# Reproducible Research

Stefan Wyder

stefan.wyder@evolution.uzh.ch

September 2017

Universität Zürich UZH

URPP Evolution in Action

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

One of the main principles of the scientific method

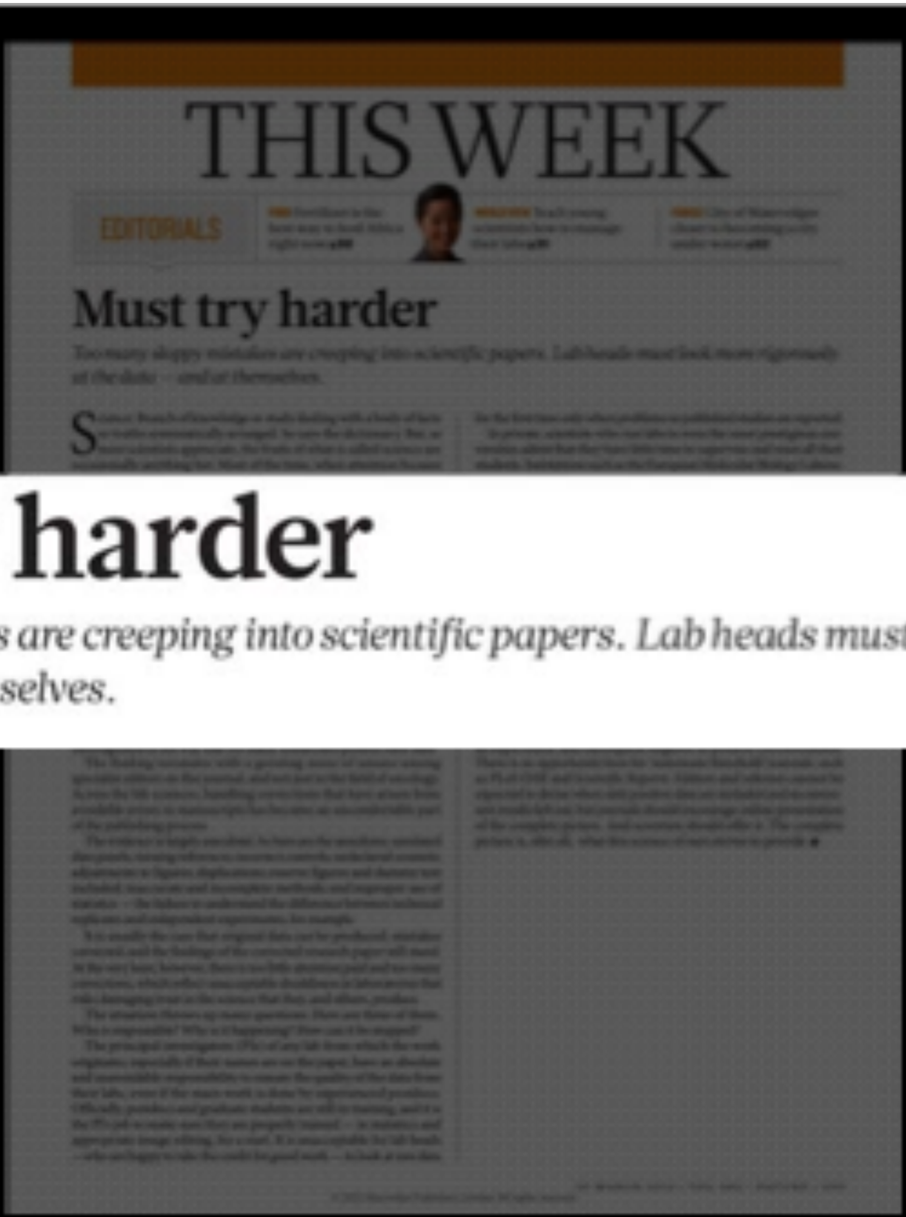
# Definition of Reproducible Research

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 S. Noble



THIS WEEK

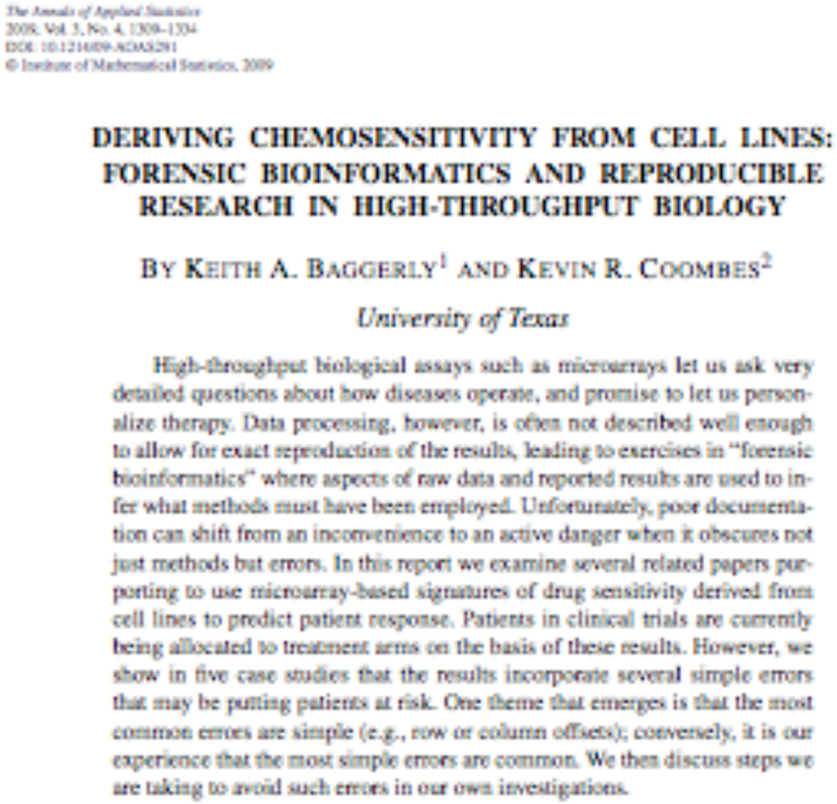
EDITORIALS

# Must try harder

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

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

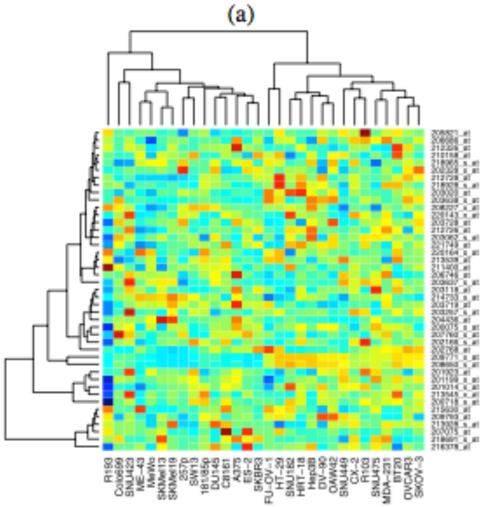
BY KEITH A. BAGGERLY[1] AND KEVIN R. COOMBES[2]

*University of Texas*

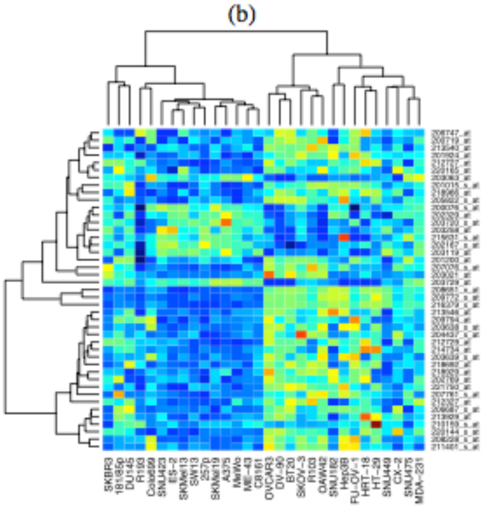
High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic 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.

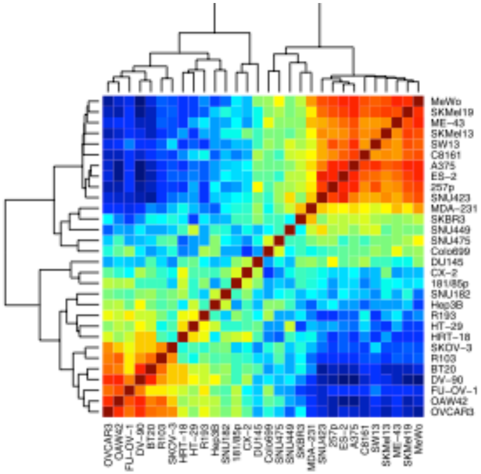
# Reconstructing heatmap for cisplatin signature



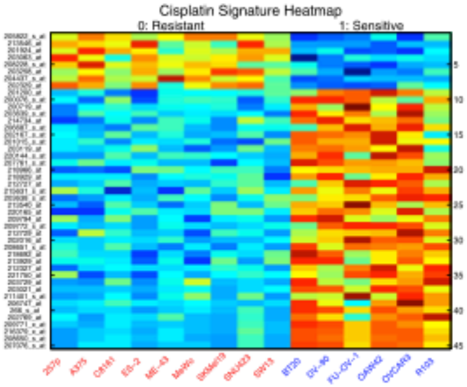
orig data: no structure



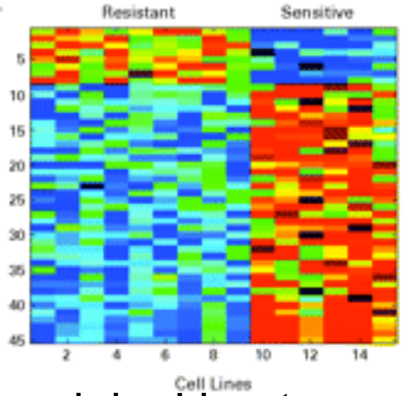
offsetting by one
(indexing error)



pairwise sample correlations
to detect label switches



reconstructed heatmap



original heatmap
Hu et al. (2007)

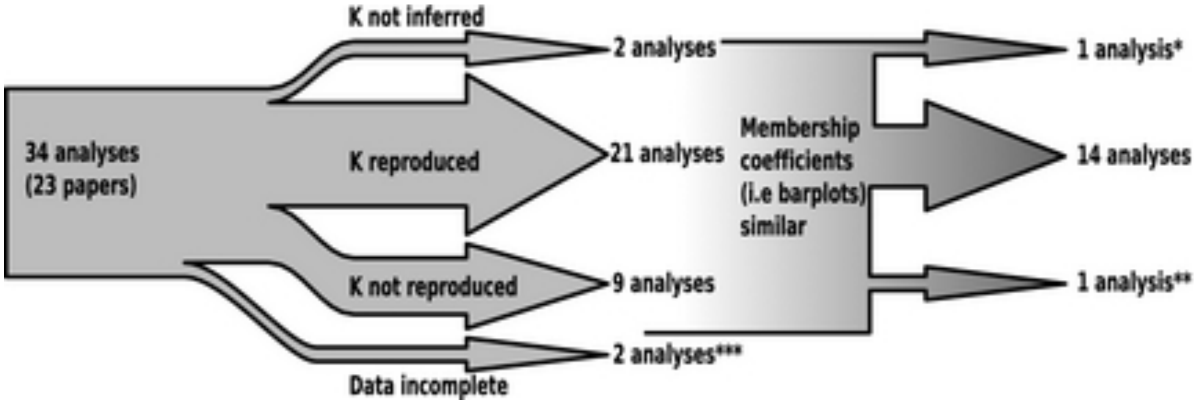
# Conclusions of Baggerly and Coombes

- The most common problems are simple:
  - confounding in the experimental design
  - mixing up sample labels
  - mixing up the gene labels (off-by-one errors)
  - mixing up the group labels (sensitive/resistant)
- Most of these mix-ups involve simple switches or offsets.
- These mistakes are easy to make, particularly if working with Excel/oocalc
- Or if working with 0/1 labels instead of names

These easy-to-fix errors are often hidden (incomplete documentation and code not shared)

# 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

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

| 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 |

- 20% of papers in leading genomics journals with supplementary Excel gene lists have erroneous gene name conversions

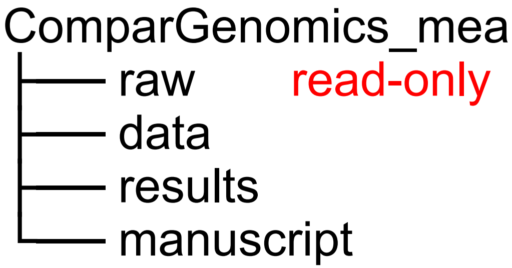
- 39.7% of Excel files with gene lists deposited to NCBI GEO (4321 screened) contain gene name errors

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

# Practical aspects

- How do organize your folders?
- How you name your files?

# Folder organization



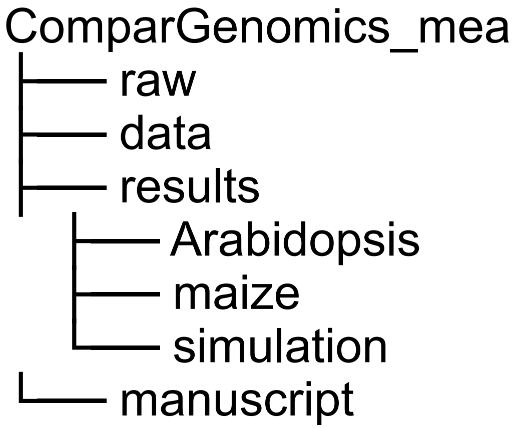
CompareGenomics_mea

raw read-only

data

results

manuscript



ComparGenomics_mea

raw

data

results

Arabidopsis

maize

simulation

manuscript

- Result folders contain scripts
- Each file exists only once - use aliases (shortcuts / symbolic links)

# …different tastes…

**Descriptions**

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

**Dates**

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

**Numbered**

```
ComparGenomics_mea
├── raw
├── data
└── results
    ├── analysis1
    ├── analysis2
    └── analysis3
```

**Descriptions 2**

```
ComparGenomics_mea
├── raw
├── trimmed_data
├── mapped_data
└── results
    ├── Arabidopsis
    ├── maize
    └── simulation
```

# Description files

- We can never fit in a single filename all the metadata to describe a file
- Each folder contains a Readme.txt/Notebook/Description file that describes the folder
- Format: plain text or Markdown

```
CompparGenomics_mea
        Readme.txt            High-level description: Introduction,why,thoughts,conclusions
├── raw
        SampleDescription.txt
├── data
├── results
       Readme.txt             low-level description, describes all analyses, links to plots
    ├── Arabidopsis
    ├── maize
    └── simulation
└── manuscript
```

# 3 Principles for file naming

1. machine readable

2. human readable

3. plays well with default ordering
   - put something numeric first

Jenny Brian

# File naming example

| | |
|---|---|
| 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

# File naming example 2

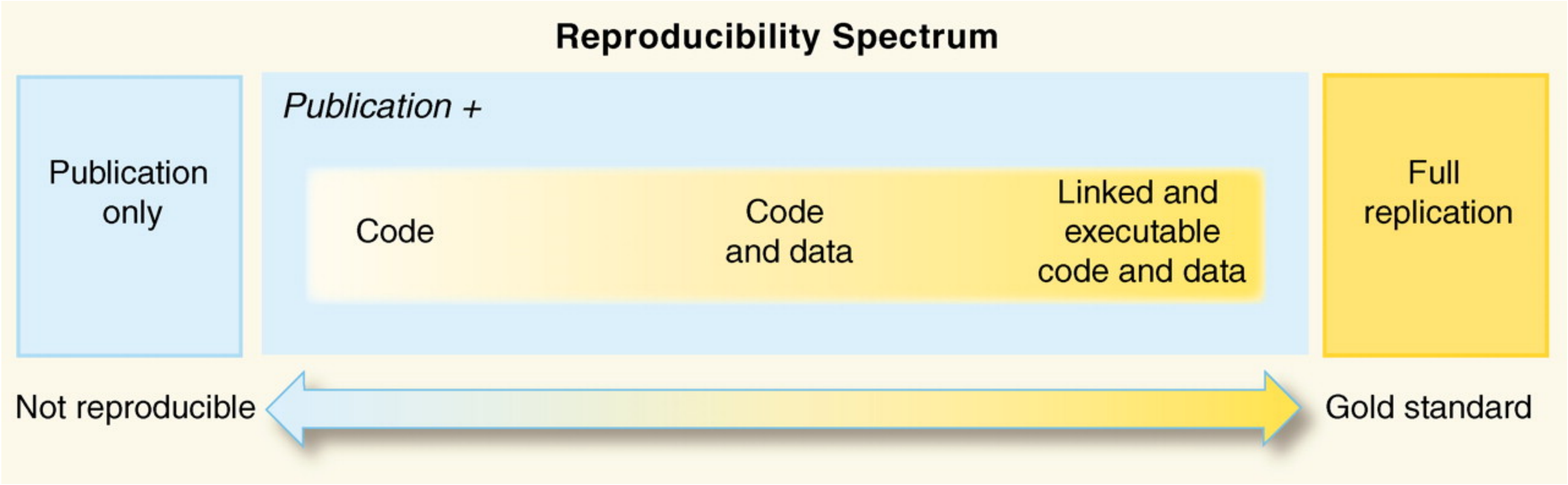
2017-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2017-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2017-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2017-06-26_BRAFWTNEGASSAY_FFPE-CRC-1-41-A01.csv
2017-06-26_BRAFWTNEGASSAY_FFPE-CRC-1-41-A02.csv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

Machine-readable

- easy to search for later
- easy to narrow file lists based on names
- easy to extract info from file names (e.g. by splitting)
- no spaces, punctuation, accents

Jenny Brian

# 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

Partial reproducibility is better than nothing!

# Research as an iterative process

Everything you do, you will ~~probably~~ have to do over again

modifications in preprocessing / analysis
more / new data
new group members

Reproducible research -> less friction, time-saver in the longer run

# Handy tools

- Electronic notebooks
  allow to mix text, code and plots in the same document

  R: rmarkdown, knitr, Jupyter Notebook
  Python: Jupyter Notebook

- "Workflow managers"
  run analyses using GUI
  take care of many aspects (history, ..)

  Galaxy, sushi (FGCZ)

- Virtualization
  Virtual machines, Docker

# 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

**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/

# Dynamic reports

- Combining thoughts and code

- Dynamic creation of figures
  One can delete all figures and recreate them at will

- R: rmarkdown, knitr, Jupyter Notebook
  python: Jupyter/IPython Notebook

- Demo
  http://rmarkdown.rstudio.com/lesson-1.html
  http://jupyter.org/

- Problems:
  - reduced development environment (<-> Rstudio)
  - long-running calculations
  - big files
  - code to wrangle/polish data might be long