

Reproducible research tools to increase efficiency and reduce errors

Dominic LaRoche

May 4, 2015



Outline

- Introduction and Motivation
- The Tools of the Trade
- Examples
- Benefits to You and HTG



Section 1

Introduction



What does it mean to have "reproducible" research?



What does it mean to have "reproducible" research?



What does it mean to have "reproducible" research?

- 1.) Data generating experiment
 - Can the same data be generated again?
 - This is an ongoing topic in the annals of *Nature* and *Science*



What does it mean to have "reproducible" research?

- 1.) Data generating experiment
 - Can the same data be generated again?
 - This is an ongoing topic in the annals of *Nature* and *Science*
- 2.) Can the analysis and results be reproduced given the same data?
 - The focus of my talk today



What does it mean to have "reproducible" research?

- 1.) Data generating experiment
 - Can the same data be generated again?
 - This is an ongoing topic in the annals of *Nature* and *Science*
- 2.) Can the analysis and results be reproduced given the same data?
 - The focus of my talk today
 - Three questions:
 - 1.) Can you reproduce the results right now?
 - 2.) Can you reproduce the results a year from now?
 - 3.) Can someone else reproduce the results in 5 years?



Why Reproducible Research?

This is the alternative!

```
# 🎳 Project
                                                                                                                                             📗 Data
4 M Data
                                                                                                                                             ■ EcoProject
 ► 🌇 CodingSeqs
                                                                                                                                             HPC Scripts
   N ProteinSegs
                                                                                                                                             abylobayes3.31
 ► NecioFiles
                                                                                                                                             I sk_windows
 > 3 SimChainsV4
                                                                                                                                               blackcapProtiens.fasta.fasta
4 🌇 EcoProject
                                                                                                                                             budgeriger
 > BayesTrees.1.3
                                                                                                                                             Desper et al 1999-Inferring Tre
   Bird Papers
                                                                                                                                             ECDFPlot
  # 1 FcoProjectData
                                                                                                                                             COL553 Poster VI
    🕌 BestHitsFiles
                                                                                                                                             Flycatcher
     andGenesFastaMSAV4
                                                                                                                                             14 Hoban et al 2012-Computer si
     ■ ClustsW
                                                                                                                                               LesserWTProteins.fasta
     CompleteReciprocalFastaFilesForAlignment
                                                                                                                                             14 Li and Durbin 2011-Inference of
   > N ECOLpipelineOutput/species
                                                                                                                                             TJ Lynch and Force 2000-The pro
    BCOLpiplineLinux
                                                                                                                                             Male_mallard_duck_2
   > 🎉 ECOLRproject
                                                                                                                                             map-migration-large
   ⊳ 

 Gblocks
                                                                                                                                             Marchini and Howie 2010-Gen
     h1AlignmentsForSim
                                                                                                                                             (III) meanDeltaSSLSdrift
     MMMR.
                                                                                                                                             meanDeltaSSLSdriftTree
   N M HPChkun
                                                                                                                                             migrationCartoon
   > N programsForHPC
                                                                                                                                             Nekamura et al-2004-Bissed b
     RecipBestHitFastaFilesForPipelineV4
                                                                                                                                             new 2
     revisedfastafilesforpipeline
                                                                                                                                             NewTree
    Rpublication
                                                                                                                                             14 Nucl. Acids Res. 2010 - Abascal
   > N SLR
                                                                                                                                             Til Parker et al 2013-Genome-wid
    SpeciesTree
                                                                                                                                             Personing Falcon 5 oboto
  FigTree v1.4.0
                                                                                                                                               phylobayes3.3f.tar
   3 Final Report
                                                                                                                                               phylobayes3,3f,tar.oz
    Monarch Pagers
                                                                                                                                             TI PNAS-2009-Cactor-8986-91-51
I MPC Scripts
                                                                                                                                               ProjectNotes
> 3 phylobayes3.3f
                                                                                                                                              nun sable multiproc
```

de run sable multiproc



Why Reproducible Research?

This is the alternative! # 🎳 Project III Data a 🅌 Data ■ EcoProject ► 🌇 CodingSeqs HPC Scripts N ProteinSegs abylobayes3.31 ► NecioFiles I sk_windows > N SimChainsV4 blackcapProtiens.fasta.fasta 4 🌇 EcoProject budgeriger > BayesTrees.1.3 Desper et al 1999-Inferring Tre Bird Papers ECDFPlot # 1 FcoProjectData @ ECOL553 Poster VI 🕌 BestHitsFiles Flycatcher andGenesFastaMSAV4 14 Hoban et al 2012-Computer si M ClusteW LesserWTProteins.fasta CompleteReciprocalFastaFilesForAlignment 14 Li and Durbin 2011-Inference of > B ECOLpipelineOutput7species Turch and Force 2000-The pro BCOLoiplineLinux Male_mallard_duck_2 > B ECOLRamiect map-migration-large b M Ghlorks Marchini and Howie 2010-Gen h1AlignmentsForSim @ meanDeltaSSLSdrift MMMR. meanDeltaSSLSdriftTree N M HPChkun migrationCartoon > N programsForHPC Nekamura et al-2004-Bissed b RecipBestHitFastaFilesForPipelineV4 fill new 2 revisedfastafilesforpipeline NewTree Roublication 14 Nucl. Acids Res. 2010 - Abascal > N SLR TJ Parker et al 2013-Genome-wid SpeciesTree Personing Falcon 5 oboto FigTree v1.4.0 phylobayes3.3f.tar 3 Final Report phylobayes3.3f.tar.co Monarch Pagers TI PNAS-2009-Cactor-8986-91-51 I MPC Scripts ProjectNotes > 🎉 phylobayes3.3f

I dare you to try and reproduce my results! (or check them for accuracy)



What if I need to make a change?



What if I need to make a change?

- Noticed an error
- Received new or additional data



What if I need to make a change?

- Noticed an error
- Received new or additional data

How many different places do I need to update?



What if I need to make a change?

- Noticed an error
- Received new or additional data

How many different places do I need to update? → Wasted Time



What if I need to make a change?

- Noticed an error
- Received new or additional data

How many different places do I need to update? \rightarrow Wasted Time

What if we need to make a change in a year (or two)?



What if I need to make a change?

- Noticed an error
- Received new or additional data

How many different places do I need to update? \rightarrow Wasted Time

What if we need to make a change in a year (or two)?

What did I do again???



What do we need to do to make analyses reproducible?



What do we need to do to make analyses reproducible?

Organization (where are my car keys?)

What do we need to do to make analyses reproducible?

- Organization (where are my car keys?)
 - Standard folder structure and naming conventions

Introduction

Code–Report pairing



What do we need to do to make analyses reproducible?

- Organization (where are my car keys?)
 - Standard folder structure and naming conventions
 - Code–Report pairing
- Literate Programming
 - Well documented code
 - "Human readable"



What do we need to do to make analyses reproducible?

- Organization (where are my car keys?)
 - Standard folder structure and naming conventions
 - Code–Report pairing
- Literate Programming
 - Well documented code
 - "Human readable"
- Version control
 - Tracked changes from project inception
 - software and package versions controlled



Section 2

The Vision



We can use tools to help with organization and version control



We can use tools to help with organization and version control

- Organization
 - LATEX or Rmarkdown typesetting language + knitr (with R studio)
 - HTG R Package



We can use tools to help with organization and version control

- Organization
 - LATEX or Rmarkdown typesetting language + knitr (with R studio)
 - HTG R Package
- Version control
 - Git
 - Checkpoint

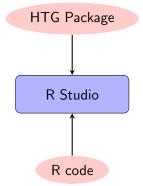


We can use tools to help with organization and version control

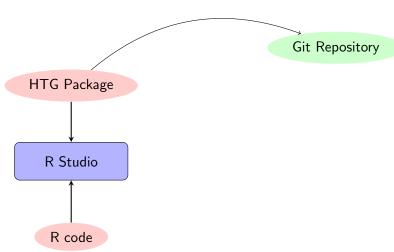
- Organization
 - LATEX or Rmarkdown typesetting language + knitr (with R studio)
 - HTG R Package
- Version control
 - Git
 - Checkpoint

Use these tools together to create a seamless analysis \rightarrow report pipeline.

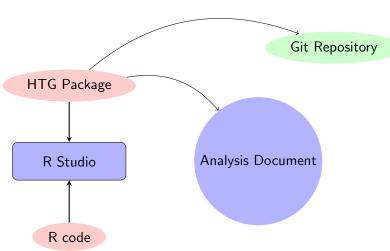




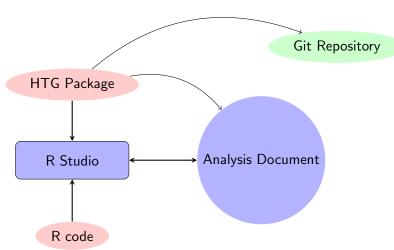




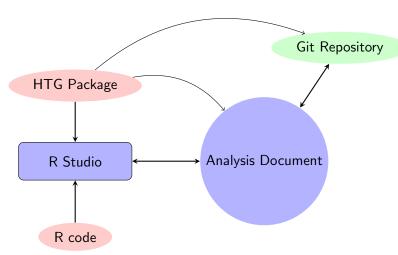




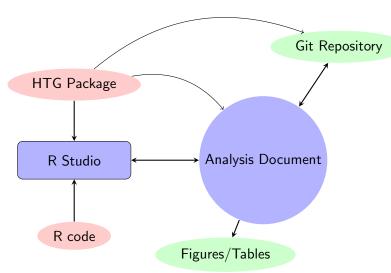




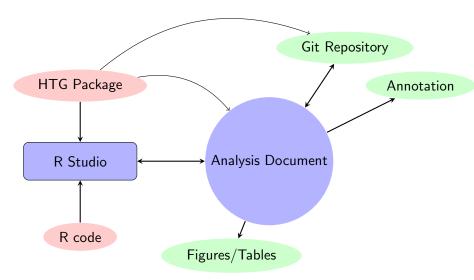




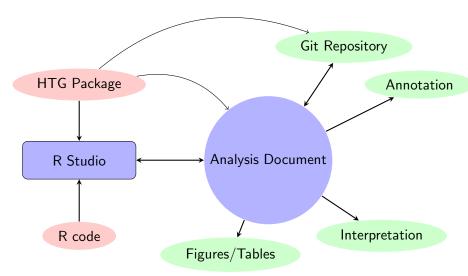




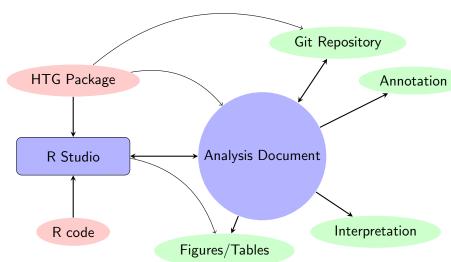














Subsection 1

The Tools



R-Studio

R-Studio is a friendly code editor:

- Built in editor support for:
 - R code
 - C++
 - JavaScript
 - CSS
 - DCF
 - Markdown
 - HTML
 - Tex
 - Python
- It can also handle:
 - PERL
 - SAS ...etc
- Can enable emacs key-bindings



R-Studio

R-Studio combined with LATEXor Rmarkdown can make:

- Reports
- Articles (formatted to most journals)
- Presentations
- Web pages and more...



R-Studio

R-Studio combined with LATEXor Rmarkdown can make:

- Reports
- Articles (formatted to most journals)
- Presentations
- Web pages and more...

(This presentation was made with R-studio)



- Tracks changes to all documents in the folder or subfolders
 - This includes figures, pictures, emails, etc.



- Tracks changes to all documents in the folder or subfolders
 - This includes figures, pictures, emails, etc.
- Can retrieve any previous version of a file



- Tracks changes to all documents in the folder or subfolders
 - This includes figures, pictures, emails, etc.
- · Can retrieve any previous version of a file
- No need to create multiple copies of a document



- Tracks changes to all documents in the folder or subfolders
 - This includes figures, pictures, emails, etc.
- Can retrieve any previous version of a file
- No need to create multiple copies of a document
 - Changes are tracked through commits



- Tracks changes to all documents in the folder or subfolders
 - This includes figures, pictures, emails, etc.
- Can retrieve any previous version of a file
- No need to create multiple copies of a document
 - Changes are tracked through commits
 - Each commit is attached to a (detailed) message about the changes/ status of the document



- Tracks changes to all documents in the folder or subfolders
 - This includes figures, pictures, emails, etc.
- Can retrieve any previous version of a file
- No need to create multiple copies of a document
 - Changes are tracked through commits
 - Each commit is attached to a (detailed) message about the changes/ status of the document
- Can make "clones" so that anyone can get the most current version of everything with a single command



- Tracks changes to all documents in the folder or subfolders
 - This includes figures, pictures, emails, etc.
- Can retrieve any previous version of a file
- No need to create multiple copies of a document
 - Changes are tracked through commits
 - Each commit is attached to a (detailed) message about the changes/ status of the document
- Can make "clones" so that anyone can get the most current version of everything with a single command
- Can specify items you don't want tracked



- Tracks changes to all documents in the folder or subfolders
 - This includes figures, pictures, emails, etc.
- Can retrieve any previous version of a file
- No need to create multiple copies of a document
 - Changes are tracked through commits
 - Each commit is attached to a (detailed) message about the changes/ status of the document
- Can make "clones" so that anyone can get the most current version of everything with a single command
- Can specify items you don't want tracked
- Can "Fork" a project into a new project



Checkpoint is an R package to keep track of package versions

• Git doesn't track changes to the computing environment



Checkpoint is an R package to keep track of package versions

- Git doesn't track changes to the computing environment
- Checkpoint creates a snapshot of all R packages used in the analysis



Checkpoint is an R package to keep track of package versions

- Git doesn't track changes to the computing environment
- Checkpoint creates a snapshot of all R packages used in the analysis
- Package versions are maintained on a separate server (using a git style change tracker).



Checkpoint is an R package to keep track of package versions

- Git doesn't track changes to the computing environment
- Checkpoint creates a snapshot of all R packages used in the analysis
- Package versions are maintained on a separate server (using a git style change tracker).
- Can access the working version of any CRAN package at any time
 - No need to worry that, ahem, someone is using a really old version of the package
 - Can keep your system up-to-date without worrying about breaking an analysis



HTG Private Package

The HTG package eases implementation of reproducible research

- MakeProject()
 - Single command creates version controlled project folder
 - Creates analysis/report templates with either Rmarkdown or LATEX
 - Creates README file for the project to facilitate communication and collaboration
 - Creates Rproject file



HTG Private Package

The HTG package eases implementation of reproducible research

- MakeProject()
 - Single command creates version controlled project folder
 - Creates analysis/report templates with either Rmarkdown or LATEX
 - Creates README file for the project to facilitate communication and collaboration
 - Creates Rproject file
- CloneProject()
 - Clones existing version controlled project to your computer



HTG Private Package

The HTG package eases implementation of reproducible research

- MakeProject()
 - Single command creates version controlled project folder
 - Creates analysis/report templates with either Rmarkdown or LATEX
 - Creates README file for the project to facilitate communication and collaboration
 - Creates Rproject file
- CloneProject()
 - Clones existing version controlled project to your computer
- Can be a place to keep other standard functions



The Glue

R-studio makes it easy to implement these tools

- Weaves code and text together to make reports
 - Works with Rmarkdown or LATEX
 - Rmarkdown lowers the bar for making reports

The Glue

R-studio makes it easy to implement these tools

- Weaves code and text together to make reports
 - Works with Rmarkdown or LATEX
 - Rmarkdown lowers the bar for making reports

The Vision

- Has easy interface with Git
 - Can see change history and associated hashes
 - has simple pull, commit and push cammands etc.



Outside Groups

How do we maintain reproducibility outside our group?



Outside Groups

How do we maintain reproducibility outside our group? Shiny applications

- Provide easy-to-use program to perform a specific task
- Eliminates use of other programs and click-through analyses
- Can incorporate documentation (exported code and report)
- Demo...



Section 3

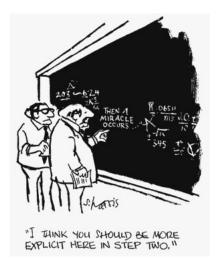
Conclusions



Benefits

- Standardized file structure
- Access to all previous versions of everything
- HTG package eases set-up and saves time
- Anyone can pick up where you left off or find important figures and documents





Questions?