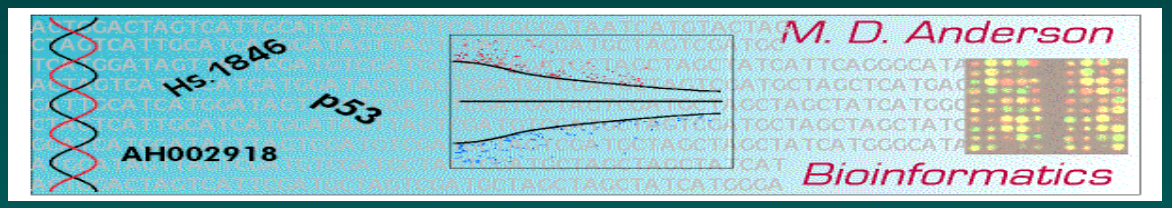
Welcome to Reproducible Research!

Keith A. Baggerly and Roger D. Peng
Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org

SISBID, July 20, 2015



Who We Are

Roger

1. RR editor for *Biostatistics*
2. Instructor for Coursera RR Course

Keith

1. Bioinformatics and Computational Biology
2. Forensic Bioinformatics

Why We're Here

The foundation of *science* is being able to *replicate* someone else's reported results using qualitatively similar methods.

That said, replication can be expensive and/or impractical (esp w big data), so a more workable goal may be to *reproduce* the results reported using the data and code supplied or described.

This is reproducible research (RR)

We're going to try learning new habits, both by increasing awareness (it's important!) and providing tools (it's not that hard!)

What We Want You to Take Away (Awareness)

Why/How practicing RR can help you

Why RR is extremely important with big data, and what can happen when it breaks down

Why the need for RR applies to many areas with big data

Where a lot of big data can be found: e.g., TCGA, TumorPortal, GEO, CCLE

What some of the most common problems are

Why the time to plan for repro is when the project starts

What We Want You to Take Away (Tools)

How to use R/Rstudio/knitr/rmarkdown to produce html/pdf/word reports

How to assemble R packages for sharing, reuse, and publication

How to use git for version control

How to share git repositories with others

How to structure reports to improve clarity and progress

Where to learn more

Some Notable People

John Claerbout - the spiritual father of RR

David Donoho - owner of the best RR quote

Hadley Wickham - developer of some of the best packages (devtools, roxygen2)

Yihui Xie - developer of some of the best packages (knitr, rmarkdown)

JJ Allaire - instigator of RStudio

Where We're Going (Day 1)

Background, semantics, and general habits

RR done wrong

Using R/Rstudio/knitr/rmarkdown

Writing R Packages

RR and Big Data

Where We're Going (Day 2)

What makes replication hard?

Using git

Sharing git/Using github

Doing research right (and reproducibly)

Where We're Going (Day 3)

Writing reports

Describing real data

Some Stuff we Won't Touch

Makefiles

Latex

YAML

Lyx

Other programming languages

Pipelines, NGS data, BAM files

An aside -
while we can't do everything, *we're still learning too*.
If you see ways we could do things better, or things we should mention but haven't, please let us know!

Some Places to Learn More

Karl Broman's Tools for RR Course

Roger's Coursera course and notes (2013)

Christopher Gandrud's book (2e, 2015)

Yihui Xie's book (2e, 2015)

Hadley Wickham's R Packages book (2015)

Some Problems

[Potti et al (2006)]

[Dressman et al (2007)]

[Baggerly and Coombes (2009)]

[Begley and Ellis (2012)]

[Ioannidis et al (2009)]

[McShane IOM testimony (2010)]; see Jan 28, 2011 entry

[Retraction Watch]

[Tabak and Collins (2014)]