Reproducible Research.. Why we love R & Bioconductor

Aedín Culhane

aedin@jimmy.harvard.edu

Boston Bioconductor Course, Oct 24/25th http://bosbioc.wordpress.com/

My R Course Website http://bcb.dfci.harvard.edu/~aedin/

My HSPH homepage http://www.hsph.harvard.edu/research/aedin-culhane/

Reproducible Research

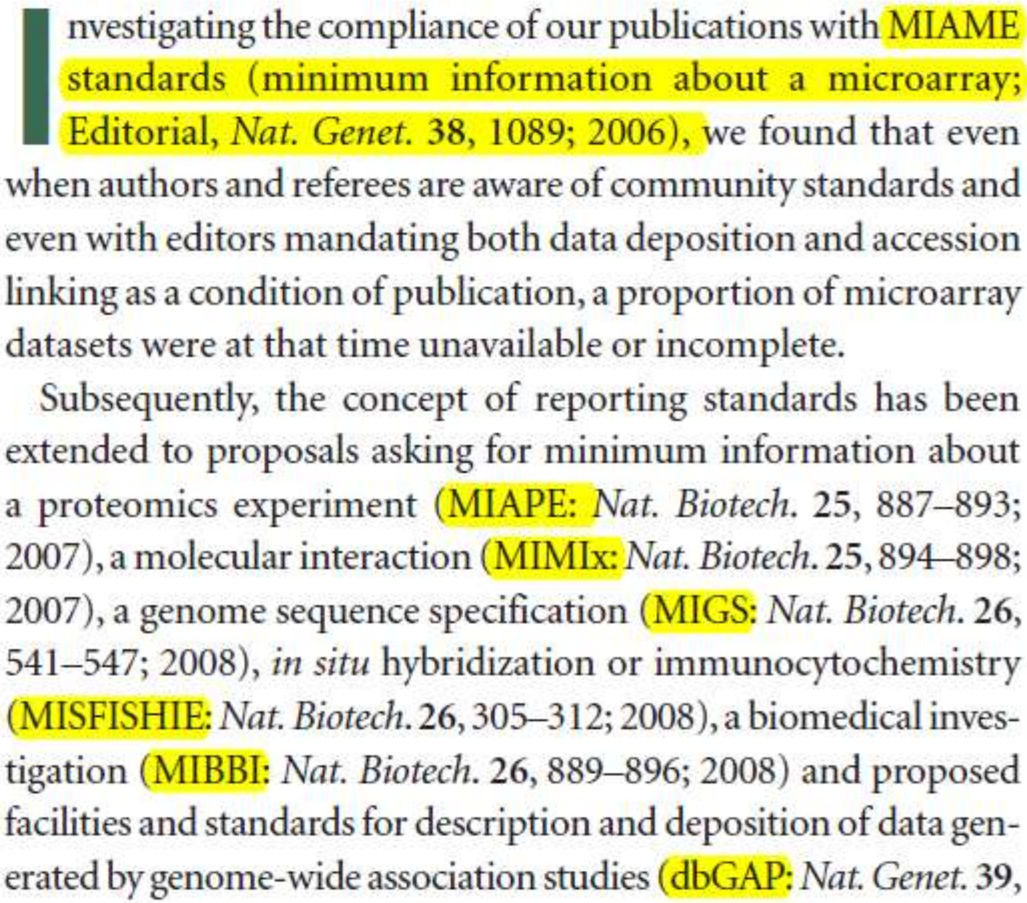
- Can I repeat the analysis described in Paper Y
  - Can I get their data?
  - Can I understand their methods?
  - Did they fully describe the methods?
  - Are the methods reproducible
- Can it be repeated on an independent (validation) dataset

Reproducibility: Importance of Sharing

How to share?

1. Data Standards

2. Data Repositories

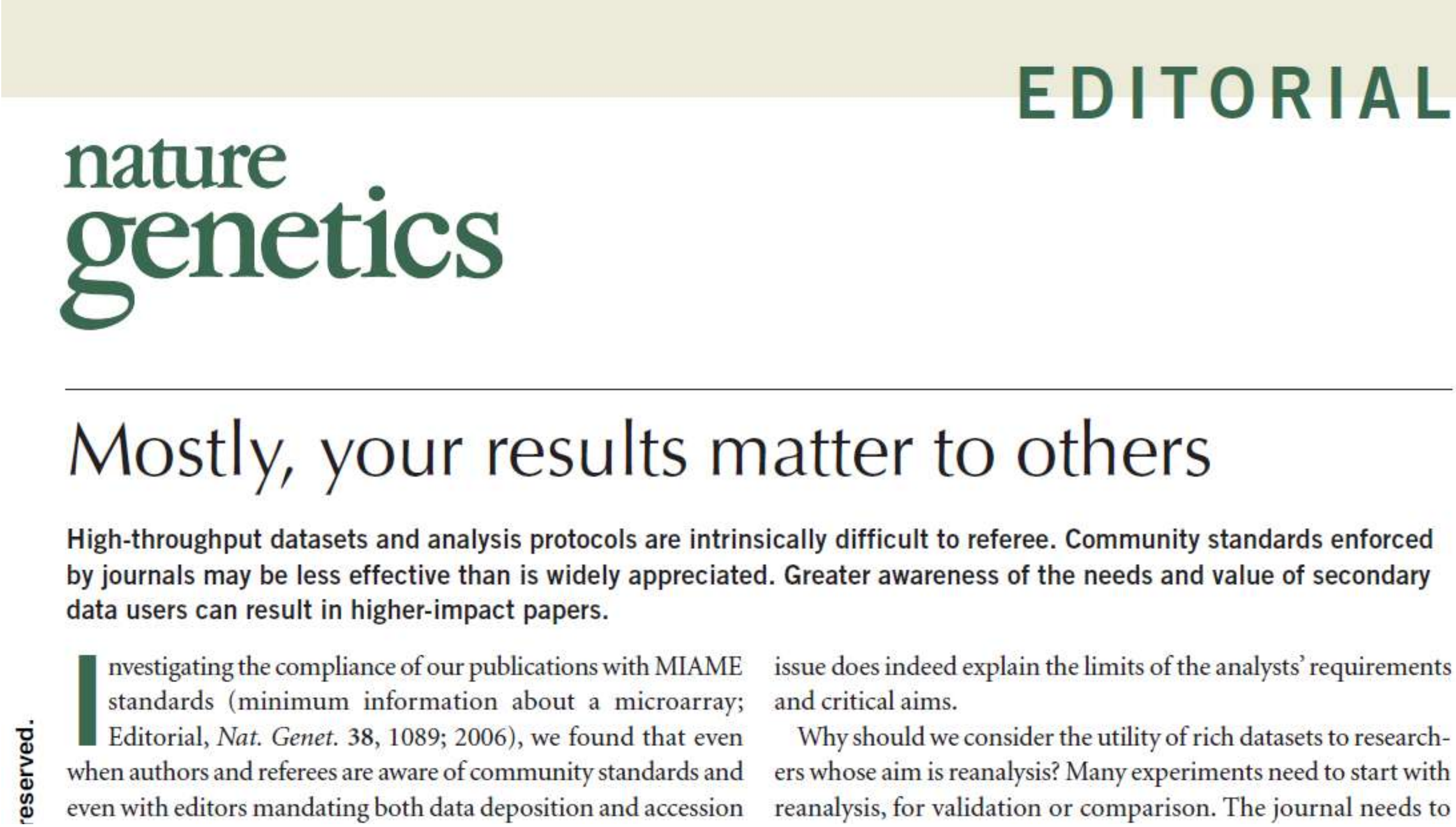


Investigating the compliance of our publications with MIAME standards (minimum information about a microarray; Editorial, *Nat. Genet.* **38**, 1089; 2006), we found that even when authors and referees are aware of community standards and even with editors mandating both data deposition and accession linking as a condition of publication, a proportion of microarray datasets were at that time unavailable or incomplete.

Subsequently, the concept of reporting standards has been extended to proposals asking for minimum information about a proteomics experiment (MIAPE: *Nat. Biotech.* **25**, 887–893; 2007), a molecular interaction (MIMIx: *Nat. Biotech.* **25**, 894–898; 2007), a genome sequence specification (MIGS: *Nat. Biotech.* **26**, 541–547; 2008), *in situ* hybridization or immunocytochemistry (MISFISHIE: *Nat. Biotech.* **26**, 305–312; 2008), a biomedical investigation (MIBBI: *Nat. Biotech.* **26**, 889–896; 2008) and proposed facilities and standards for description and deposition of data generated by genome-wide association studies (dbGAP: *Nat. Genet.* **39**,

Reproducible Analysis

- Research papers have minimal space to describe methods…
  - “Analyst reads paper. Finds algorithms described in English sentences that occupy minimal amounts of space in the methods section.
  - Analyst acquires public data from the paper. Makes wild guesses at actual algorithms and parameters. Is unable to reproduce the reported results.”
    - Keith Baggerly



EDITORIAL

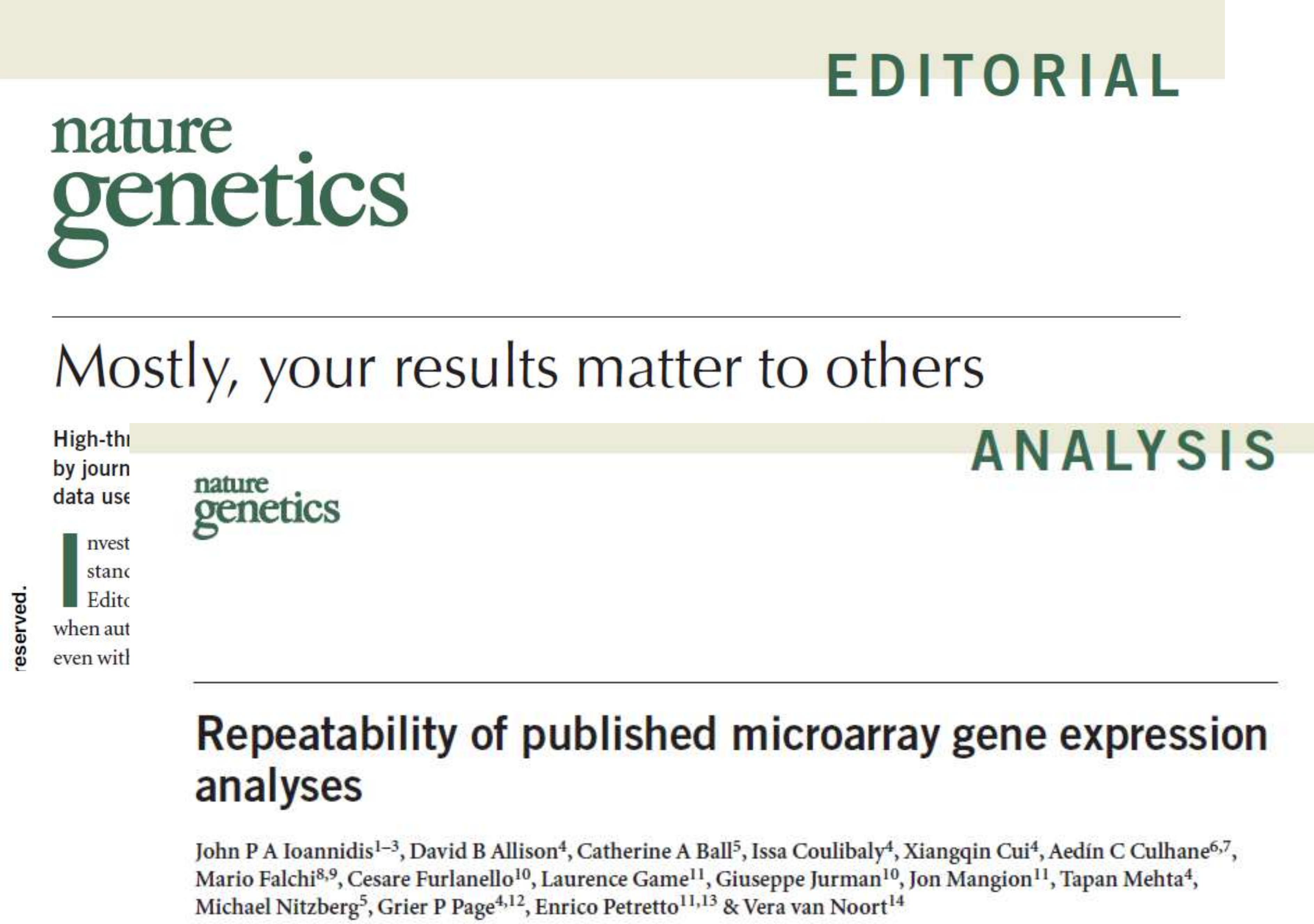
nature genetics

Mostly, your results matter to others

High-throughput datasets and analysis protocols are intrinsically difficult to referee. Community standards enforced by journals may be less effective than is widely appreciated. Greater awareness of the needs and value of secondary data users can result in higher-impact papers.

Investigating the compliance of our publications with MIAME standards (minimum information about a microarray; Editorial, *Nat. Genet.* **38**, 1089; 2006), we found that even when authors and referees are aware of community standards and even with editors mandating both data deposition and accession issue does indeed explain the limits of the analysts' requirements and critical aims.

Why should we consider the utility of rich datasets to researchers whose aim is reanalysis? Many experiments need to start with reanalysis, for validation or comparison. The journal needs to



EDITORIAL

nature genetics

Mostly, your results matter to others

High-thr
by journ
data use

Invest
stanc
Edito
when aut
even witl

reserved.

ANALYSIS

nature genetics

Repeatability of published microarray gene expression analyses

John P A Ioannidis[1–3], David B Allison[4], Catherine A Ball[5], Issa Coulibaly[4], Xiangqin Cui[4], Aedín C Culhane[6,7], Mario Falchi[8,9], Cesare Furlanello[10], Laurence Game[11], Giuseppe Jurman[10], Jon Mangion[11], Tapan Mehta[4], Michael Nitzberg[5], Grier P Page[4,12], Enrico Petretto[11,13] & Vera van Noort[14]

NATURE GENETICS | VOLUME 41 | NUMBER 2 | FEBRUARY 2009

Repeatability of published microarray gene expression analyses

- Selected articles published in *Nature Genetics between* January 2005 and December 2006 that had used profiling with microarrays

- Of the 56 items retrieved electronically, 20 articles were considered potentially eligible for the project

- The four teams were from
  - University of Alabama at Birmingham (UAB)
  - Stanford/Dana-Farber (SD)
  - London (L) and Ioannina/Trento (IT)

- Each team was comprised of 3-6 scientists who worked together to evaluate each article.

Methods

- Two articles were re-analysis of data
- Analysis of 18 articles
- Both cDNA and oligonucleotide array data
- Each randomly allocated to 2 teams

Methods

- Attempted to replicate 1 figure from each article
  - download data
  - Normalize and preprocess data
  - perform analysis described to reproduce figure

- No communications with authors of article permitted

Results

- Result could be reproduced n=2
- Reproduced with discrepancy n=6
- Could not be reproduced n=10
  - No data n=4 (no data n=2, subset n=1, no reporter data n=1)
  - Confusion over matching of data to analysis (n=2)
  - Specialized software required and not available (n=1)m
  - Raw data available but could not be processed n=2

Reproducibility of Analysis

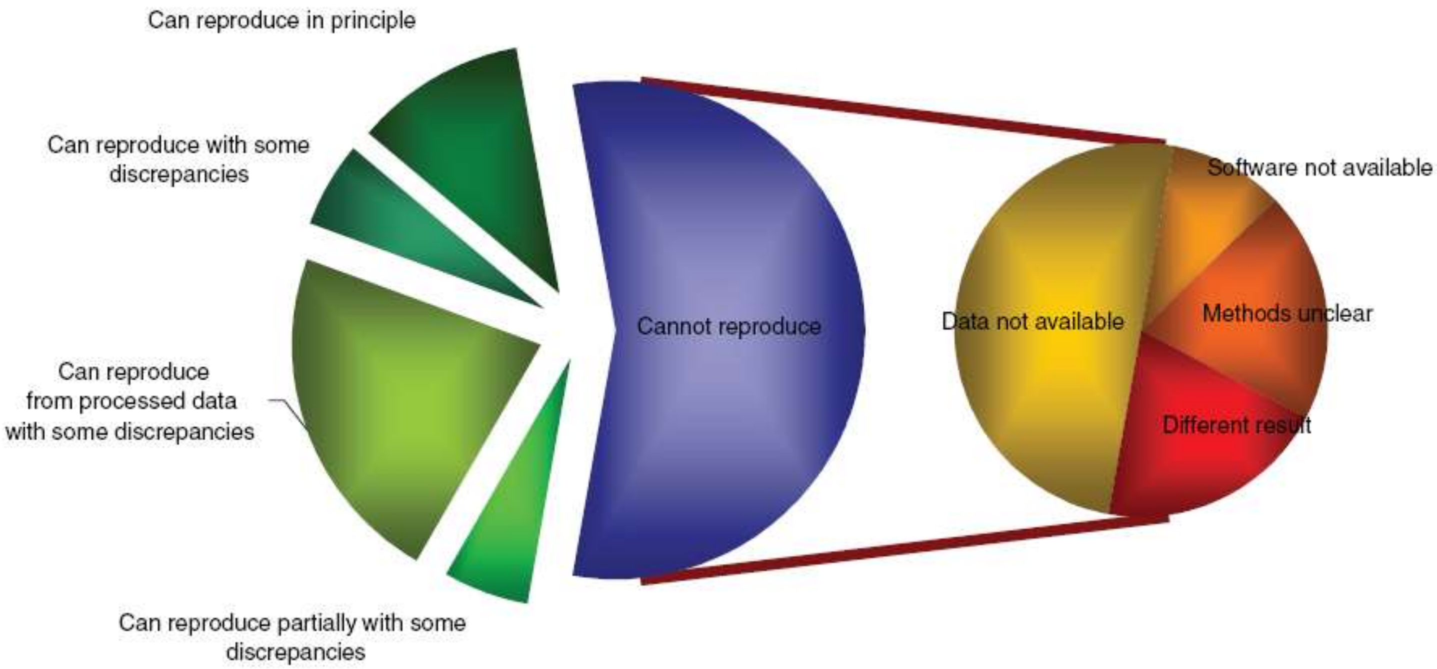
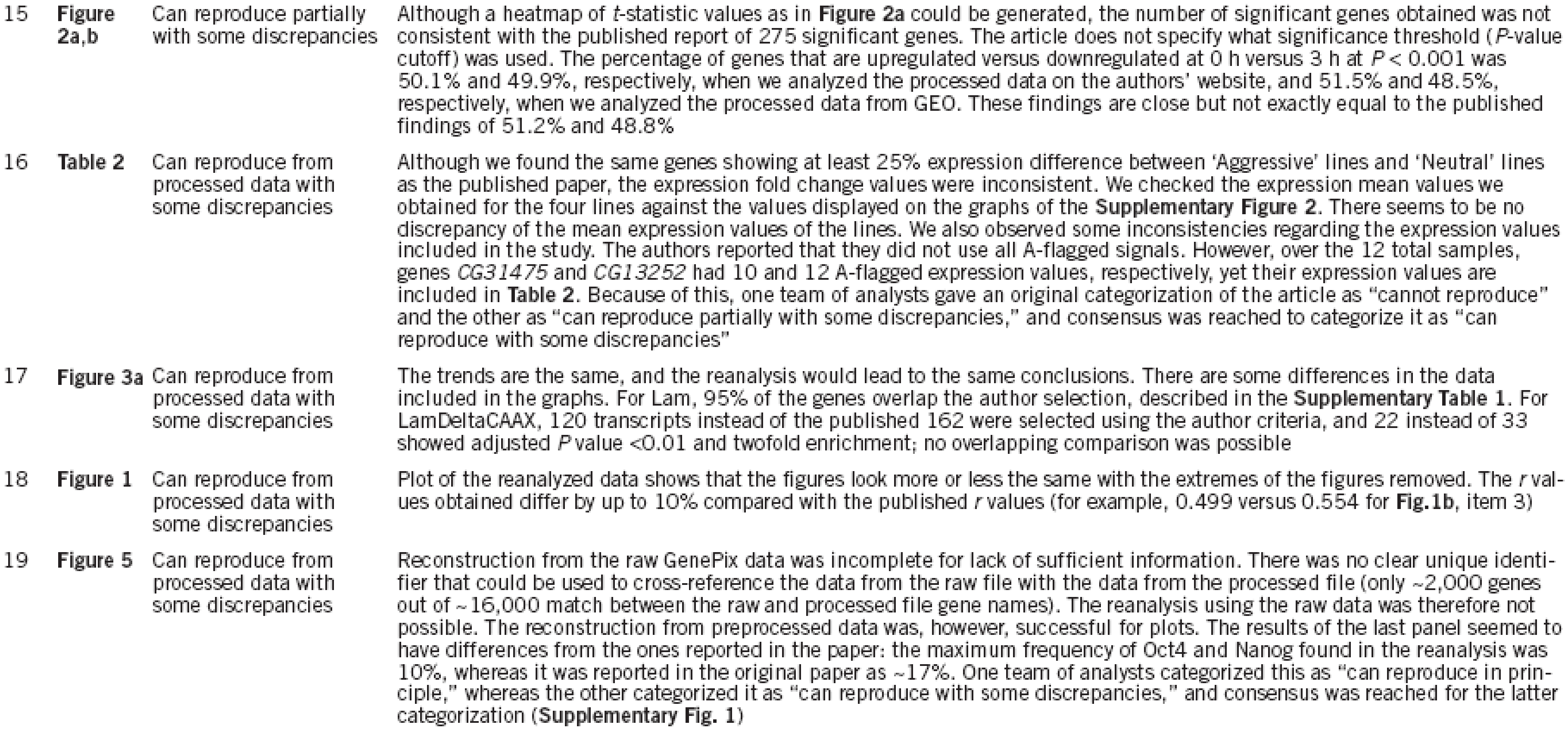


Figure 1 Summary of the efforts to replicate the published analyses.

Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al, (2009) **Repeatability of published microarray gene expression Analyses**. *Nature Genetics* 41(2):149

Problems in reproducing analysis



15	**Figure 2a,b**	Can reproduce partially with some discrepancies	Although a heatmap of *t*-statistic values as in **Figure 2a** could be generated, the number of significant genes obtained was not consistent with the published report of 275 significant genes. The article does not specify what significance threshold (*P*-value cutoff) was used. The percentage of genes that are upregulated versus downregulated at 0 h versus 3 h at $P < 0.001$ was 50.1% and 49.9%, respectively, when we analyzed the processed data on the authors' website, and 51.5% and 48.5%, respectively, when we analyzed the processed data from GEO. These findings are close but not exactly equal to the published findings of 51.2% and 48.8%
16	**Table 2**	Can reproduce from processed data with some discrepancies	Although we found the same genes showing at least 25% expression difference between 'Aggressive' lines and 'Neutral' lines as the published paper, the expression fold change values were inconsistent. We checked the expression mean values we obtained for the four lines against the values displayed on the graphs of the **Supplementary Figure 2**. There seems to be no discrepancy of the mean expression values of the lines. We also observed some inconsistencies regarding the expression values included in the study. The authors reported that they did not use all A-flagged signals. However, over the 12 total samples, genes *CG31475* and *CG13252* had 10 and 12 A-flagged expression values, respectively, yet their expression values are included in **Table 2**. Because of this, one team of analysts gave an original categorization of the article as "cannot reproduce" and the other as "can reproduce partially with some discrepancies," and consensus was reached to categorize it as "can reproduce with some discrepancies"
17	**Figure 3a**	Can reproduce from processed data with some discrepancies	The trends are the same, and the reanalysis would lead to the same conclusions. There are some differences in the data included in the graphs. For Lam, 95% of the genes overlap the author selection, described in the **Supplementary Table 1**. For LamDeltaCAAX, 120 transcripts instead of the published 162 were selected using the author criteria, and 22 instead of 33 showed adjusted *P* value <0.01 and twofold enrichment; no overlapping comparison was possible
18	**Figure 1**	Can reproduce from processed data with some discrepancies	Plot of the reanalyzed data shows that the figures look more or less the same with the extremes of the figures removed. The *r* values obtained differ by up to 10% compared with the published *r* values (for example, 0.499 versus 0.554 for **Fig.1b**, item 3)
19	**Figure 5**	Can reproduce from processed data with some discrepancies	Reconstruction from the raw GenePix data was incomplete for lack of sufficient information. There was no clear unique identifier that could be used to cross-reference the data from the raw file with the data from the processed file (only ~2,000 genes out of ~16,000 match between the raw and processed file gene names). The reanalysis using the raw data was therefore not possible. The reconstruction from preprocessed data was, however, successful for plots. The results of the last panel seemed to have differences from the ones reported in the paper: the maximum frequency of Oct4 and Nanog found in the reanalysis was 10%, whereas it was reported in the original paper as ~17%. One team of analysts categorized this as "can reproduce in principle," whereas the other categorized it as "can reproduce with some discrepancies," and consensus was reached for the latter categorization (**Supplementary Fig. 1**)

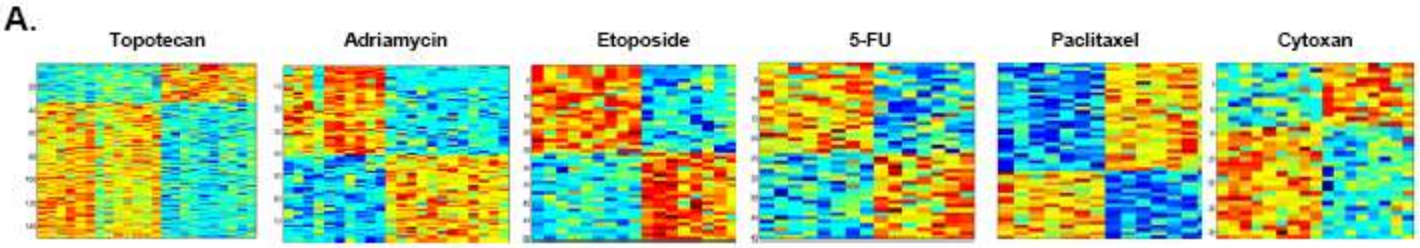
So it this important?

- Well yes…
  - Genomics studies are increasing the basis for new clinical trials and treatment decisions
  - What if there is a problem
  - In review of manuscript reviewers may not spot.. And what if it get published..

The case of
Keith Baggerly vs Anil Potti

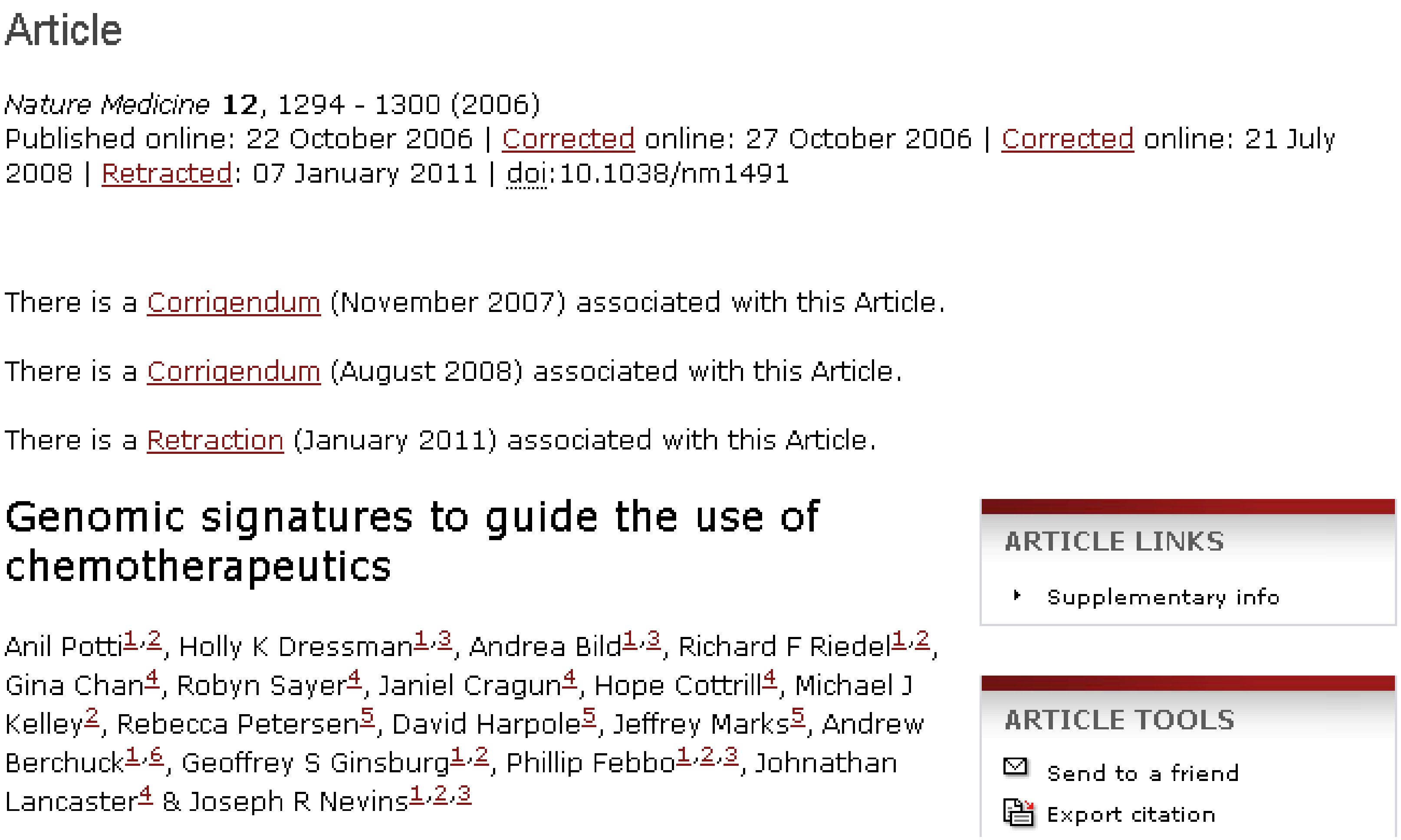
http://www.nytimes.com/2010/07/21/health/research/21cancer.html

- Gene predictor of tumor reponse to chemotherapy



  - http://youtu.be/Q4ZMIUBeLoY

- Baggerly, a statistician at M.D. Anderson in Houston tried to reproduce data from Potti et al., and publicly questioned errors in it
  - http://videolectures.net/cancerbioinformatics2010_baggerly_irrh/



Article

Nature Medicine **12**, 1294 - 1300 (2006)
Published online: 22 October 2006 | Corrected online: 27 October 2006 | Corrected online: 21 July 2008 | Retracted: 07 January 2011 | doi:10.1038/nm1491

There is a Corrigendum (November 2007) associated with this Article.

There is a Corrigendum (August 2008) associated with this Article.

There is a Retraction (January 2011) associated with this Article.

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1,2,3], Johnathan Lancaster[4] & Joseph R Nevins[1,2,3]

ARTICLE LINKS

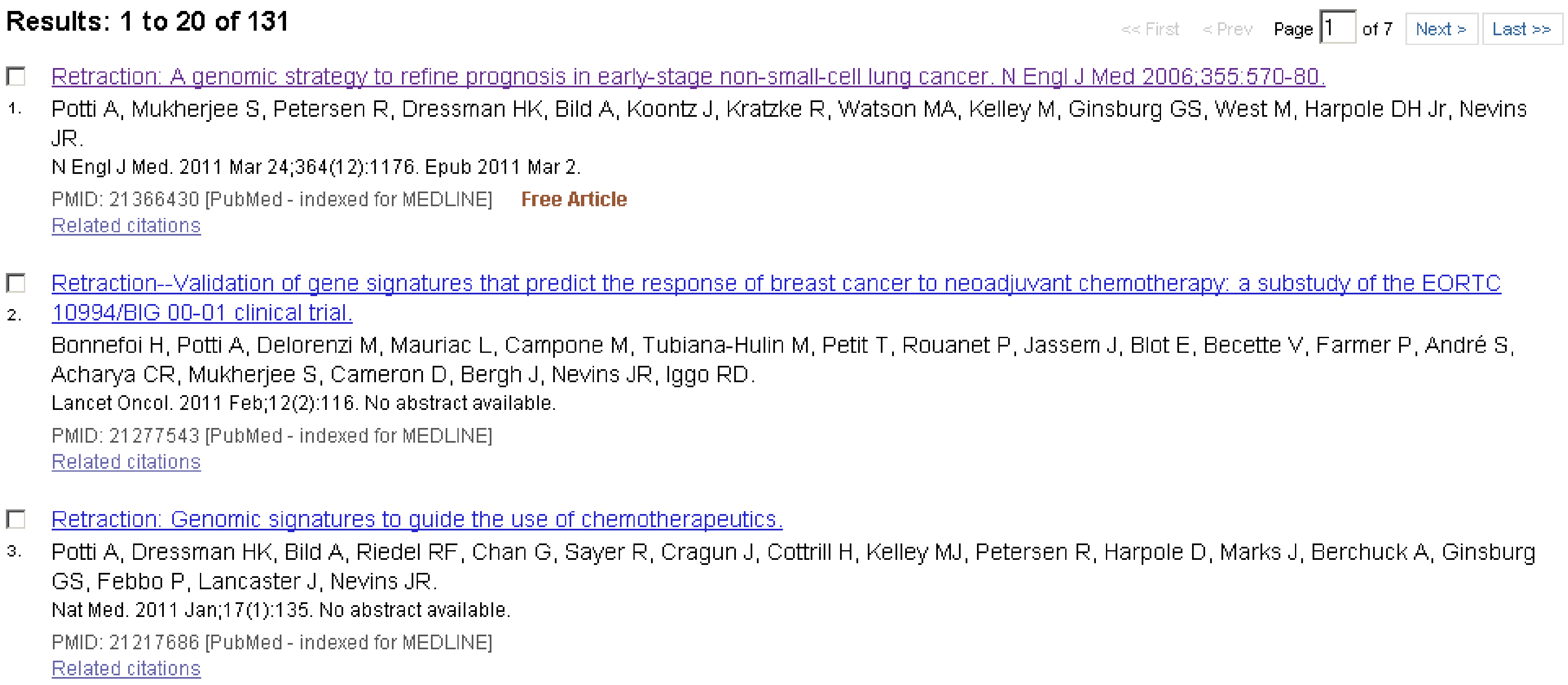
- Supplementary info

ARTICLE TOOLS

Send to a friend

Export citation

Results is his suspension, and suspension of three clinical trials, lung cancer (n=2) and breast cancer



Results: 1 to 20 of 131

□ 1. Retraction: A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. N Engl J Med 2006;355:570-80.
Potti A, Mukherjee S, Petersen R, Dressman HK, Bild A, Koontz J, Kratzke R, Watson MA, Kelley M, Ginsburg GS, West M, Harpole DH Jr, Nevins JR.
N Engl J Med. 2011 Mar 24;364(12):1176. Epub 2011 Mar 2.
PMID: 21366430 [PubMed - indexed for MEDLINE] **Free Article**
Related citations

□ 2. Retraction--Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial.
Bonnefoi H, Potti A, Delorenzi M, Mauriac L, Campone M, Tubiana-Hulin M, Petit T, Rouanet P, Jassem J, Blot E, Becette V, Farmer P, André S, Acharya CR, Mukherjee S, Cameron D, Bergh J, Nevins JR, Iggo RD.
Lancet Oncol. 2011 Feb;12(2):116. No abstract available.
PMID: 21277543 [PubMed - indexed for MEDLINE]
Related citations

□ 3. Retraction: Genomic signatures to guide the use of chemotherapeutics.
Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR.
Nat Med. 2011 Jan;17(1):135. No abstract available.
PMID: 21217686 [PubMed - indexed for MEDLINE]
Related citations

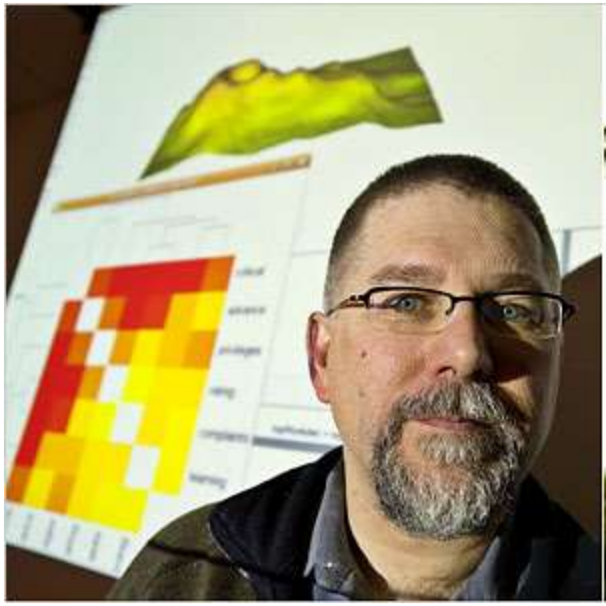
When issues of reproducibility arise

- ``Remember that microarray analysis you did six months ago? We ran a few more arrays. Can you add them to the project and repeat the same analysis?''
- ``The statistical analyst who looked at the data I generated previously is no longer available. Can you get someone else to analyze my new data set using the same methods (and thus producing a report I can expect to understand)?''
- ``Please write/edit the methods sections for the abstract/paper/grant proposal   I am submitting based on the analysis you did several months ago.''

From Keith Baggerly

Why we love Bioconductor and R

- So many methods available
- Easy to view code and get to nuts and bolts
- Leading edge (Methods published in R)
- Great user support through website/mailing list
- Can easily provide code to collaborators or journals
- Reproducible research



- Open source, development- flexible, extensible
- Large number of statistical and numerical methods
- High quality visualization and graphical tools
- Extended by a very large collection of rapidly developing packages
- **Facilitates reproducible research**

The New York Times

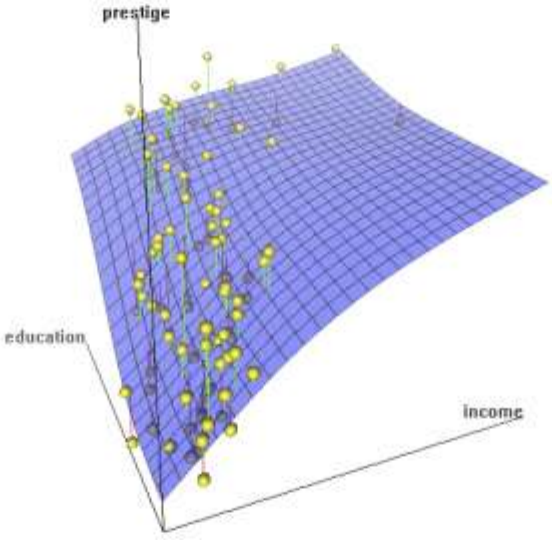
Jan 2009
Data Analysts Captivated by R’s Power

"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.”

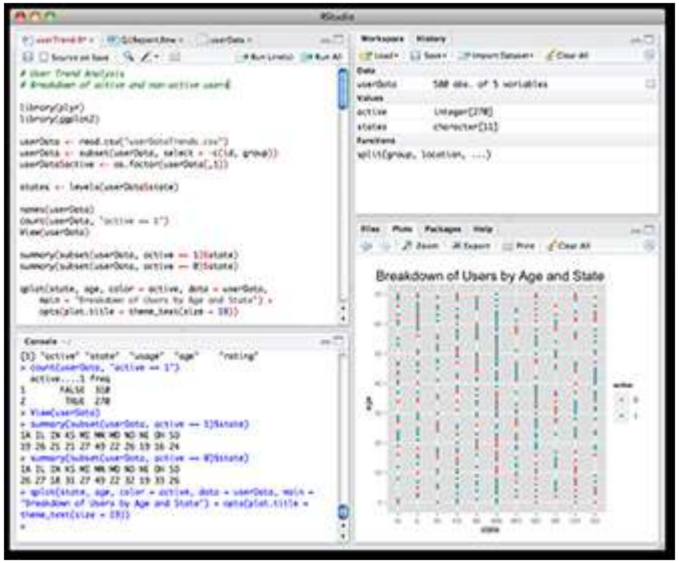
Forbes

Nov 10 2010
Names You Need to Know in 2011: R Data Analysis Software

"R is rapidly augmenting or replacing other statistical analysis packages at universities"



RStudio



```
# User Trend Analysis
# Breakdown of active and non-active users

library(plyr)
library(ggplot2)

userData <- read.csv("userDataTrends.csv")
userData <- subset(userData, select = -c(id, group))
userData$active <- as.factor(userData[,1])

states <- levels(userData$state)

names(userData)
count(userData, "active == 1")
View(userData)

summary(subset(userData, active == 1)$state)
summary(subset(userData, active == 0)$state)

qplot(state, age, color = active, data = userData,
    main = "Breakdown of Users by Age and State") +
    opts(plot.title = theme_text(size = 18))
```

```
Data
userData     500 obs. of 5 variables
Values
active       integer[270]
states       character[11]
Functions
split(group, location, ...)
```

```
[1] "active" "state"  "usage"  "age"    "rating"
> count(userData, "active == 1")
  active....1 freq
1       FALSE  350
2        TRUE  270
> View(userData)
> summary(subset(userData, active == 1)$state)
IA IL IN KS MI MN MO ND NE OH SD
19 26 21 21 27 49 22 26 19 16 24
> summary(subset(userData, active == 0)$state)
IA IL IN KS MI MN MO ND NE OH SD
26 27 18 31 27 49 22 32 19 33 26
> qplot(state, age, color = active, data = userData, main =
"Breakdown of Users by Age and State") + opts(plot.title =
theme_text(size = 18))
>
```

- 4 windows
  -Editor, Console, History, Files/plots
- Code completion
- Easy access to help (F1)
- One step Sweave pdf generation
- Searchable history
- Keyboard Shortcuts
  - http://www.rstudio.org/docs/using/keyboard_shortcuts

- Nice short cut button to build Sweave docs

Sweave

- R embedded in Latex producing pdf or html files in which the R code works!!!

- More help
  - http://www.stat.umn.edu/~charlie/Sweave/
  - http://bioinformatics.mdanderson.org/SweaveTalk/sweaveTalkb.pdf
  - http://www.r-bloggers.com/getting-started-with-sweave-r-latex-eclipse-statet-texlipse/

Where does it come from?

- Where does the name come from?
  - Sweave is to **weave** in S. R is a dialect of S.

- See Friedrich Leisch's (2002) *Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis*

- Literate programming (from Donald Knuth) the author/inventor of TeX with basic idea
  - Programs are useless without descriptions.
  - Descriptions should be *literate*, not comments in code or typical reference manuals.
  - The code in the descriptions *should work*. Thus it is necessary to extract the real working code from the literary description.

- http://www-cs-faculty.stanford.edu/~knuth/taocp.html

Quick Start to Sweave

- Already know R and LaTeX

- Prepare a LaTeX document. Give it an ``Rnw" extension instead of ``tex". Say it is called ``myfile.Rnw'‘
- Insert an R code chunk starting with << >>=
- Terminate the R code chunk with an @ sign followed by a space.
- Run Sweave to produce Tex, and Stangle to extract R code

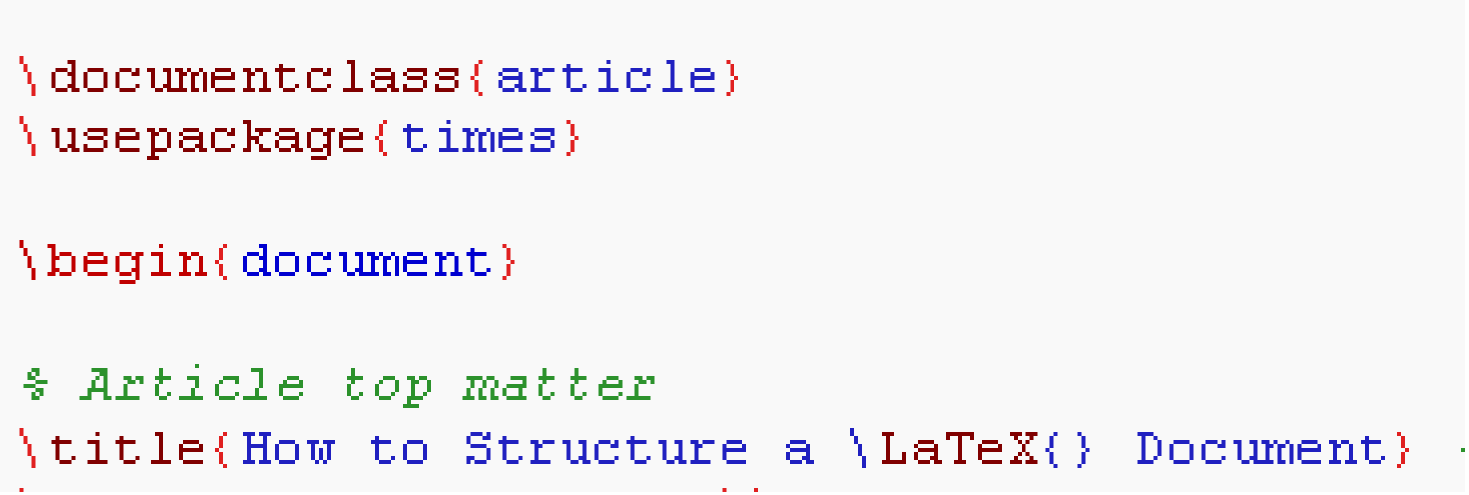
```
<<easySweave>>=
x <- mean(1:10)
print(x)
@
```

Background: Tex and $\LaTeX$

- TeX is a computer program for typesetting documents, created by D. E. Knuth.

- $\LaTeX$, written by L. B. Lamport, is one of a number of `dialects' of TeX. It is particularly suited to the production of articles and books

Typical LaTeX

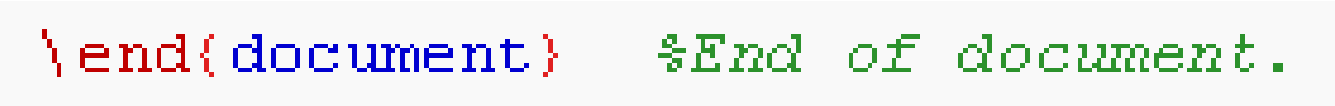


```
\documentclass{article}
\usepackage{times}

\begin{document}

% Article top matter
\title{How to Structure a \LaTeX{} Document}
```

Blah blah blah blah.....



```
\end{document}  %End of document.
```

http://en.wikibooks.org/wiki/LaTeX/simple.tex

Sweave (.Rnw file)

Simple html “dump”

```
library(R2HTML)
HTMLStart(outdir=".", filename="demoR2HTML")
summary(cars)
plot(cars, xlab = "Speed (mph)", ylab = "Stopping
   distance (ft)", las = 1, log = "xy")
HTMLplot(Caption=“a plot”, Height=400)
HTMLStop()
```

./demoR2HTML_main.html

Sweave works in a html document

Create a basic html document and process with Sweave

Sweave(“filename.rnw”, **driver=RweaveHTML)**

```
<html>
<head>
<title>Sweave and html</title>
</head>
<body>

Blah blah

<<SweaveCode>>=
1+2
sum(1:10)
@

blah blah
</html>
```

Exercise

- Use the template on the website and create a .rnw document
- With a title “I Love CSHL and R” and your name
- Create a section called “My Research”, write 2 lines of which describes your lab
- Create a section called “Data and Calculations” which include 3 R code chunks to:
  - Make a vector of random data (n=100, hint: use rnorm)
  - print the mean and sd of your data
  - plot a histogram of the data

Please feel free to contact me

aedin@jimmy.harvard.edu