Reproducible Research in *R* / *Bioconductor*

Martin T. Morgan mtmorgan@fhcrc.org
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

22 November 2013

Reproducibility

Long-term

- Returning to analysis after days, weeks, months of other activity

Multi-participant: communicating with...

- Other statisticians / bioinformaticians
- Biologists and others without specialized statistical knowledge

Science: reproducibility...

- Facilitates third-party verification
- Allows critical assessment
- Challenging, even in high-profile journals requiring archived raw data (Ioannidis *et al.*, 2009, Nat Genet 41: 149-155).

Reproducible Research: Case Study

Original research

- Potti *et al.*, 2006; Hsu *et al.*, 2007
- NCI60 cell line drug sensitivity signature
- Clinical trial allocation

Reproducibility

- Baggerly & Coombes, 2009
- Off-by-one cisplatin gene signature
- Four 'interesting' genes not supported by analysis (two not on array)

References

- Potti *et al.* 2006 Nat Med 12: 1294-1300; (retracted)
- Hsu *et al.* 2007 J Clin Oncol 25: 4350-4357. (retracted)
- Baggerly & Coombes 2009 Ann Appl Stat 3: 1309-1334

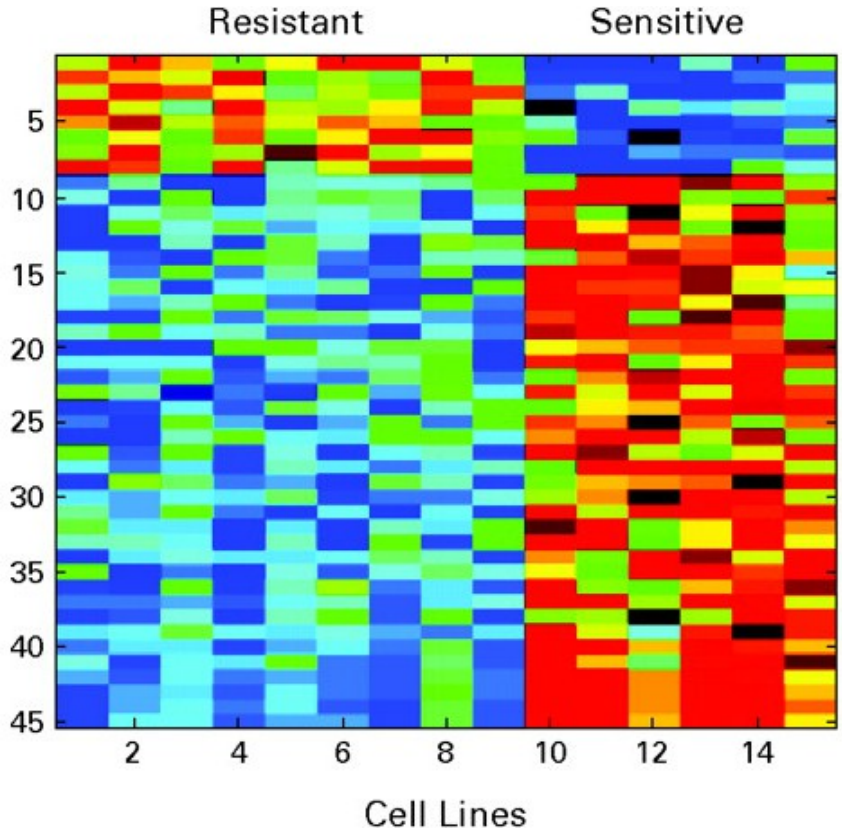
Reproducible Research: Case Study

Original research

- Potti *et al.*, 2006; Hsu *et al.*, 2007
- **NCI60 cell line drug sensitivity signature**
- Clinical trial allocation

Reproducibility

- Baggerly & Coombes, 2009
- Off-by-one cisplatin gene signature
- Four 'interesting' genes not supported by analysis (two not on array)



Hsu *et al.*, cisplatin, fig. 1a

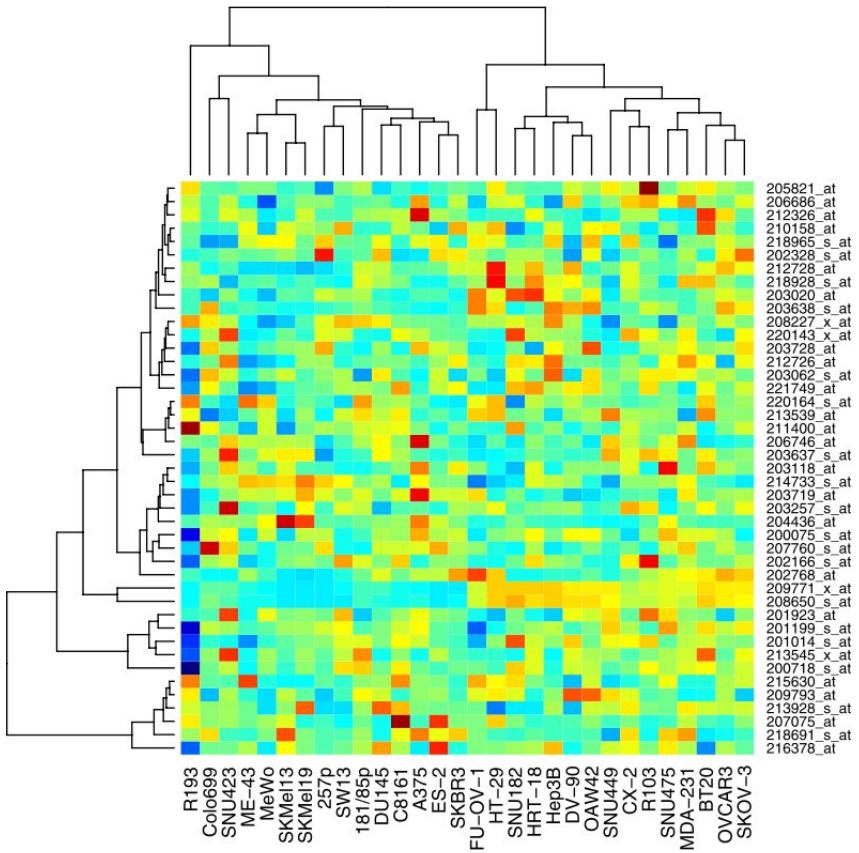
Reproducible Research: Case Study

Original research

- Potti *et al.*, 2006; Hsu *et al.*, 2007
- NCI60 cell line drug sensitivity signature
- Clinical trial allocation

Reproducibility

- Baggerly & Coombes, 2009
- **Off-by-one cisplatin gene signature**
- Four 'interesting' genes not supported by analysis (two not on array)



Baggerly & Coombes, fig. 2a

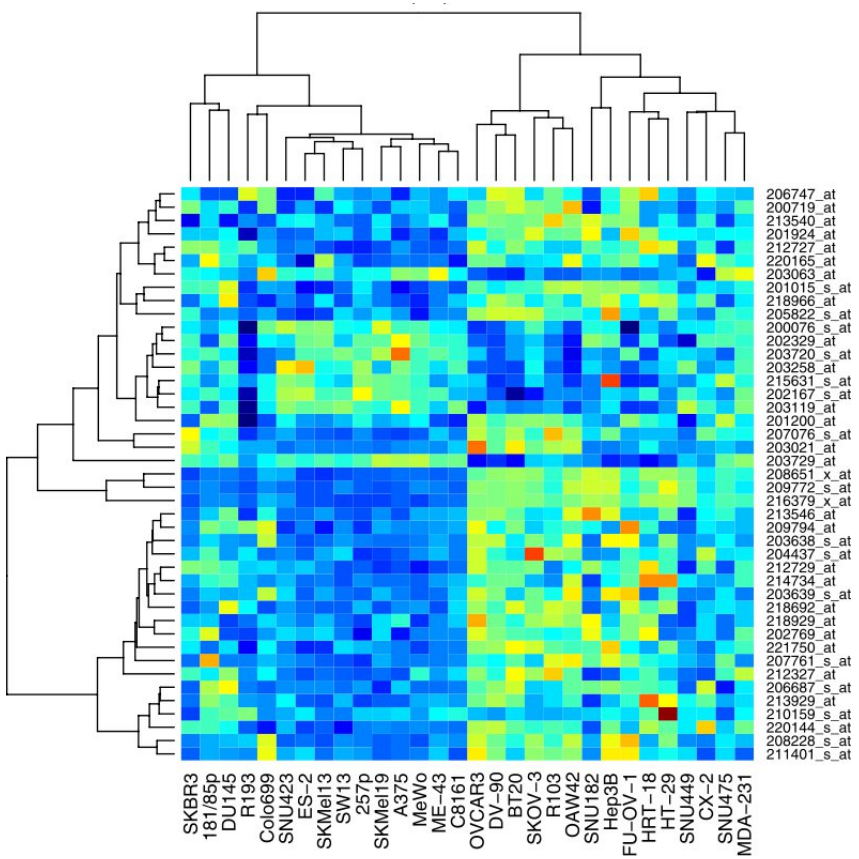
Reproducible Research: Case Study

Original research

- Potti *et al.*, 2006; Hsu *et al.*, 2007
- NCI60 cell line drug sensitivity signature
- Clinical trial allocation

Reproducibility

- Baggerly & Coombes, 2009
- **Off-by-one cisplatin gene signature**
- Four 'interesting' genes not supported by analysis (two not on array)



Baggerly & Coombes, fig. 2b

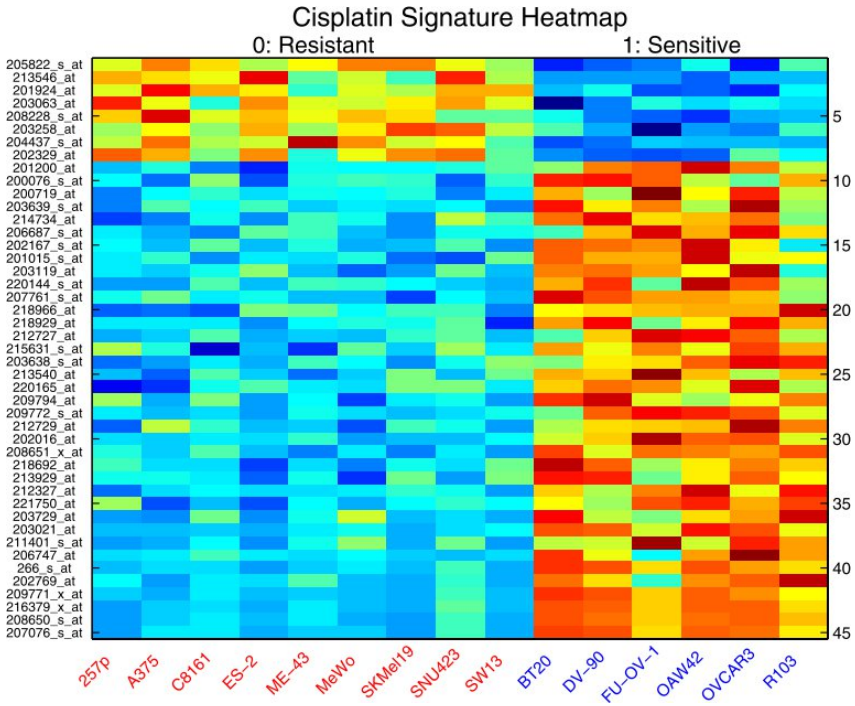
Reproducible Research: Case Study

Original research

- Potti *et al.*, 2006; Hsu *et al.*, 2007
- NCI60 cell line drug sensitivity signature
- Clinical trial allocation

Reproducibility

- Baggerly & Coombes, 2009
- **Off-by-one cisplatin gene signature**
- Four 'interesting' genes not supported by analysis (two not on array)



Baggerly & Coombes, fig. 2d

Reproducible Research: Case Study

Original research

- Potti *et al.*, 2006; Hsu *et al.*, 2007
- NCI60 cell line drug sensitivity signature
- Clinical trial allocation

Reproducibility

- Baggerly & Coombes, 2009
- Off-by-one cisplatin gene signature
- Four 'interesting' genes not supported by analysis (two not on array)

. . . 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 – Baggerly & Coombes, 2009

Reproducibility: *R* / *Bioconductor*

Script-based Data transformations *necessarily* documented

'Literate programming' Text documents embed scripts, scripts *evaluated* when text document processed

Versioned software and repositories Record which package versions used, and retrieve from *Bioconductor* archives

Integrated data containers Sample descriptions and expression data in a single object. Subsetting expression data automatically subsets sample descriptions

Packages Combine code and documentation into a versioned package for archiving and distribution

References

[1] J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort. Repeatability of published microarray gene expression analyses. *Nat. Genet.*, 41(2):149–155, Feb 2009. URL http://dx.doi.org/10.1038/ng.295.

[2] Christopher Gandrud. *Reproducible Research With R and Rstudio*, volume 13. Chapman & Hall/CRC, 2013.

[3] A Morin, J Urban, PD Adams, I Foster, A Sali, D Baker, and P Sliz. Shining light into black boxes. *Science*, 336(6078): 159–160, 2012.