Is your work reproducible?

Best Practices When Working With Computers

Stefan Wyder & Önder Kartal

September 2017

Universität Zürich[UZH]

URPP Evolution in Action

Reproducible Research

“Research is reproducible if it can be reproduced by others”

One of the main principles of the scientific method

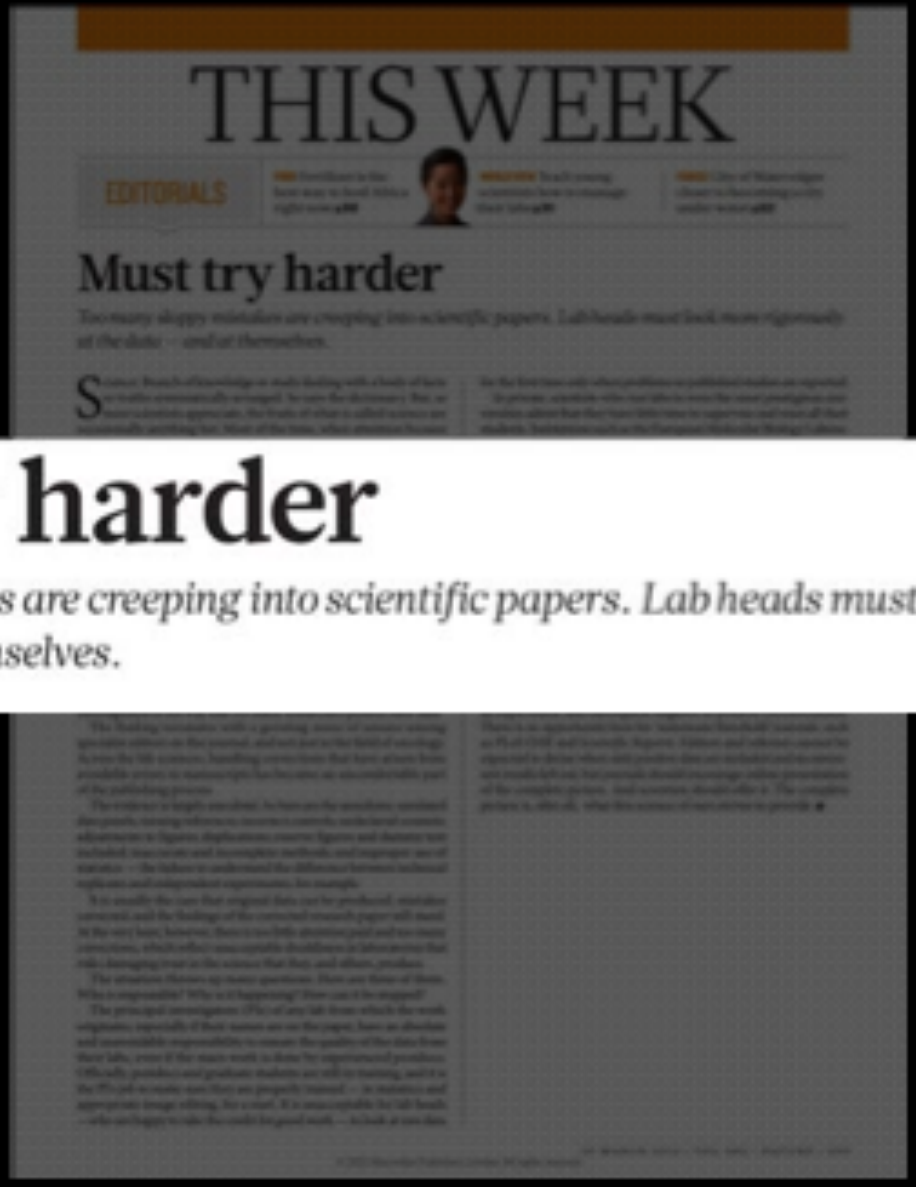
Definition of Reproducible Research

A complete description of the data and the analysis of that data — including computer programs — so the results can be exactly reproduced by others.

Amstat News, 1 January 2011

Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.

William Stafford Noble


THIS WEEK

EDITORIALS

Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

Nature 483, 509 (29 March 2012)

Forensic Bioinformatics

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

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

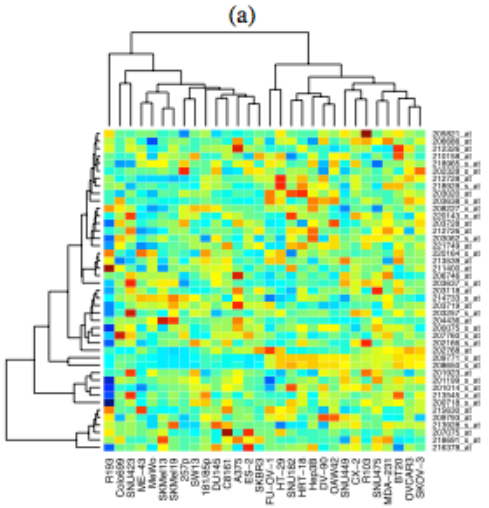
BY KEITH A. BAGGERLY[1] AND KEVIN R. COOMBES[2]

University of Texas

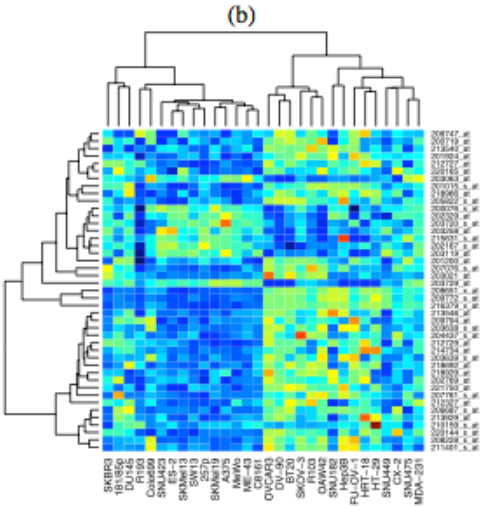
High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

Keith Baggerly, Ph.D.

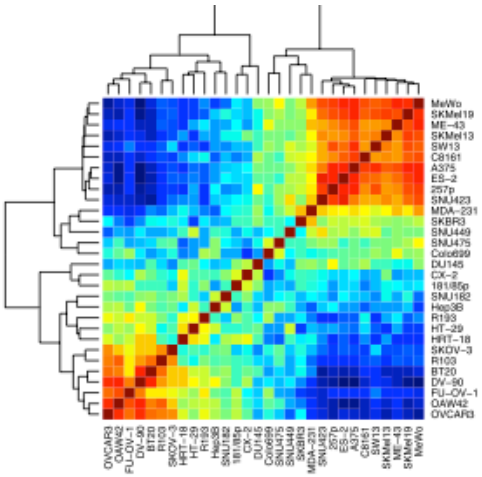
Reconstructing heatmap for cisplatin signature



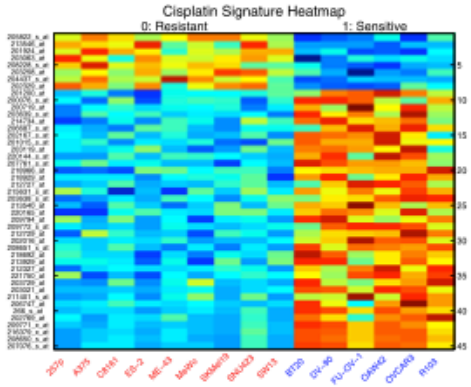
orig data: no structure



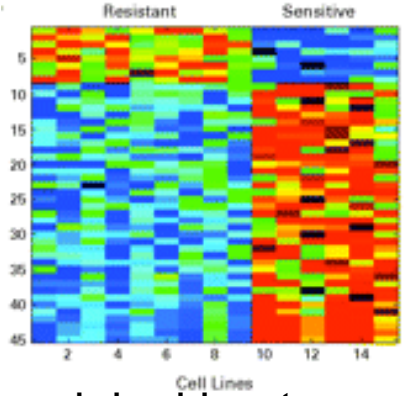
offseting by one
(indexing error)



pairwise sample correlations
to detect label switches



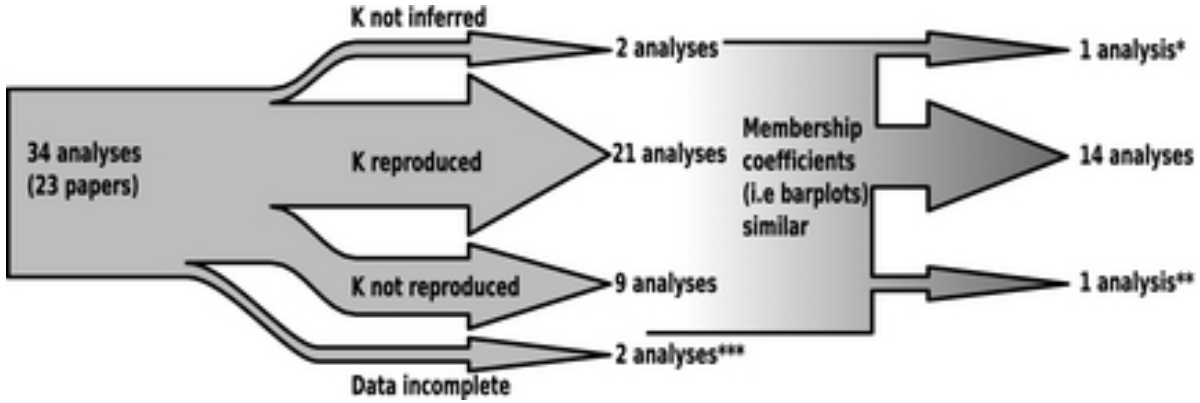
reconstructed heatmap



original heatmap
Hu et al. (2007)

Population Genomics

UBC Reproducibility Group could not reproduce the results in 30% of published analyses using the population genetic package STRUCTURE, using the same data as provided by the authors



http://onlinelibrary.wiley.com/doi/10.1111/j.1365-294X.2012.05754.x/abstract

Gene name errors are widespread

- 20% of papers in leading genomics journals with supplementary Excel gene lists have erroneous gene name conversions
- 40% of Excel files with gene lists deposited to NCBI GEO (4321 screened) contain gene name errors

Automatic conversion of gene symbols to dates and floating-point numbers

Gene symbol	converted to
SEPT2 (Septin 2)	2-Sep
SEPT2	2006/09/02
MARCH1 [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase]	1-Mar
2310009E13	2.31E+13

http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7

Data Management Plan (DMP)

"Plan the life cycle of data: long-term perspective by outlining how data will be generated, collected, documented, shared and preserved"

asked for by funding agencies
e.g. SNF http://www.snf.ch/en/theSNSF/research-policies/open_research_data

Some Tips for Better Data Management, Processing and Analysis

1. Consistent project organization

```
ComparativeGenomics_medea
├── raw                 read-only
├── data                contains scripts
├── results
└── manuscript
```

```
ComparativeGenomics_medea
├── raw
├── data
├── results
│   ├── Arabidopsis     contains scripts
│   ├── maize           contains scripts
│   └── simulation      contains scripts
└── manuscript
```

separation of data, method, output

…different tastes…

Descriptions

```
ComparativeGenomics_medea
├── raw
├── data
└── results
    ├── Arabidopsis
    ├── maize
    └── simulation
```

Dates

```
ComparativeGenomics_medea
├── raw
├── data
└── results
    ├── 2017-05-10
    └── 2017-05-17
```

Mixed

```
ComparativeGenomics_medea
├── raw
├── data
│   └── 2017-05-17
│       ├── Arabidopsis
│       └── maize
└── results
    ├── 2017-05-17
    │   ├── Arabidopsis
    │   └── maize
    └── 2017-05-20
        └── simulation
```

(data&results -> experiments)

Descriptions 2

```
ComparativeGenomics_medea
├── 00_RawFiles
├── 01_Trimming
├── 02_Mapping
└── 03_SNPanalysis
    ├── Arabidopsis
    ├── maize
    └── simulation
```

2. Names matter

NO

myabstract.docx
Joe's Filenames Use Spaces and Punctuation.xlsx
figure 1.png
fig 2.png
JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

YES

2014-06-08_abstract-for-sla.docx
joes-filenames-are-getting-better.xlsx
fig01_scatterplot-talk-length-vs-interest.png
fig02_histogram-talk-attendance.png
1986-01-28_raw-data-from-challenger-o-rings.txt

Jenny Brian

no spaces, punctuation, accents

2. File/Folder naming

- machine readable
- human readable
- plays well with default ordering (some numbers first)

2017-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2017-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2017-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2017-06-26_BRAFWTNEGASSAY_FFPE-CRC-1-41-A01.csv
2017-06-26_BRAFWTNEGASSAY_FFPE-CRC-1-41-A02.csv
2017-06-26_BRAFWTNEGASSAY_FFPE-CRC-1-41-A03.csv

Jenny Brian

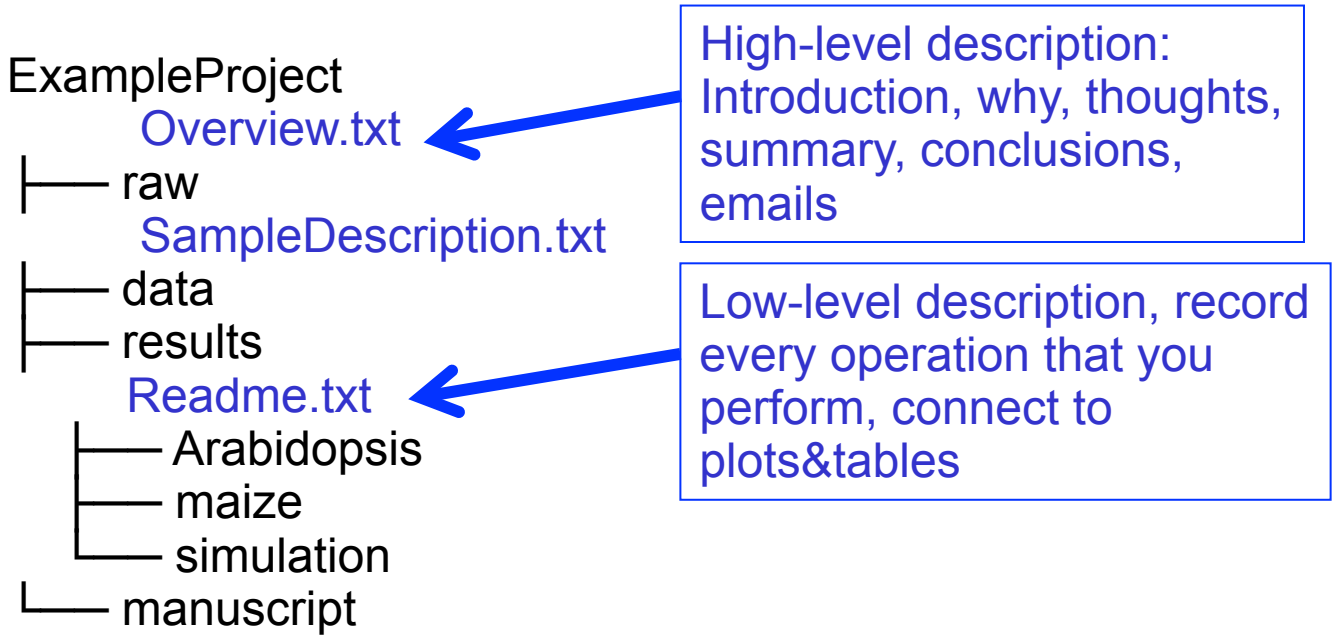
2. File/Folder naming cont.

01_marshal-data.r	01.r
02_pre-dea-filtering.r	02.r
03_dea-with-limma-voom.r	03.r
04_explore-dea-results.r	04.r
90_limma-model-term-name-fiasco.r	90.r
helper01_load-counts.r	helper01.r
helper02_load-exp-des.r	helper02.r
helper03_load-focus-statinfo.r	helper03.r
helper04_extract-and-tidy.r	helper04.r

Jenny Brian

3. For every result, keep track of how it was produced

- Many steps
- Pre- and post-processing steps are often critical to reproduce results
- Each folder contains a Readme.txt/Notebook/Description file that describes the folder
- relatively verbose!
- Dated entries



4. Use plain text formats

Untitled — Edited 58 Words

```
# Markdown demo

### human-readable

### easy formatting

It's very easy to make some words **bold** and
other words *italic* with Markdown. You can
even [link to Google!](http://google.com)

[Link to different Markdown document](docs/
CONTRIBUTING.md)

### also images

If you want to embed images, this is how you do
it:

![Logo](https://github.com/swyder/
Reproducible_Research/raw/master/
Logo_URPP_kl2.png)
```

Markdown demo

human-readable

easy formatting

It's very easy to make some words **bold** and other words *italic* with Markdown. You can even link to Google!

Link to different Markdown document

also images

If you want to embed images, this is how you do it:

- plain text (comma or tab delimited) or Markdown
- doesn't require special software to read: xls > csv/tsv, doc > txt
- minimum: keep a copy of your data files in plain text format
- designed to produce documents from human readable text
- Easy conversion -> pdf / html / doc / docx

5. Avoid manual data manipulation steps

- Manual manipulations (e.g. Excel)
  - Inefficient
  - Error-prone
  - Difficult to reproduce

- Scripts are also documentations

- Research is iterative, automating tasks saves time in the longer run
  - workflow modifications (e.g. reviewers' request)
  - new data
  - more data
  - new group members
  - collaborations

6. Keep raw files read-only

- Leave untouched original raw files (e.g. never open using Excel)
- Write-protect raw files once created
- Store them in a separate folder

- You know the raw data are in the right format if you:
  1. Ran no software on the data
  2. Did not modify any of the data value
  3. Did not remove any data from the data set
  4. Did not summarize the data in any way

  e.g. the strange binary fille (?) your measurement machine spits out, hand-entered number you collected at the microscope, FASTQ sequencing reads

7. Backup your data & documentation

"There are 2 kinds of people: people who backup their computers, and people who have never lost everything"

- One day, your hard / flash drive will fail!
  HD crash, malware, theft, fire/flooding, coffee spills, user error ...

- Choose >1 method, make multiple copies:
  - local: Institute's NAS, external HD
  - off-site: cloud/Switchdrive, DVD/HD stored at home
- Backup software:
  Time Machine (Mac), tarball (Linux)
- Test Backup&Restoring (e.g. overwrite with empty file)

- Backup <-> Syncing: versioning
- Syncing: Switchdrive, dropbox, Google drive, rsync/rsnapshot (Linux)

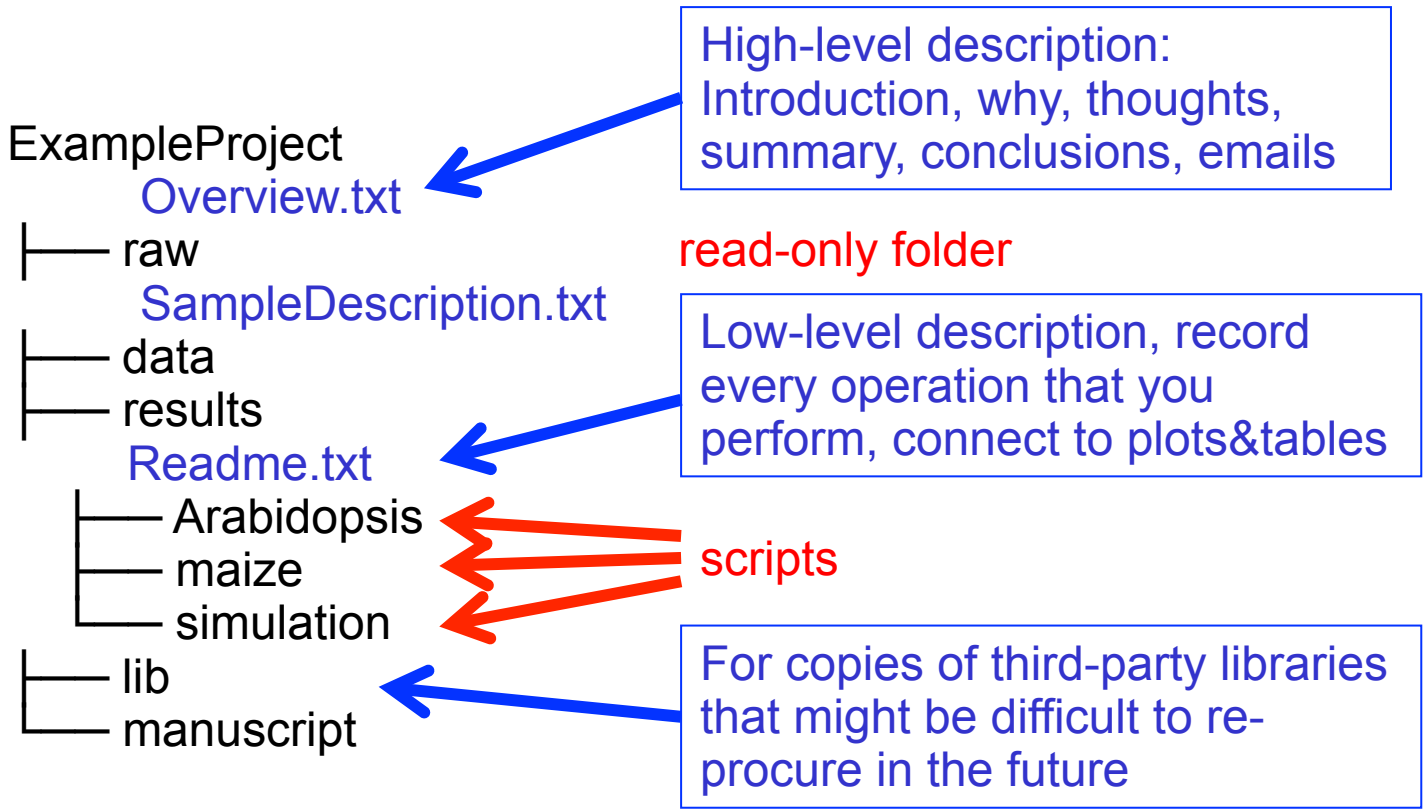
8. Deposit data and metadata onto public repositories in formats that anyone can find, access and reuse without restriction

FAIR data principles

- Findable — rich metadata, unique identifier
- Accessible — long-term storage, open, non-commercial
- Interpretable — controlled vocabulary, well-defined structure
- Re-usable — sufficiently well-described (what for,who, when,limitations,variable names), license

SNF recommends 4 general (Dryad, EUDAT, Harvard Dataverse, Zenodo) + field-specific repositories

Summary



- Reproducible research is hard work - trade-off
- less friction to repeat, share project with group member - increased productivity
- partial reproducibility is better than nothing!

Sources & Links

Presentation by Frédéric Schütz (SIB Lausanne)

Good Enough Practices in Scientific Computing
http://arxiv.org/abs/1609.00037

Ten Simple Rules for Reproducible Computational Research
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285

A Quick Guide to Organizing Computational Biology Projects
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424

More things about reproducibility

Article collection from Nature
http://www.nature.com/news/reproducibility-1.17552

Nature poll about reproducibility crisis
http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Practical Data Science for Stats - a PeerJ Collection
https://peerj.com/collections/50-practicaldatascistats/

Data organization

A Hormone_Data_raw.csv

PatientID	Cortisol	IGF1	...	Hormone50
ID1	17.4	327	...	33.5
ID2	18.1	412	...	44.2
⋮	...	...	...	...
ID40	20.2	264	...	28.6

B Demographic_Data.csv

PatientID	Age	Sex	BMI	CollectionDate	Diagnosis
ID1	45	female	18	2016-09-25	control
ID2	12	female	17	2016-09-25	diabetes
⋮	...	...	...	...	...
ID40	40	male	22	2016-09-29	control

C CodeBook.docx:

Study Design:

Experimental Question: This study looks to determine whether or not there are differences in hormone levels in individuals with diabetes relative to healthy controls.

Sample Details: 20 individuals with diabetes and 20 unrelated age- and sex-matched controls were included for study. Individuals were recruited to the study using flyers posted throughout Johns Hopkins Hospital and online recruitment through www.website.com. Informed consent was obtained from all study participants. Blood was drawn by a single phlebotomist in clinic X and all samples processed on the same day they were collected by company Y.

Code Book/Data dictionary:

Variable	Description	Units	CodingNotes	OtherNotes
Age	Age At Blood Draw	years	numerical	Taken from electronic medical record
Sex	Self-reported	'male', 'female'	2-level factor	Confirmed using electronic medical record
BMI	weight/height	kg/m^2	numerical	Measured day of blood draw
Collection Date	Date of Blood Draw	date	YYYY-MM-DD	Collection of blood by phlebotomist
Diagnosis	Individual diagnosis	'diabetes', 'control'	2-level factor	`diabetes` = Type 2 Diabetes. Confirmed by medical record.
Cortisol	Stress Hormone	µg/dL	numerical	Required fasting and to be measured in the AM (8-10am)
IGF1	Insulin-Like Growth Factor 1	ng/dL	numerical	Did not require fasting, but taken at the same time as other measures
⋮	⋮	⋮	⋮	⋮
Hormone50	Hormone Name	ng/dL	numerical	Hormone Details

D Pseudocode.docx:

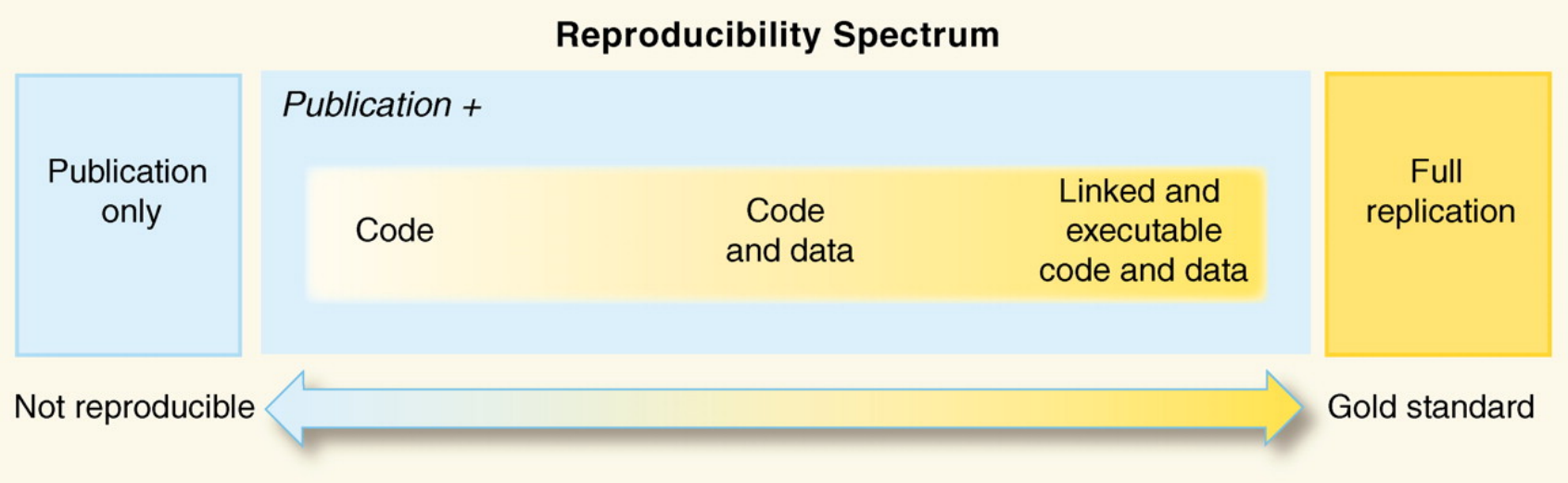
1. Values for hormone levels were received from company Y and input into Horome_Data_raw.csv. No processing has been done on these values.
2. Values in Demographic_Data.csv were obtained upon visit to clinicX. Data were extracted from electronic medical record and input into Excel by Jane Doe.

https://peerj.com/preprints/3139/

High-level description / Lab notebook

- complete picture of the development of the project over time
- contains a prose description of the experiment (driver script contains all the gory details)
- notes from conversations and paste e-mail text
- record your observations, conclusions, and ideas for future work
- The URL can also be provided to PI / remote collaborators to give them status updates on the project
- Particularly when an experiment turns out badly, it is tempting simply to link the final plot or table of results and start a new experiment. Before doing that, it is important to document how you know the experiment failed, since the interpretation of your results may not be obvious to someone else reading your lab notebook.

The spectrum of reproducibility



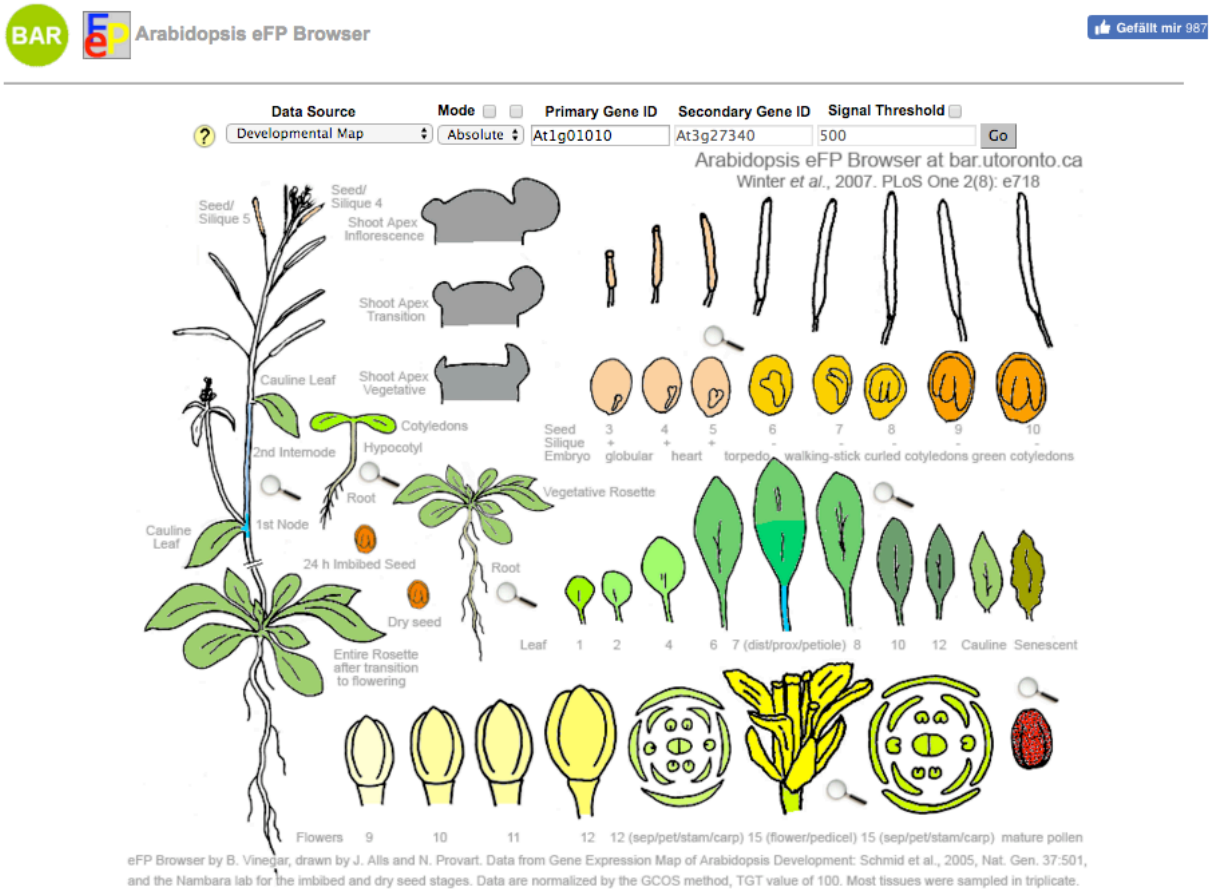
http://science.sciencemag.org/content/334/6060/1226.full

- A minority of the papers available today provide code and data
- Making articles reproducible takes time and effort (e.g. recreating figures as in publication)

- Partial reproducibility is better than nothing!

Why open data?

- Interrograte and Reuse - Hypothesis Creation
- Aggregating - Metaanalysis / systematic reviews
- Aggregating and integration - Databases / Information Resources (e.g. TAIR, Araport, InterMINE, iPlant)



Toronto BAR