DEALING WITH HIGH DIMENSIONAL DATA

SITC WINTER SCHOOL
FEBRUARY 21, 2019

Leslie Cope, PhD
Division of Biostatistiics and Bioinformatics
Sidney Kimmel Comprehensive Cancer Center at JHU
cope@jhu.edu

With thanks to Rob Scharpf, PhD

WHAT IS BIG DATA?

Calling Bullshit
@callin_bull

Follow

Big Data: (n): the belief that a big enough pile of horseshit will, with probability one, somewhere contain a pony.

(thanks to @mlipsitch)

4:37 PM - 14 Feb 2017

WHAT IS BIG DATA?

big data *n*. *Computing* (also with capital initials) data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.

1980—2012

http://www.oed.com

HIGH DIMENSIONAL MOLECULAR DATA

- Too big to inspect
- Requires computational tools for even the simplest manipulation
- Generally requires programming skills
  - at least comfort reading and modifying scripts

IF YOU LEARN ONE LANGUAGE…

WHY R?

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Events

Course material

Course material from many previous events is available.

Upcoming

BioC 2019: Where Software and Biology Connect
24 - 27 June 2019 — New York, USA

Previous (recent)

EuroBioc2018

Bioconductor version 3.8 (Release)

Autocomplete biocViews search:

- Software (1649)
  - AssayDomain (661)
  - BiologicalQuestion (668)
  - Infrastructure (360)
  - ResearchField (728)
  - StatisticalMethod (572)
  - Technology (1049)
  - WorkflowStep (884)
- AnnotationData (941)
- ExperimentData (360)
- Workflow (23)

GOALS

I hope to address two main ideas:

- What are some of the unique opportunities (and challenges) offered by high dimensional data
- What is reproducible research and why is it (especially) a concern for high dimensional data

YOU CAN'T STEP IN THE SAME RIVER ONCE

Sequencing reads

Raw Data

+ Methods

+ Annotations

= Data

Various workflows

Knowledge Bases

Gene Expression
Mutations
TCR
Patient Biomarkers

STANDARDS FOR REPORTING DATA

GEO deposit procedures enable and encourage submitters to supply MIAME and MINSEQE compliant data. All GEO submission procedures are designed to closely follow the MIAME and MINSEQE checklists; if you provide all requested information, your submission will be compliant.

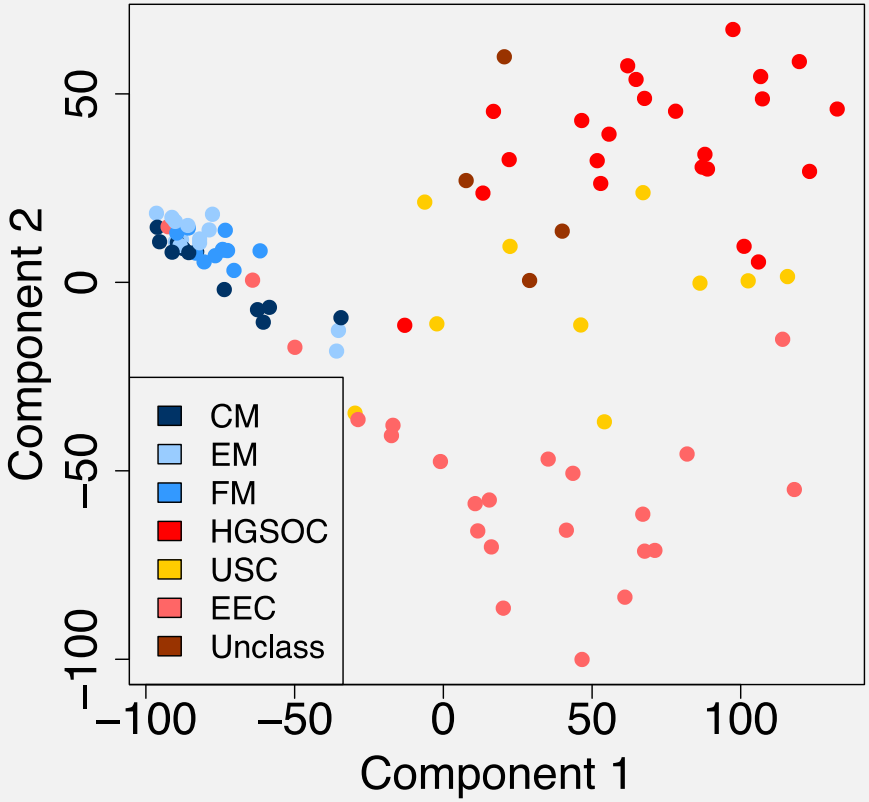
The six most critical elements contributing towards MIAME are:

- Raw data for each assay (e.g., CEL or FASTQ files)
- Final processed (normalized) data for the set of assays in the study (e.g., the gene expression data count matrix used to draw the conclusions in the study)
- Essential sample annotation (e.g., tissue, sex and age) and the experimental factors and their values (e.g., compound and dose in a dose response study)
- Experimental design including sample data relationships (e.g., which raw data file relates to which sample, which assays are technical, which are biological replicates)
- Sufficient annotation of the array or sequence features examines (e.g., gene identifiers, genomic coordinates)
- Essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

Gene Expression Omnibus Submission Guidelines

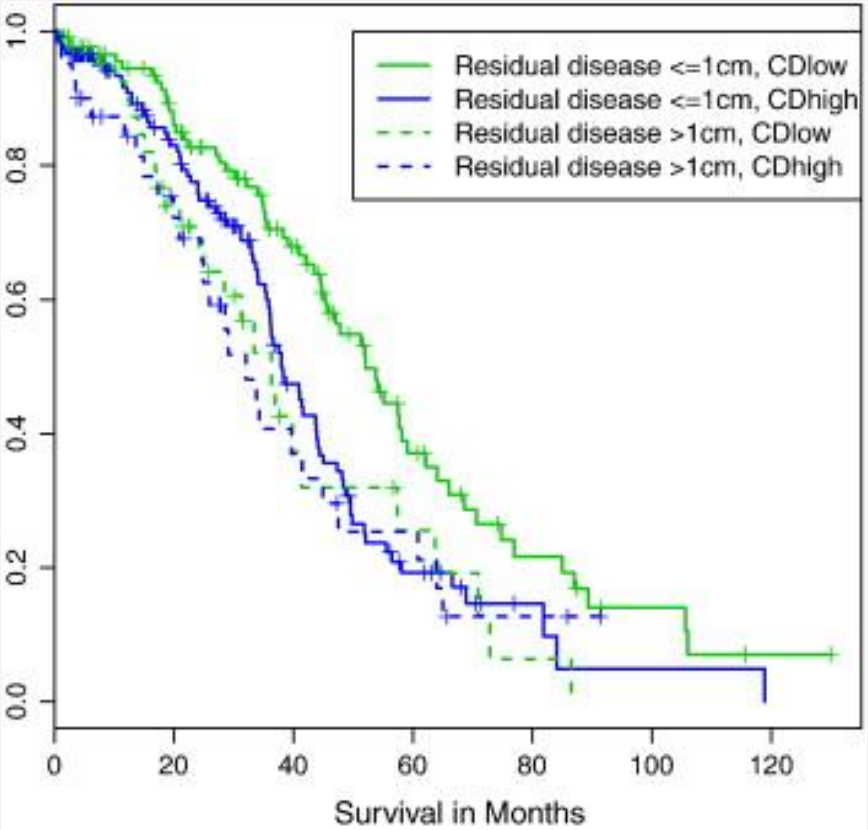
ANALYSIS OF HIGH DIMENSIONAL DATA

HIGH DIMENSIONAL DATA DOESN'T ALWAYS REQUIRE SPECIALIZED STATISTICS



Pearson, K. (1901). *Philosophical Magazine*. **2** (11): 559–572

WE APPLY FAMILIAR METHODS TO LOW DIMENSIONAL SUMMARIES



OR BREAK PROBLEMS DOWN INTO MANY, LOW DIMENSIONAL ANALYSES

Gene	logFC	t	P.Value	adj.P.Val
CCNA2	-1.9032858	-66.257167	4.1703E-15	8.6053E-11
NENF	-3.8532791	-50.970523	6.4895E-14	5.3616E-10
SPIN4	1.45010331	49.0466208	9.703E-14	5.3616E-10
ANKRD20A2	2.42809525	48.725257	1.0393E-13	5.3616E-10
METTL6	-1.8804462	-46.158899	1.8294E-13	7.5499E-10
FDXR	2.75450305	44.0121708	3.0084E-13	1.0346E-09

CHALLENGES

MULTIPLE TESTS

Gene	logFC	t	P.Value	adj.P.Val
CCNA2	-1.9032858	-66.257167	4.1703E-15	8.6053E-11
NENF	-3.8532791	-50.970523	6.4895E-14	5.3616E-10
SPIN4	1.45010331	49.0466208	9.703E-14	5.3616E-10
ANKRD20A2	2.42809525	48.725257	1.0393E-13	5.3616E-10
METTL6	-1.8804462	-46.158899	1.8294E-13	7.5499E-10
FDXR	2.75450305	44.0121708	3.0084E-13	1.0346E-09

MULTIPLE TESTS

Suppose my "scientific method" was to:

- Pull a slip of paper from my ”random hypothesis" hat
- Carry out a related experiment and analyze the data
- Discard or Report, depending on whether $p < .05$
- Repeat

I will get positive results – lots if I can work quickly.

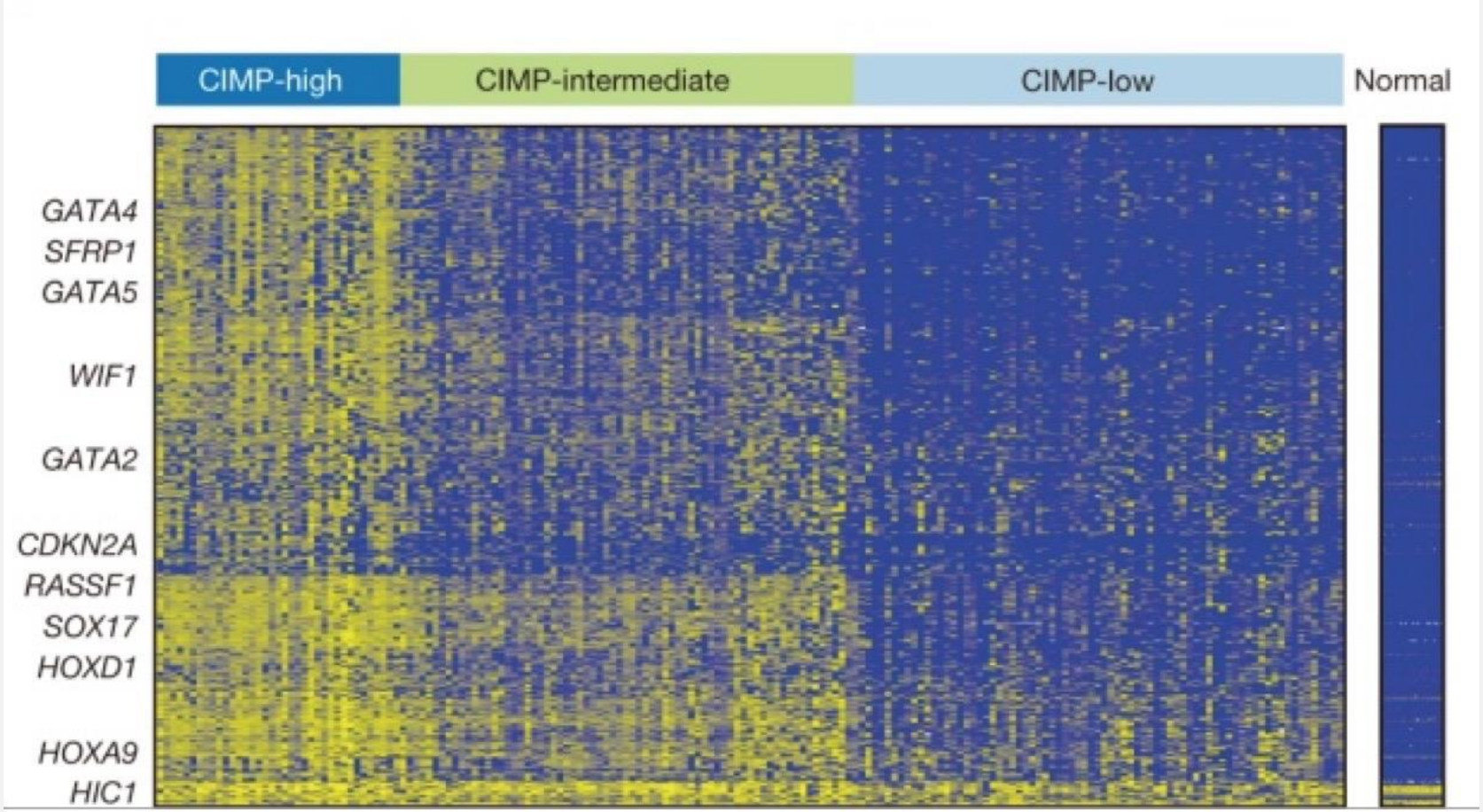
MULTIPLE TESTS

- Large Search Space + Small Sample Size = Inflated Performance in training data and unsuccessful testing
- Training set of 5 cancer samples, 5 normal samples
- 10 coins as candidate biomarkers
- Expected training set error for best of 10 coins/biomarkers is 26%
- Expected test set error is of course 50%

OPPORTUNITIES

CLUSTERING

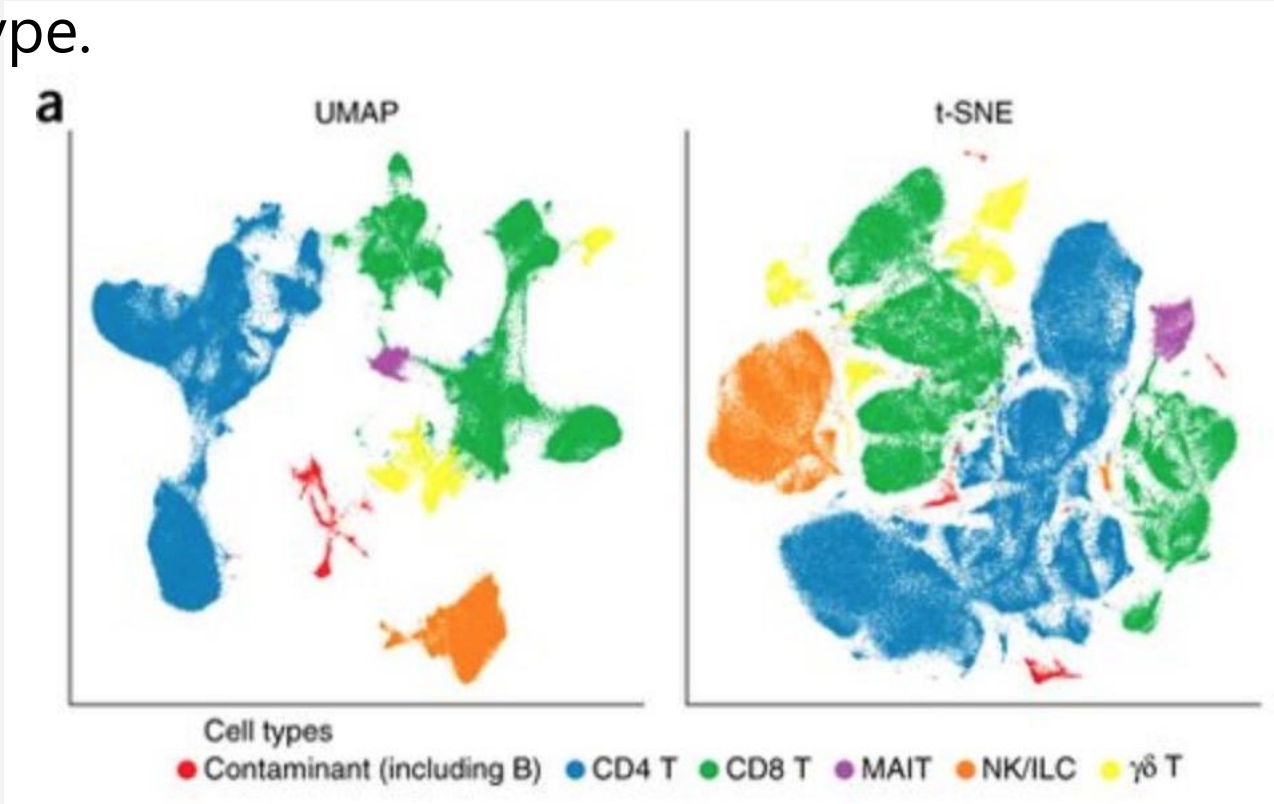
Patterns of DNA methylation in TCGA lung adenocarcinoma



EA Collisson *et al. Nature* **000**, 1-8 (2014) doi:10.1038/nature13385

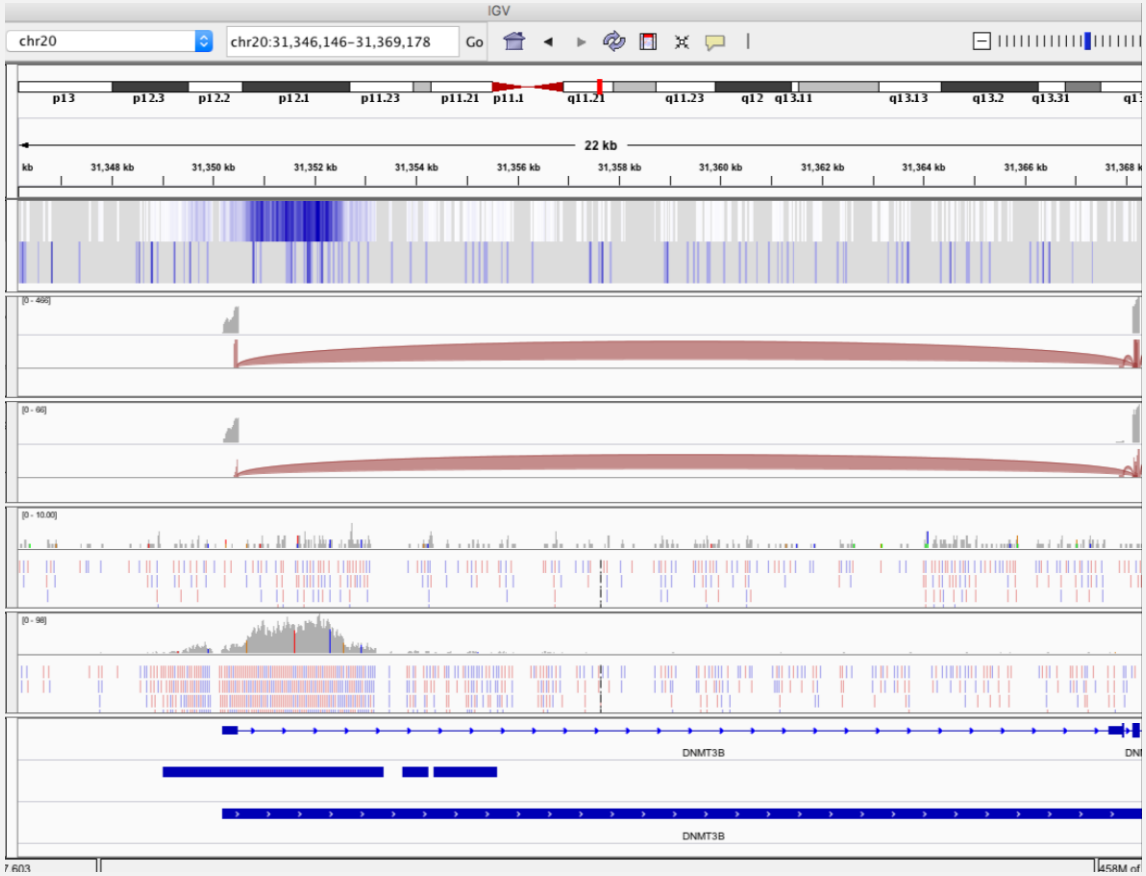
IDENTIFYING STRUCTURE

Unsupervised dimension reduction reveals single cell data structures associated with cell type.



Becht et al. ,*Nature Biotechnology* **volume37**, pages38–44 (2019)

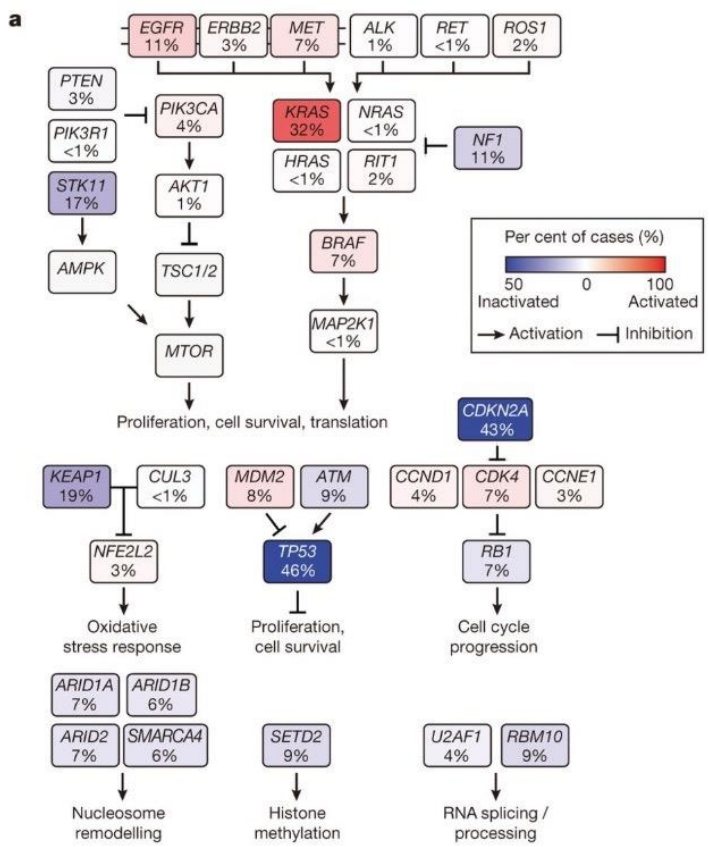
INTEGRATION

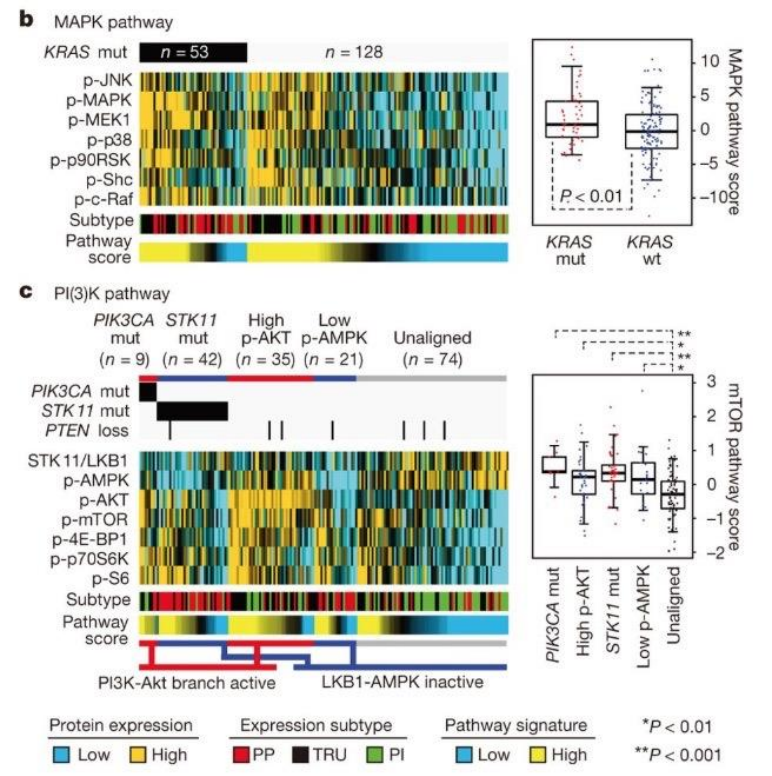


FUNCTIONAL INFERENCE

Figure 4 : Pathway alterations in lung adenocarcinoma.

From: Comprehensive molecular profiling of lung adenocarcinoma

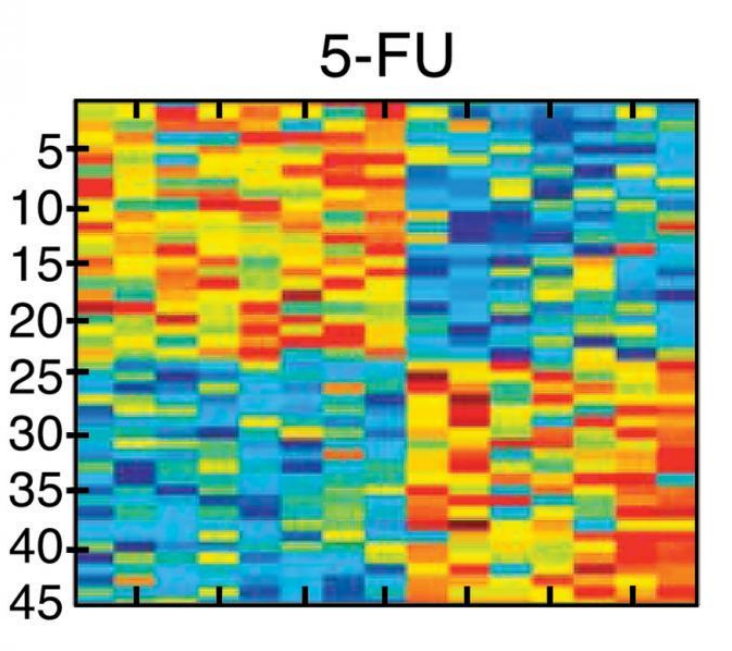




EA Collisson *et al. Nature* **000**, 1-8 (2014)doi:10.1038/nature13385

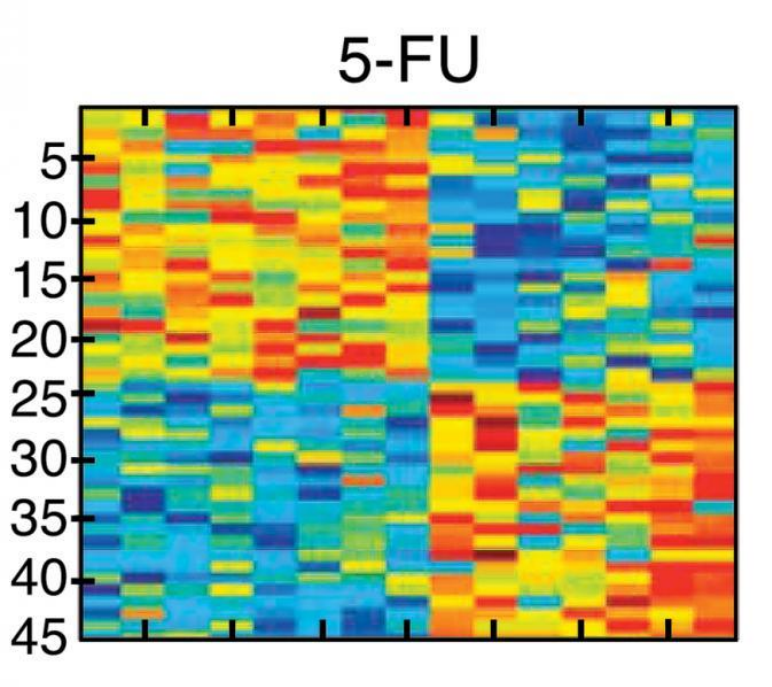
REPRODUCIBLE RESEARCH

"THE INCIDENT"

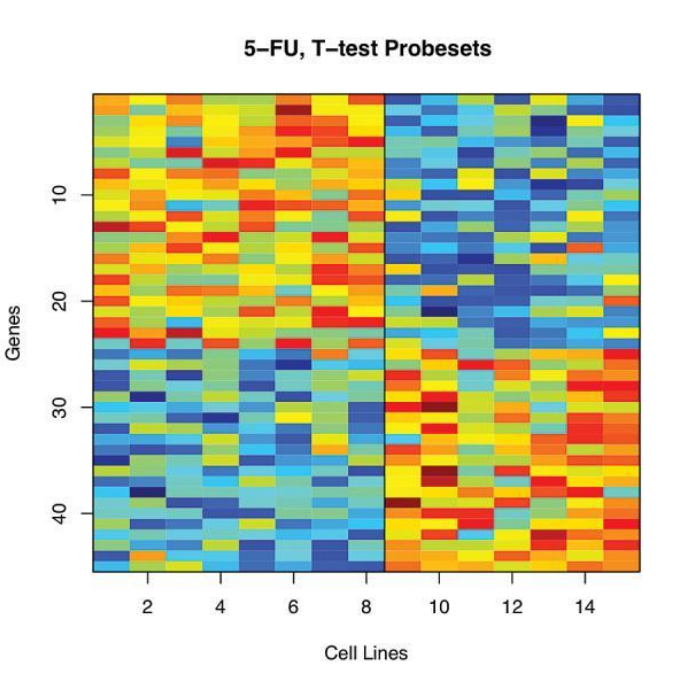


Potti et al. Nature Medicine 2006; 12: 1294-1300.

THE IMAGE CAN BE REPRODUCED (APPROX.)

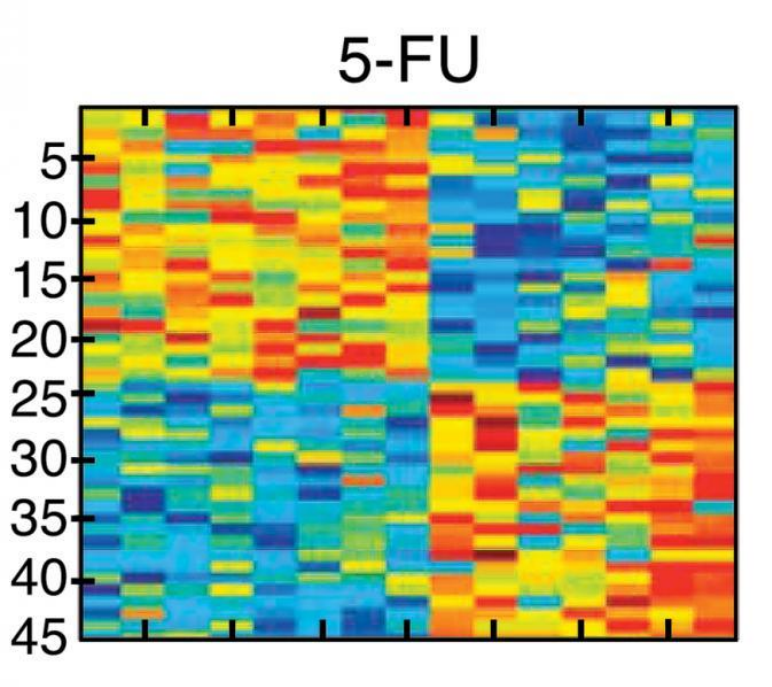


Nature Medicine 2006; 12: 1294-1300.

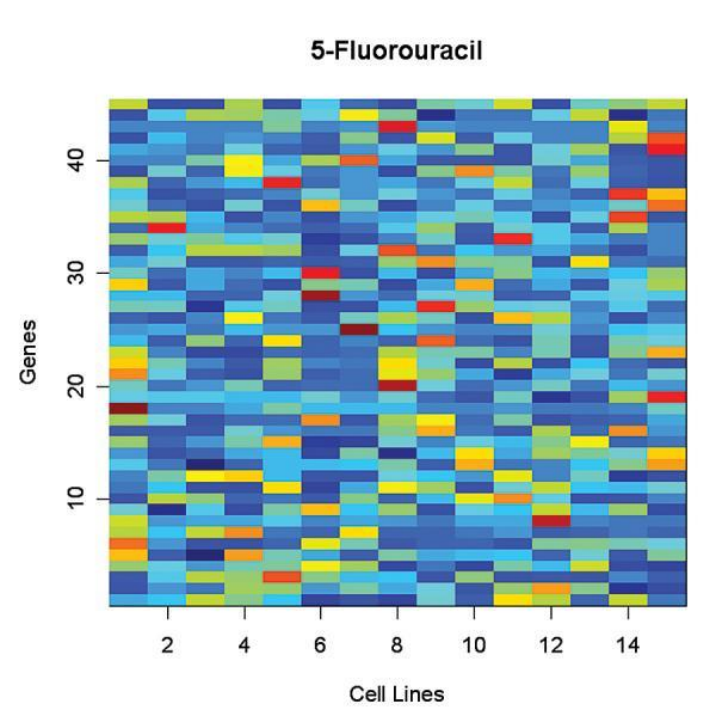


Baggerly, Keynote Address,
Council of Scientific Editors Meeting
May 3, 2011, Baltimore MD

BUT NOT USING THE REPORTED GENE SIGNATURE



Nature Medicine 2006; 12: 1294-1300.



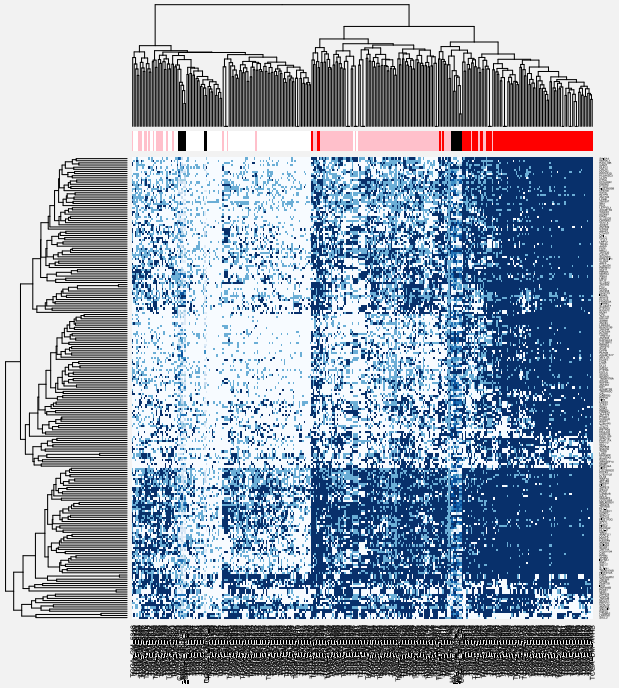
Baggerly, Keynote Address, Council of Scientific Editors Meeting May 3, 2011, Baltimore MD

THE EXPLANATION "OFF BY ONE ROW"

Published Probe IDs	Actual Probe IDs
1881_at	1882_at
31321_at	31322_at
31725_at	31726_at
32307_at	32308_at
…	…

Baggerly Keynote Address,
Council of Scientific Editors Meeting
May 3, 2011, Baltimore MD

RECURRING NIGHTMARE



“Can we add the new cell lines to this figure?”

RECURRING NIGHTMARE

```
pdf("heatmap.pdf")
heatmap(betaPerGene2[cands,],
col=cols,ColSideColors=colColors)
dev.off()
```

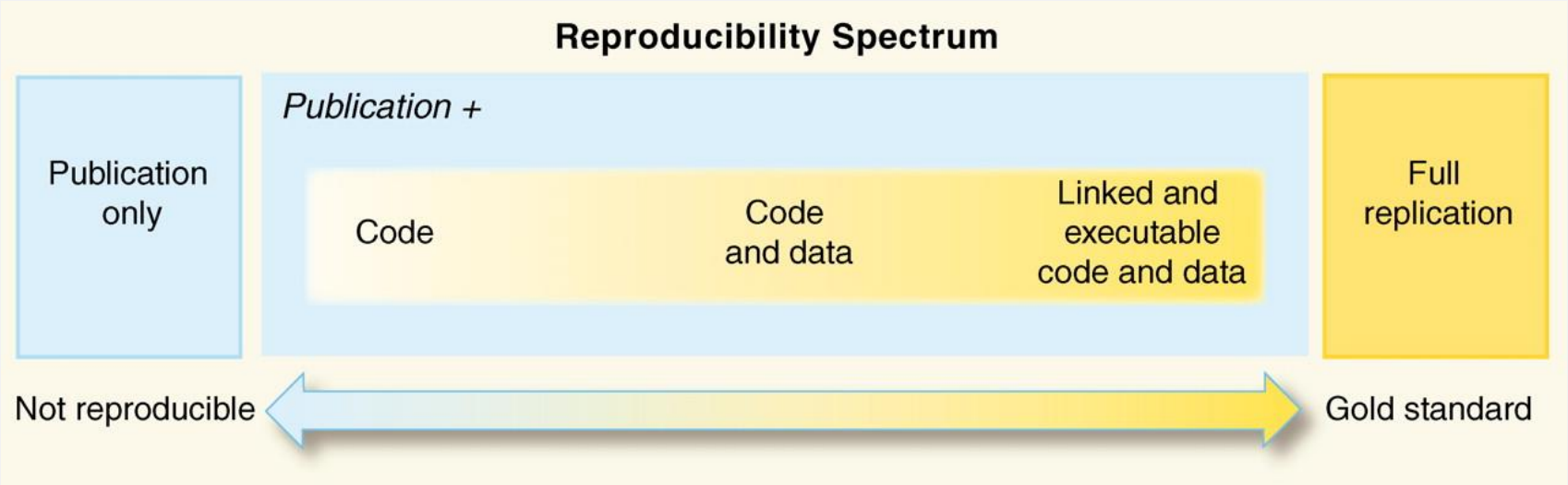
REPRODUCIBLE RESEARCH

Reproducible - Replicable

REPRODUCIBLE RESEARCH

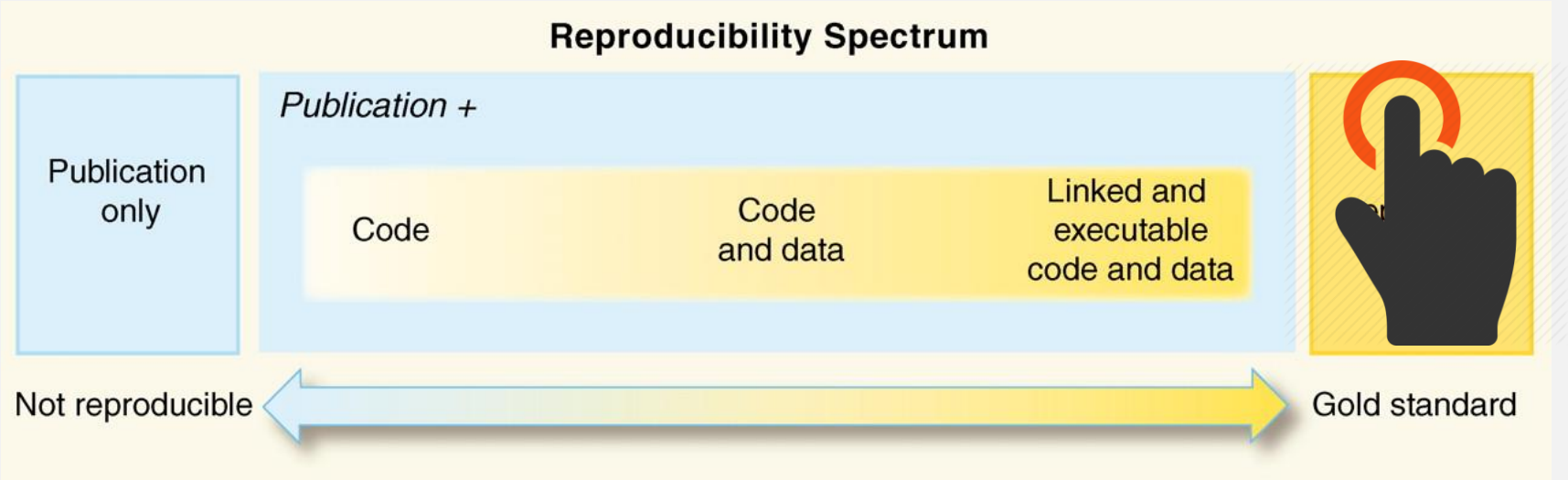
Reproducible - Correct

WHAT DOES REPRODUCIBLE RESEARCH MEAN



Roger Peng, *Science* 02 Dec 2011

WHAT DOES REPRODUCIBLE RESEARCH MEAN



ACHIEVING THE GOAL

Literate Code

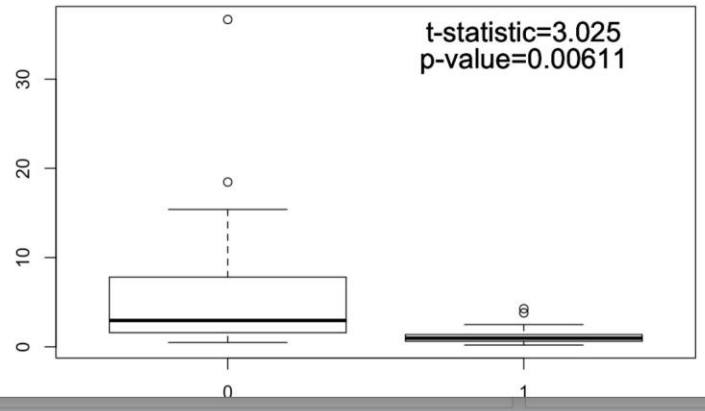
The gold standard for reproducible research requires that source data,

ID	value	status	age
S21	1.518	0	64
S90	0.966	1	38
S35	1.761	0	38

as well as the code used to analyze it,

```
test.results=t.test(value~status,data=mydata)
```

be packaged with the manuscript that results from that analysis.



ACHIEVING THE GOAL

```
#### The gold standard for reproducible research requires that source data,
```{r readData, echo=F}
mydata=read.table("./mydata.txt",header=T)
kable(mydata[1:3,])
```

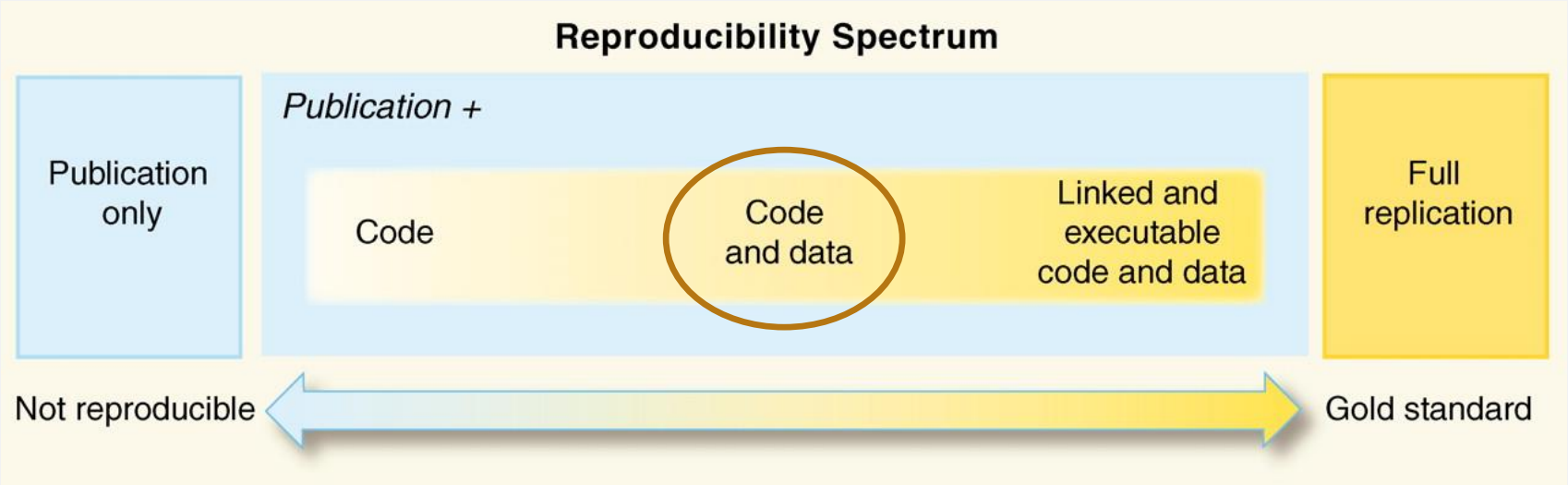
as well as the code used to analyze it,
```{r analyzeData, echo=TRUE}
test.results=t.test(value~status,data=mydata)
```

be packaged with the manuscript that results from that analysis.
```{r plot,echo=F}
boxplot(value~status,data=mydata)
text(2,32,paste0("t-statistic=",round(summary(test.results$statistic),5)),p
os=3)
text(2,30,paste0("p-value=",round(summary(test.results$p.value),5)),pos=3)
```
```

CONFESSION

I have never published a paper where the final manuscript can be reproduced from included data by compiling literate code

WHAT IS PRACTICAL?



Roger Peng, *Science* 02 Dec 2011

RULE 1: CODE EVERYTHING

```
#### first work out sample names
## open core sample shee
sampleSheet=read.csv("96 Illumina  Epic Sample sheet  _TLWang _01192017.csv"

### and phenodata
library(xlsx)
pData=read.xlsx("sample diagnosis for data analysis_01302017.xlsx",sheetI=1)
pSamps=gsub(" ","",as.character(pData[,1]))

### find matches
aSamps=gsub(" ","",as.character(sampleSheet[,1]))
pSamps=gsub(" ","",as.character(pData[,1]))
matches=sapply(pSamps,grep,x=aSamps,fixed=T)
```

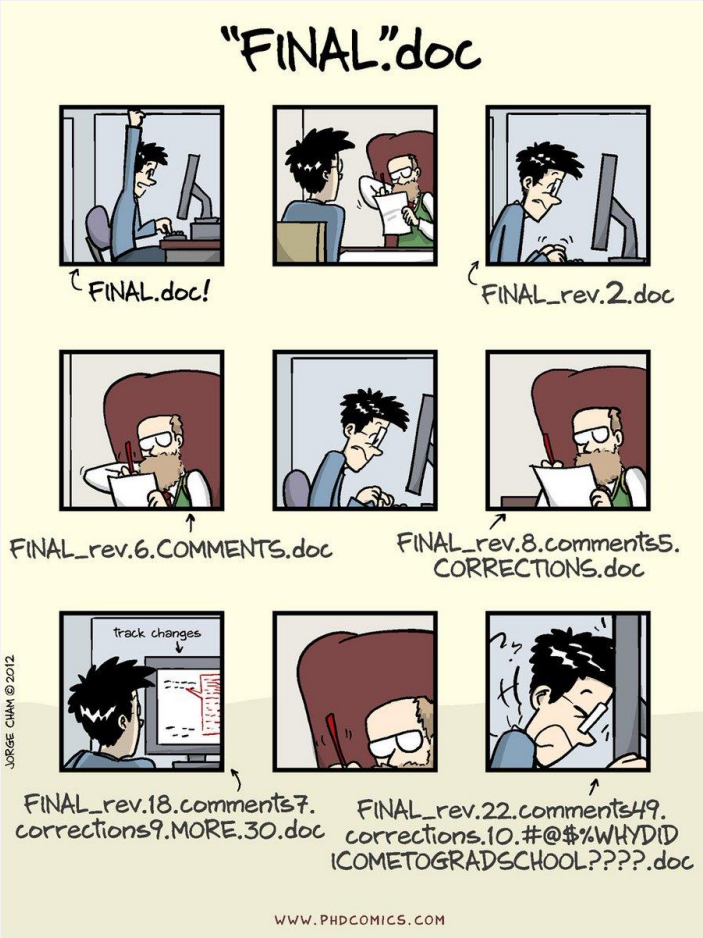
RULE 2: ORGANIZE DATA

"…one should *organize the data in the form that you expect to be made public*, and work from those files."

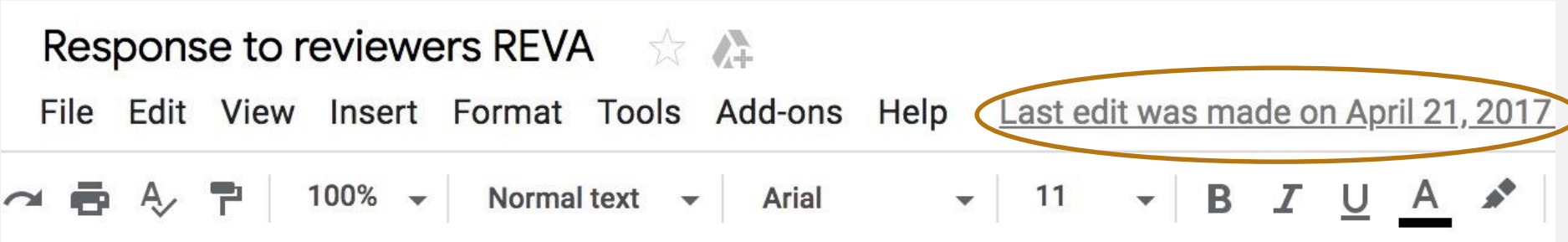
-Karl Broman

https://kbroman.org/blog/2015/09/09/reproducibility-is-hard/

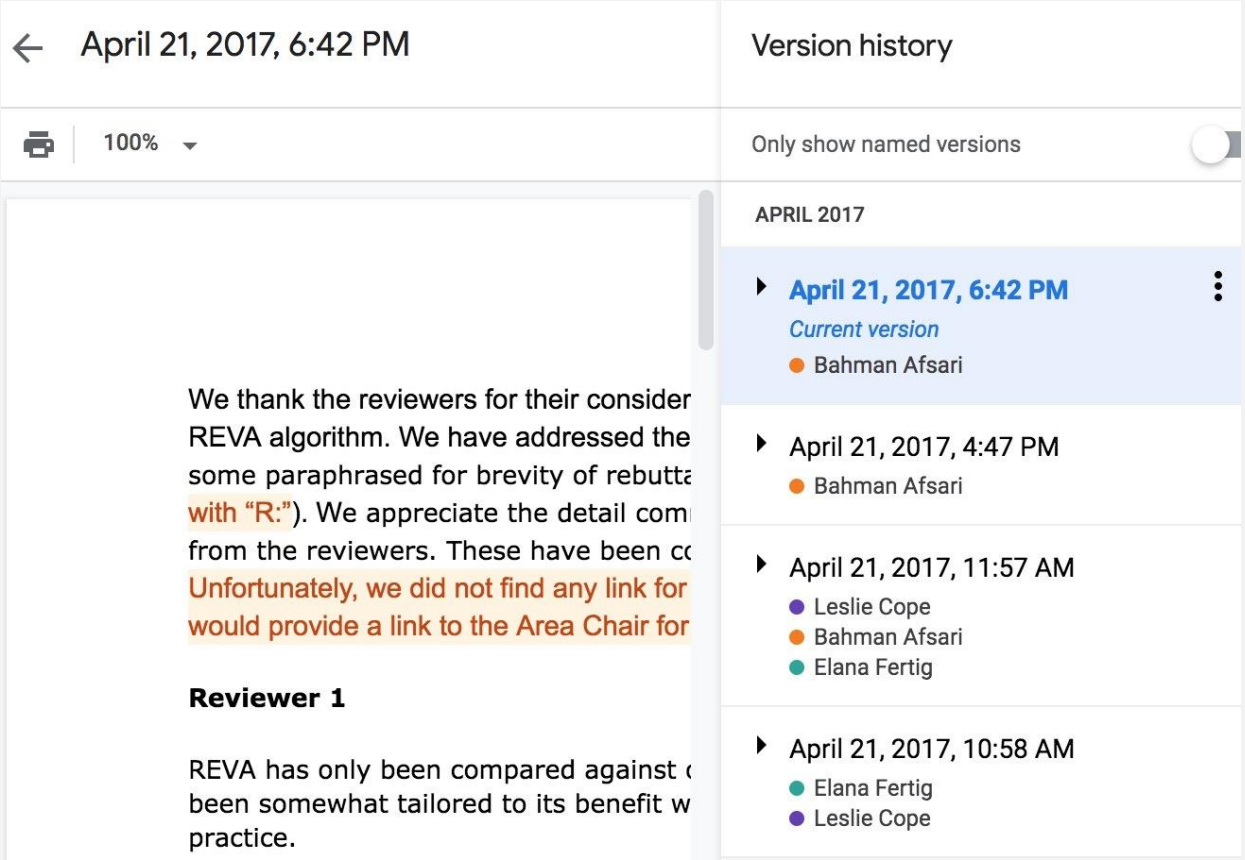
RULE 3: MAINTAIN VERSION CONTROL



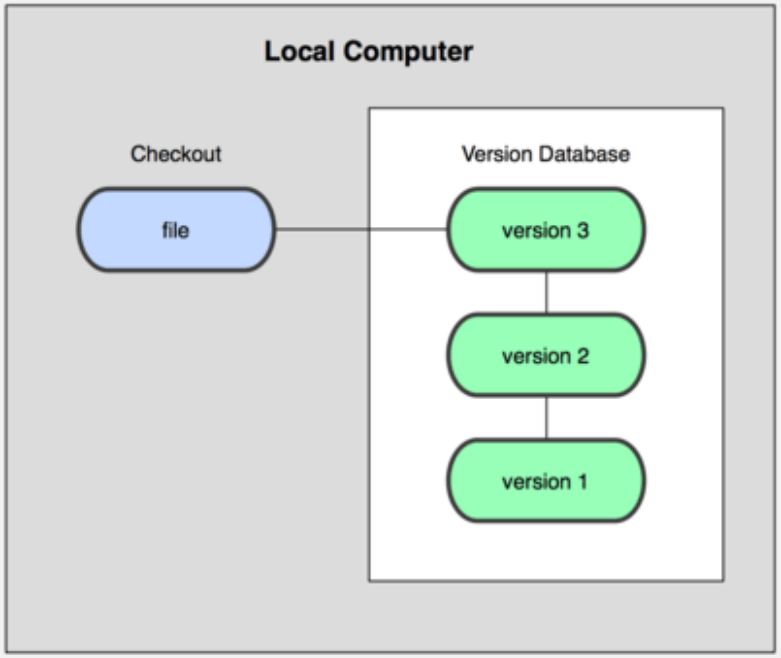
RULE 3: MAINTAIN VERSION CONTROL



RULE 3: MAINTAIN VERSION CONTROL



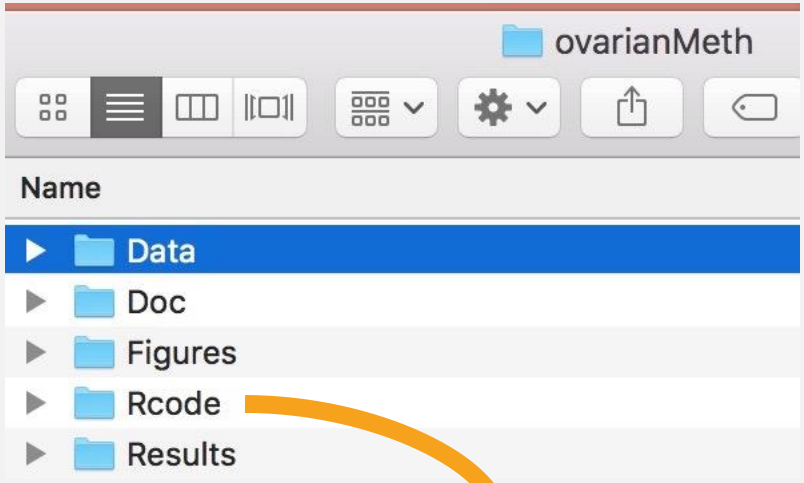
RULE 3: MAINTAIN VERSION CONTROL

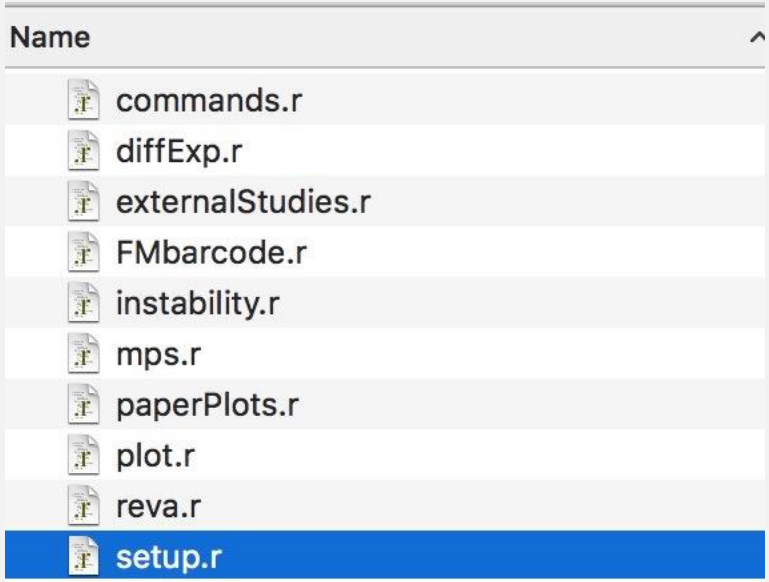


Chacon and Straub, *Git Pro 2nd Ed.* 2014, https://git-scm.com/book/en/v2

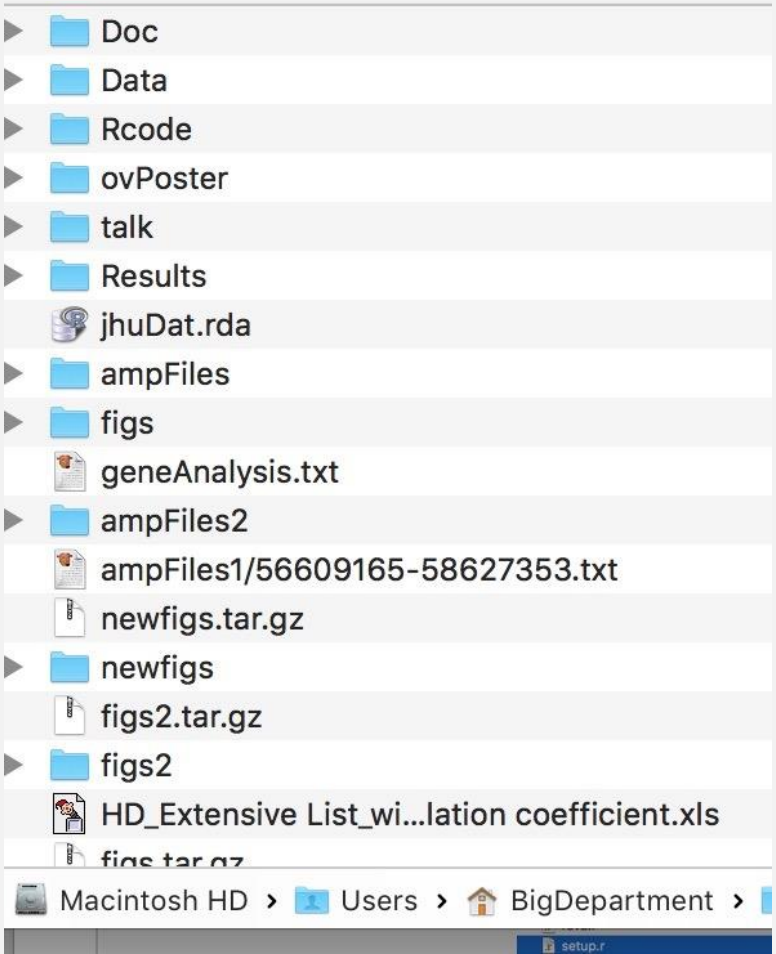
RULE 4: DO EVERYTHING THE SAME WAY, EVERY TIME

RULE 4: DO EVERYTHING THE SAME WAY, EVERYTIME





RULE 4: DO EVERYTHING THE SAME WAY, EVERYTIME



GOLDEN RULE: BE KIND TO FUTURE YOU

- Files
  - Use meaningful, searchable names for data files, figures, tables
  - Organize input data, work, and results in clearly labeled folders
  - Include simple text files describing work flows, file organization,
- Code
  - Use meaningful, searchable variable names
  - Comment code generously

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

-Keith Baggerly

RESOURCES

- General
  - ROpenSci's Reproducibility guide
- Online Course
  - Reproducible Research (part of the Johns Hopkins data science specialization at Coursera)
- Books
  - Dynamic documents with R and knitr
  - Reproducible research with R and Rstudio
  - Implementing reproducible research