceexchange.com/#/cancer-biology

Reproducibility initiatives

nature.com

Enhancing reproducibility

New reporting standards for Nature journal authors are intended to improve transparency and reproducibility.

Difficulty in reproducing published biomedical research studies has become a matter of increasing concern that, if unaddressed, will waste limited research funding and may erode public support for research. *Nature Methods* is therefore adopting new editorial measures in an attempt to improve the consistency and quality of reporting in submitted manuscripts.

We join *Nature* and the other Nature research journals ure legend for readers interested in seeing them firsthand. We also continue to encourage authors to use resources for sharing detailed methods and reagent descriptions by providing direct online linking between primary research articles and Protocol Exchange (http://www.nature.com/protocolexchange/), an open resource into which authors can deposit the detailed step-by-step experimental protocols used in their study.

https://www.ncbi.nlm.nih.gov/pubmed/23762900

NEWS

REPRODUCIBLE RESEARCH

ADDRESSING THE NEED FOR DATA AND CODE SHARING IN COMPUTATIONAL SCIENCE

By the Yale Law School Roundtable on Data and Code Sharing

https://web.stanford.edu/~vcs/papers/RoundtableDeclaration2010.pdf

NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring shorter term, however, the checks and balances that once ensured scientific fidelity have been hobbled. This has compromised

https://www.ncbi.nlm.nih.gov/pubmed/24482835

Reproducibility initiatives

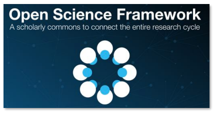
https://cos.io/

Order experiments from the world's best labs

Search for experiments, labs, etc...

Popular services: RNA Sequencing, Immunohistochemistry, Molecular Cloning, DNA Sequencing, Next Generation Sequencing, Western Blot

https://www.scienceexchange.com/



Open Science Framework

A scholarly commons to connect the entire research cycle

https://osf.io/

WHAT CAN WE DO TO ENHANCE REPRODUCIBILITY?

Steps in reproducible research

The most important tool is the mindset, when starting, that the end product will be reproducible.
– Keith Baggerly

- Can my conclusions be reproduced by others?
- Data availability
- Detailed Methods description
- Software availability

- Empirical reproducibility
- Computational reproducibility
- Statistical reproducibility

http://kbroman.org/Tools4RR/

Common approach: write report around results

Point and click approach

- Use MS Excel for data entry/cleaning/preparation, and possibly statistical analysis;
- Copy/paste to/from other programs.

Zeeberg BR et al. **Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics** *BMC Bioinformatics*. 2004
https://www.ncbi.nlm.nih.gov/pubmed/15214961

Problems

- With point-and-click, there's no way to record/save the steps that generated the results;
- Data files are kept separately from the analysis code, and from reports;
- After modifications of one of the files involved, it becomes unclear which version corresponds exactly to the reported results;
- Every time something changes, you have to regenerate the figures/results/reports by hand – very time consuming.

Better approach: write report that generates results

- Everything automated via code;
- Most raw data is attached to the code;
- Any changes in code should be version controlled;
- The full report should be self-sufficient and reproducible with a single command.

EVERYTHING AUTOMATED

Automating everything, or, why we love R

- R is an open source programming language – removes error-prone point-and-click;
- R is free to run, study, change and improve;
- R runs on Windows, Mac, and Linux;
- R has publication quality graphing capabilities;
- Extensible with a very large collection of actively developing packages;
- Excellent report-creating capabilities.

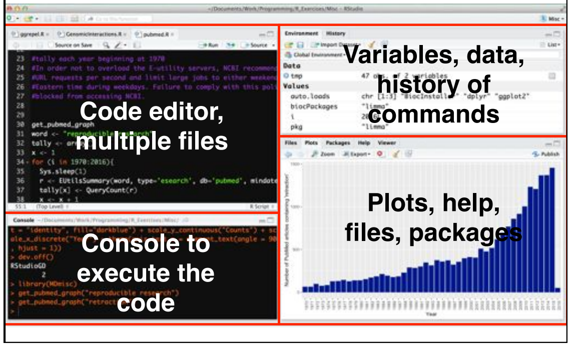
http://rpubs.com/ideltoro/147317

The New York Times

"**R is really important to the point that it's hard to overvalue it**," said Daryl Pregibon, a research scientist at Google, which uses the software widely. "**It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems.**"

http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html

R is reimagined with RStudio



```
#tally each year beginning at 1970
#In order not to overload the E-utility servers, NCBI recommend
#URL requests per second and limit large jobs to either weekend
#Eastern time during weekdays. Failure to comply with this poli
#blocked from accessing NCBI.


get_pubmed_graph
word <- "reproducible research"
tally <- ar
x <- 1
for (i in 1970:2016){
  Sys.sleep(1)
  r <- EUtilsSummary(word, type='esearch', db='pubmed', mindate
  tally[x] <- QueryCount(r)
  x <- x + 1
```

```
Data
tmp             47 obs. of 2 variables
Values
auto.loads      chr [1:3] "BiocInstaller" "dplyr" "ggplot2"
biocPackages    "limma"
i               2016
pkg             "limma"
```

```
t = "identity", fill="darkblue") + scale_y_continuous("Counts") + sc
ale_x_discrete("Ye
, hjust = 1))
> dev.off()
RStudioGD
        2
> library(MDmisc)
> get_pubmed_graph("reproducible research")
> get_pubmed_graph("retract
>
```

Code editor, multiple files

Variables, data, history of commands

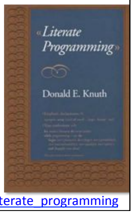
Console to execute the code

Plots, help, files, packages

CODE ATTACHED TO THE REPORT

Literate programming, or self-documenting code

- A report containing a stream of text and code chunks;
- Each code chunk loads data, computes results, shows figures;
- Each text chunk explains how the code chunks work;
- The resulting report is human- and machine readable.



https://en.wikipedia.org/wiki/Literate_programming

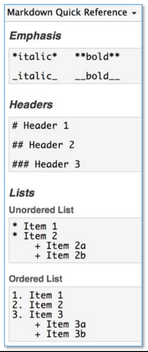
Evolution of literate programming

- **LaTeX** – document markup language;
- **Sweave** – integrates LaTeX markup and R code formatting;
- **R markdown (knitr)** – most natural way of combining text and code.

LaTeX	R Markdown
\textbf{word}	**word**

http://arxiv.org/abs/1402.1894

Markdown basics



Markdown Quick Reference ▾

Emphasis

```
*italic*   **bold**
_italic_   __bold__
```

Headers

```
# Header 1
## Header 2
### Header 3
```

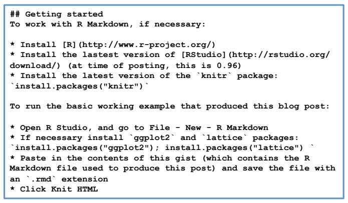
Lists

Unordered List

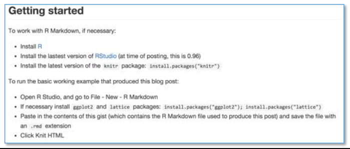
```
* Item 1
* Item 2
    + Item 2a
    + Item 2b
```

Ordered List

```
1. Item 1
2. Item 2
3. Item 3
    + Item 3a
    + Item 3b
```



```
## Getting started
To work with R Markdown, if necessary:

* Install [R](http://www.r-project.org/)
* Install the lastest version of [RStudio](http://rstudio.org/
download/) (at time of posting, this is 0.96)
* Install the latest version of the `knitr` package:
`install.packages("knitr")`

To run the basic working example that produced this blog post:

* Open R Studio, and go to File - New - R Markdown
* If necessary install `ggplot2` and `lattice` packages:
`install.packages("ggplot2"); install.packages("lattice") `
* Paste in the contents of this gist (which contains the R
Markdown file used to produce this post) and save the file with
an `.rmd` extension
* Click Knit HTML
```



Getting started

To work with R Markdown, if necessary:

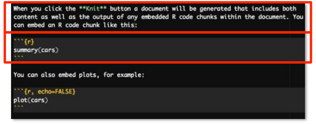
- Install R
- Install the lastest version of RStudio (at time of posting, this is 0.96)
- Install the latest version of the `knitr` package: `install.packages("knitr")`

To run the basic working example that produced this blog post:

- Open R Studio, and go to File - New - R Markdown
- If necessary install `ggplot2` and `lattice` packages: `install.packages("ggplot2"); install.packages("lattice")`
- Paste in the contents of this gist (which contains the R Markdown file used to produce this post) and save the file with an `.rmd` extension
- Click Knit HTML

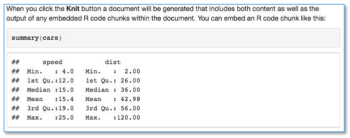
Literate programming with knitr

- Mix markdown with code

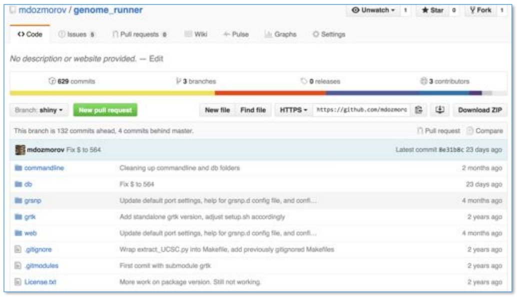


~~~
When you click the **Knit** button a document will be generated that includes both
content as well as the output of any embedded R code chunks within the document. You
can embed an R code chunk like this:

```{r}
summary(cars)
~~~

You can also embed plots, for example:

```{r, echo=FALSE}
plot(cars)
```
```

- And "knit" a report with one button



When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

VERSION CONTROL OF CHANGES

Why we need version control?

"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.corrections9.MORE.30.doc
FINAL_rev.22.comments49.corrections.10.#@$%WHYDIDICOMETOGRADSCHOOL????.doc

http://www.phdcomics.com/comics/archive.php?comicid=1531

Version control – what and when did you do

- **Git** and GitHub – version control system;
- Each project stored in its own repository;
- History of changes – track what you did;
- Ability to go back if something breaks;
- Branch out, go creative, then merge or revert the changes;
- Collaborate through merging changes from multiple people.

https://github.com/

Git basics

- Git is a command line tool;
- GitHub is a web-based storage for your project repositories;
- `Git add` – add a file to version control system;
- `Git commit` – make a snapshot of current changes;
- `Git push/pull` – send/get changes to/from GitHub.

https://github.com/
```

# Repository on GitHub



mdozmorov / genome_runner

No description or website provided. — Edit

629 commits | 3 branches | 0 releases | 3 contributors

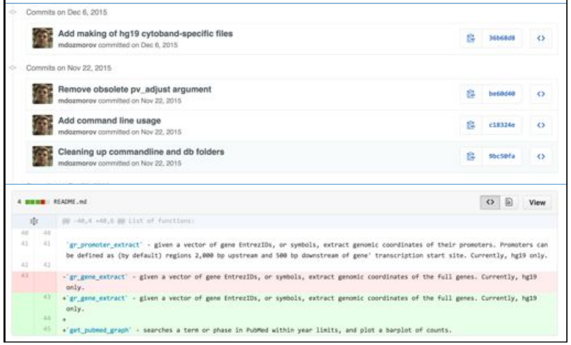
https://github.com/mdozmorov

This branch is 132 commits ahead, 4 commits behind master.

mdozmorov Fix $ to 564 — Latest commit 8e31b8c 23 days ago

| | | |
|---|---|---|
| commandline | Cleaning up commandline and db folders | 2 months ago |
| db | Fix $ to 564 | 23 days ago |
| grsnp | Update default port settings, help for grsnp.d config file, and confi… | 4 months ago |
| grtk | Add standalone grtk version, adjust setup.sh accordingly | 2 years ago |
| web | Update default port settings, help for grsnp.d config file, and confi… | 4 months ago |
| .gitignore | Wrap extract_UCSC.py into Makefile, add previously gitignored Makefiles | 2 years ago |
| .gitmodules | First comit with submodule grtk | 2 years ago |
| License.txt | More work on package version. Still not working. | 2 years ago |

# History of commits



Commits on Dec 6, 2015

Add making of hg19 cytoband-specific files
mdozmorov committed on Dec 6, 2015 — 36b68d8

Commits on Nov 22, 2015

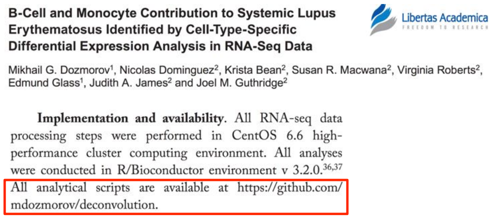
Remove obsolete pv_adjust argument
mdozmorov committed on Nov 22, 2015 — be60d40

Add command line usage
mdozmorov committed on Nov 22, 2015 — c18324e

Cleaning up commandline and db folders
mdozmorov committed on Nov 22, 2015 — 9bc50fa

```
4 README.md                                                         View
@@ -40,4 +40,6 @@ List of functions:

 `gr_promoter_extract` - given a vector of gene EntrezIDs, or symbols, extract genomic coordinates of their promoters. Promoters can
 be defined as (by default) regions 2,000 bp upstream and 500 bp downstream of gene' transcription start site. Currently, hg19 only.

-`gr_gene_extract` - given a vector of gene EntrezIDs, or symbols, extract genomic coordinates of the full genes. Currently, hg19
 only.
+`gr_gene_extract` - given a vector of gene EntrezIDs, or symbols, extract genomic coordinates of the full genes. Currently, hg19
 only.
+
+`get_pubmed_graph` - searches a term or phase in PubMed within year limits, and plot a barplot of counts.
```

Publishing with Git



B-Cell and Monocyte Contribution to Systemic Lupus Erythematosus Identified by Cell-Type-Specific Differential Expression Analysis in RNA-Seq Data

Libertas Academica

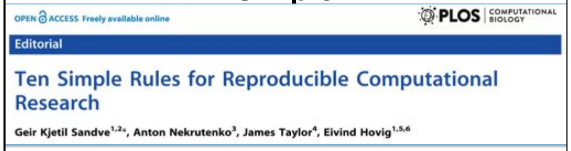
Mikhail G. Dozmorov[1], Nicolas Dominguez[2], Krista Bean[2], Susan R. Macwana[2], Virginia Roberts[2], Edmund Glass[1], Judith A. James[2] and Joel M. Guthridge[2]

Implementation and availability. All RNA-seq data processing steps were performed in CentOS 6.6 high-performance cluster computing environment. All analyses were conducted in R/Bioconductor environment v 3.2.0.[36,37] All analytical scripts are available at https://github.com/mdozmorov/deconvolution.

https://www.ncbi.nlm.nih.gov/pubmed/26512198

LEARN MORE

Reproducible research made simple



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2*], Anton Nekrutenko[3], James Taylor[4], Eivind Hovig[1,5,6]

https://www.ncbi.nlm.nih.gov/pubmed/24204232

Best Practices for Scientific Computing

Greg Wilson [*], D.A. Aruliah [†], C. Titus Brown [‡], Neil P. Chue Hong [§], Matt Davis [¶], Richard T. Guy [‖], Steven H.D. Haddock [**], Katy Huff [††], Ian M. Mitchell [‡‡], Mark D. Plumbley [§§], Ben Waugh [¶¶], Ethan P. White [***], Paul Wilson [†††]

https://www.ncbi.nlm.nih.gov/pubmed/24415924

Reproducible research in data science - a bioinformatics primer

Thank you

https://github.com/mdozmorov/presentations

Mikhail Dozmorov, Ph.D.
Department of Biostatistics, VCU
mikhail.dozmorov@vcuhealth.org
01/29/2016