GitHub Introduction

Daniel E. Weeks

Department of Human Genetics
School of Public Health
University of Pittsburgh

Git

- What is Git?
- Why should I use version control?
- How do I use Git?

Git for Humans

https://speakerdeck.com/alicebartlett/git-for-humans

Git

- What is Git?
  - A version control system
- Why should I use version control?
  - Unlimited reversions - travel back in time
  - Branches - can try out changes
  - Back up
  - Collaborate with others

Git: Why should I use version control

- Why should I use version control?
    - Professional skill: Used to coordinate changes when collaboratively developing software.
    - Allows your instructors to see the **same set of files** that you are using.
    - Enables peer review of code using GitHub Issues.
    - Enables easy distribution of slides and example code via a shared GitHub repository.
    - Enables all of us to mutually edit and view Engagement/Feedback Questions.
    - Allows your instructors to efficiently send back comments on your code.
    - Greatly eases year-to-year transfer of lectures and demo code from year to year, much better using Canvas.

Git

- What is Git?
    - A version control system

Version control

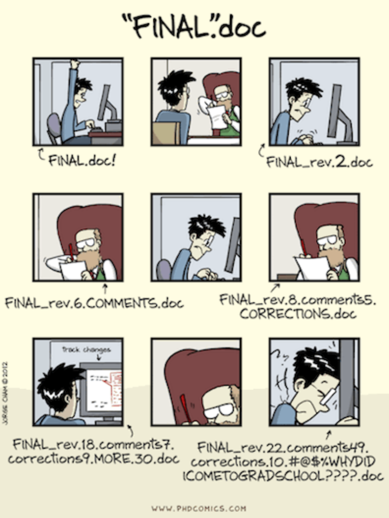
- What is version control?
- How does a version control system work?

What is version control?

Version control is a system that records changes to a file or set of files over time so that you can recall specific versions later.

Slide by Mine Çetinkaya-Rundel "Making your research reproducible: Documenting and automating your data analysis in R" https://github.com/mine-cetinkaya-rundel/2016-01-11-reproducible-research-unc

Bad



Source: Piled Higher and Deeper by Jorge Cham, http://www.phdcomics.com.

Slide by Mine Çetinkaya-Rundel "Making your research reproducible: Documenting and automating your data analysis in R" https://github.com/mine-cetinkaya-rundel/2016-01-11-reproducible-research-unc

Good

```
2013-10-14_manuscriptFish.doc
2013-10-30_manuscriptFish.doc
2013-11-05_manusctiptFish_intitialRyanEdits.doc
2013-11-10_manuscriptFish.doc
2013-11-11_manuscriptFish.doc
2013-11-15_manuscriptFish.doc
2013-11-30_manuscriptFish.doc
2013-12-01_manuscriptFish.doc
2013-12-02_manuscriptFish_PNASsubmitted.doc
2014-01-03_manuscriptFish_PLOSsubmitted.doc
2014-02-15_manuscriptFish_PLOSrevision.doc
2014-03-14_manuscriptFish_PLOSpublished.doc
```

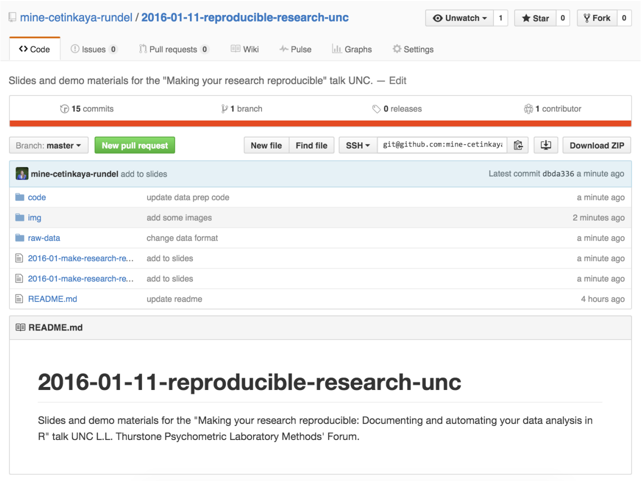
Slide by Mine Çetinkaya-Rundel "Making your research reproducible: Documenting and automating your data analysis in R" https://github.com/mine-cetinkaya-rundel/2016-01-11-reproducible-research-unc

Better - Saving everything together at once

Everytime you make a save, you zip the entire directory that your project files are in and save it with a date.

Slide by Mine Çetinkaya-Rundel “Making your research reproducible: Documenting and automating your data analysis in R” https://github.com/mine-cetinkaya-rundel/2016-01-11-reproducible-research-unc

Best - Version Control



Slides and demo materials for the "Making your research reproducible" talk UNC. — Edit

mine-cetinkaya-rundel add to slides — Latest commit dbda336 a minute ago

code	update data prep code	a minute ago
img	add some images	2 minutes ago
raw-data	change data format	a minute ago
2016-01-make-research-re...	add to slides	a minute ago
2016-01-make-research-re...	add to slides	a minute ago
README.md	update readme	4 hours ago

2016-01-11-reproducible-research-unc

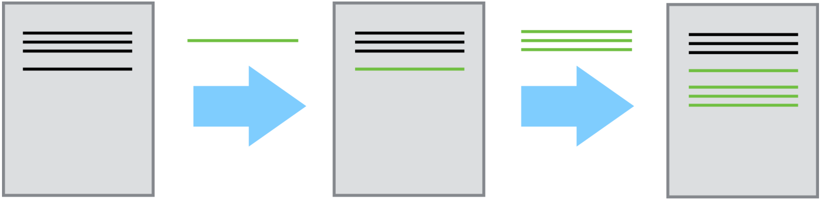
Slides and demo materials for the "Making your research reproducible: Documenting and automating your data analysis in R" talk UNC L.L. Thurstone Psychometric Laboratory Methods' Forum.

Source: https://github.com/mine-cetinkaya-rundel/2016-01-11-reproducible-research-unc/.

Slide by Mine Çetinkaya-Rundel "Making your research reproducible: Documenting and automating your data analysis in R" https://github.com/mine-cetinkaya-rundel/2016-01-11-reproducible-research-unc

How does a version control system work?

- Start with a base version of the document, save just the changes you made at each step of the way.
- Think of it as a tape: if you rewind the tape and start at the base document, then you can play back each change and end up with your latest version.

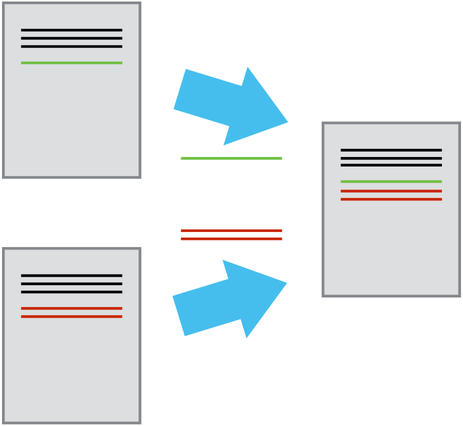


Source: Software Carpentry.

Slide by Mine Çetinkaya-Rundel "Making your research reproducible: Documenting and automating your data analysis in R" https://github.com/mine-cetinkaya-rundel/2016-01-11-reproducible-research-unc

How does a version control system work?

- “Playing back” different sets of changes onto the base document and getting different versions of the document.



Source: Software Carpentry.

Slide by Mine Çetinkaya-Rundel “Making your research reproducible: Documenting and automating your data analysis in R” https://github.com/mine-cetinkaya-rundel/2016-01-11-reproducible-research-unc

Git

- Easy to set up
- Integrated with RStudio
- GitHub's strong community: your colleagues are probably already there
- Provides tools to help enhance collaboration
- A common location to share your work

Slide by Mine Çetinkaya-Rundel "Making your research reproducible: Documenting and automating your data analysis in R" https://github.com/mine-cetinkaya-rundel/2016-01-11-reproducible-research-unc

Version control

- What is version control?
    - A system that keeps track of changes, permitting you to undo them.
    - It permits multiple people to work on the same project.
- How does a version control system work?
    - It records the set of changes from one version to the next.

git with a local repository

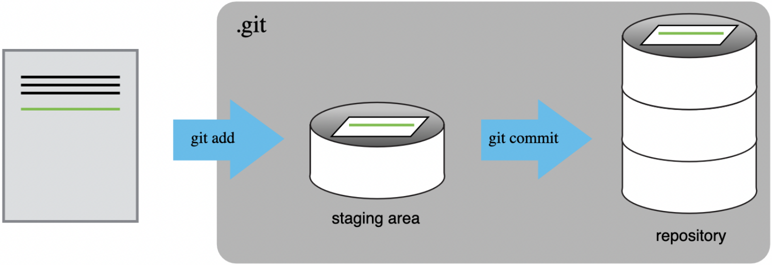


Figure from Ivan Gonzalez; Daisie Huang; Nima Hejazi; Katherine Koziar; Madicken Munk (eds): "Software Carpentry: Version Control with Git." Version 2019.06.1, July 2019, https://github.com/swcarpentry/git-novice, 10.5281/zenodo.3264950 Licensed under CC-BY 4.0 2018–2022 by The Carpentries. Licensed under CC-BY 4.0 2016–2018 by Software Carpentry Foundation.

The git add/commit process

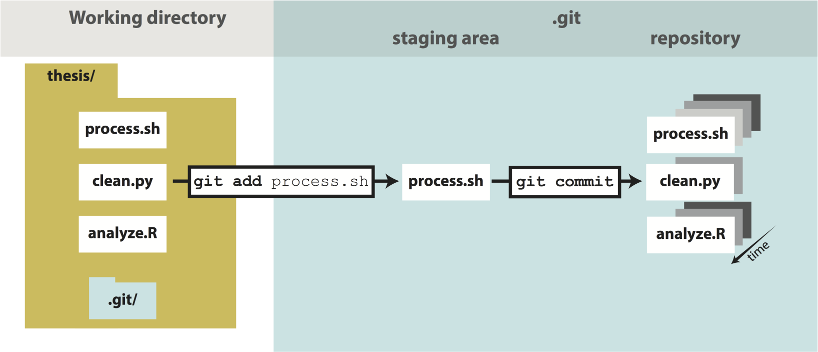


Figure 1 from Blischak JD, Davenport ER, Wilson G. A Quick Introduction to Version Control with Git and GitHub. PLoS Comput Biol. 2016 Jan 19;12(1):e1004668. doi: 10.1371/journal.pcbi.1004668. PMID: 26785377; PMCID: PMC4718703.

Working with a local repository

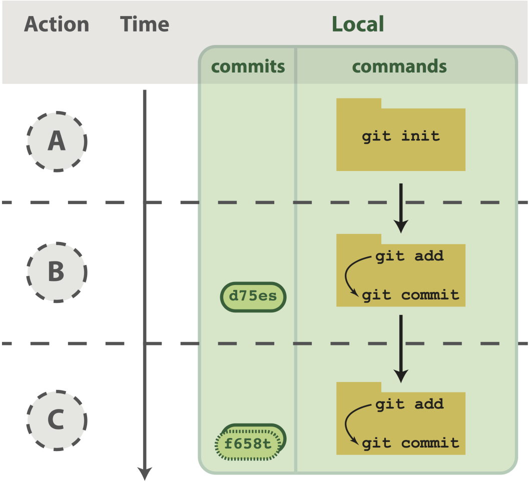


Figure 2 from Blischak et al. PLoS Comput Biol. 2016 Jan 19;12(1):e1004668.

git with a local repository

- `git add` to indicate this is a file you want to keep track of.
    - You only need to add a file once - after being added, git will keep track of subsequent changes to it.
- `git commit` to store your current set of changes in your local repository.
- `git status` to see the current status (list of changed and untracked files).

Working with both a local and remote repository as a single user

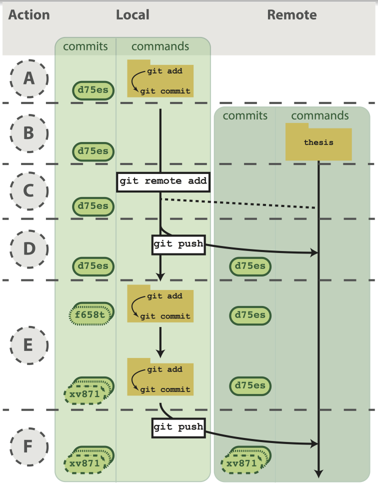


Figure 3 from Blischak et al. PLoS Comput Biol. 2016 Jan 19;12(1):e1004668.

git with both a local and remote repository

- `git pull` fetches the latest changes from the remote repository.
- `git push` pushes your committed changes to the remote repository.

Best to do `git pull`, make changes, `git commit`, `git pull`, and then `git push`.

git - best practice

When working collaboratively on the same repository, do the following:

- Do a `git pull` before starting to make any edits.
- Make your edits,
- `git commit` your changes.
- Do a `git pull` to pull down the very latest remote changes.
- Do a `git push` to push your changes.

git - file system

When you clone a git repository to your computer, it creates a folder containing all the files.

The folder also contains a hidden `.git` folder with internal information git needs to know to connect your folder to the source github.com repository.

Git commands will only work if your working directory is set to be inside of that folder.

RStudio demo

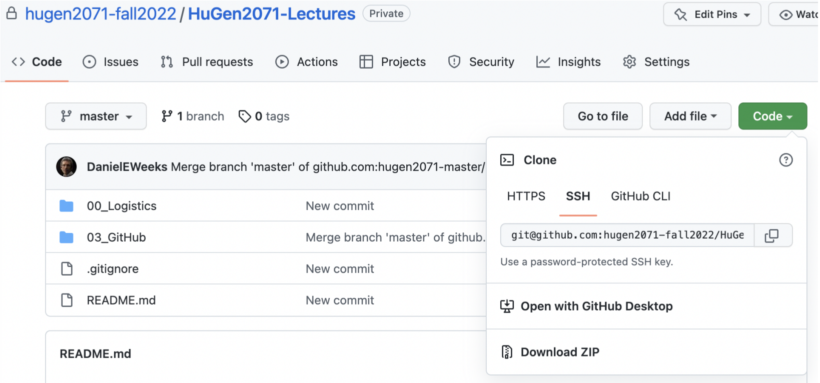
- Setup SSH keys
  - GitHub -> Settings -> SSH and GPG keys -> New SSH key
  - Provide public key to GitHub
  - For detaled instructions, see https://happygitwithr.com/ssh-keys.html
- From the Code drop down in your respoitory, select the SSH link.
- File -> New Project -> Version Control -> Git
  - Pay attention to where your project will be created
  - This is known as 'cloning' and is done only once.
- Make local changes, save, commit
- Push your local changes online to GitHub
  - Be sure to pull, commit, pull, push
- Confirm the local change propagated to the GitHub remote
  - Check at GitHub website

For details, see https://happygitwithr.com/rstudio-git-github.html

RStudio

- Open as a Project
    - Automatically sets the working directory to the project folder

Use the SSH link



hugen2071-fall2022 / HuGen2071-Lectures Private

DanielEWeeks Merge branch 'master' of github.com:hugen2071-master/

00_Logistics	New commit
03_GitHub	Merge branch 'master' of github.
.gitignore	New commit
README.md	New commit

README.md

Clone

```
git@github.com:hugen2071-fall2022/HuGe
```

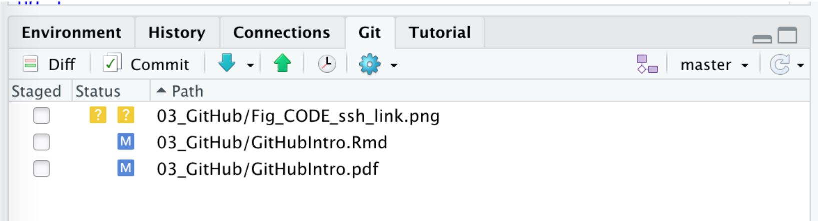
Use a password-protected SSH key.

Do not store on OneDrive or other cloud drives

- Do not clone your repository onto a OneDrive folder, as git does not work properly on cloud drives
    - Cloud drive systems typically maintain their own backup copies ('versions') and this confuses `git`

Using git

- Can use via the command line
  - See https://swcarpentry.github.io/git-novice/
- Easiest to use via RStudio Git window
  - Blue 'M' indicates modified file
  - Yellow '?' indicates a new file not yet under git control



Staged	Status	Path
□	? ?	03_GitHub/Fig_CODE_ssh_link.png
□	M	03_GitHub/GitHubIntro.Rmd
□	M	03_GitHub/GitHubIntro.pdf

Collaborative work on Git

- Dealing with push rejection
  - Always pull before you begin to work.
  - Read the error message
  - https://happygitwithr.com/push-rejected.html
- Pull, but you have local work
  - https://happygitwithr.com/pull-tricky.html
  - Merge conflicts
    - Happens when two people try to change the same lines.
    - Must be manually resolved.
    - https://happygitwithr.com/pull-tricky.html#pull-fetch-and-merge
    - https://docs.github.com/en/github/collaborating-with-issues-and-pull-requests/resolving-a-merge-conflict-using-the-command-line

Git conflict

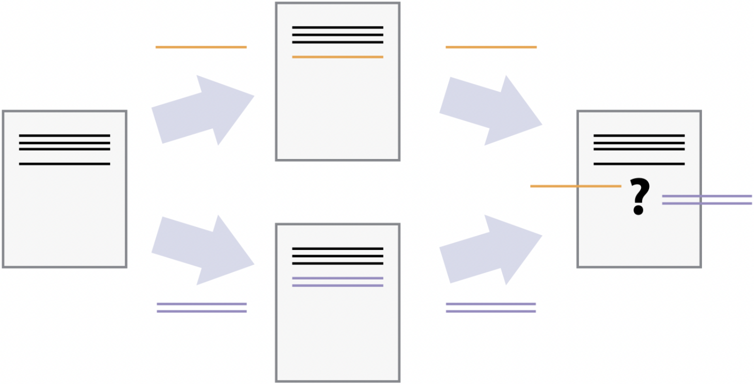


Figure from Ivan Gonzalez; Daisie Huang; Nima Hejazi; Katherine Koziar; Madicken Munk (eds): "Software Carpentry: Version Control with Git." Version 2019.06.1, July 2019, https://github.com/swcarpentry/git-novice, 10.5281/zenodo.3264950 Licensed under CC-BY 4.0 2018–2022 by The Carpentries. Licensed under CC-BY 4.0 2016–2018 by Software Carpentry Foundation.

Selection of files to track

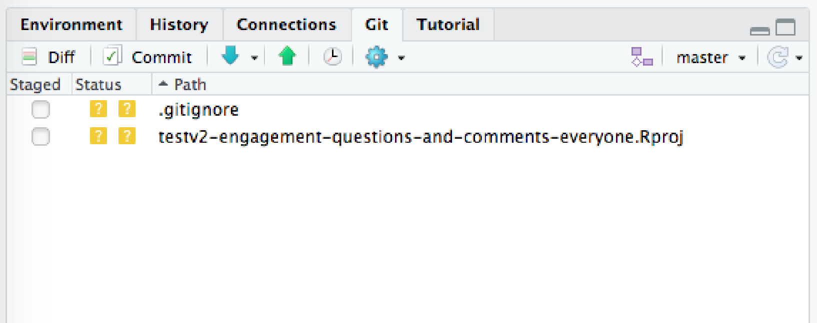
When working collaboratively, think carefully about which files need to be shared via GitHub with your collaborators.

Git is most efficient at tracking changes between different versions of text-based files.

- Do share analysis code
- Don't share static input data
- Don't check in user-specific files
- Don't check in large files
- Don't check in confidential data if using an external remote repository
  - Follow University guidelines about where research data can be stored.

Selection of files to track

When opening a repository on my machine, I saw these untracked files. Neither of these should be checked in.



Staged	Status	Path
□	? ?	.gitignore
□	? ?	testv2-engagement-questions-and-comments-everyone.Rproj

.gitignore

You can exclude files from git's list of modified files by listing their name in the .gitignore file.

Good way to simplify git's list of modified files.

Markdown

- Simple text-based formatting
  - RStudio -> Help -> Markdown Quick Reference
  - https://docs.github.com/en/github/writing-on-github/basic-writing-and-formatting-syntax

Happy Git and GitHub for the useR

https://happygitwithr.com/

git one-time commands

These commands are typically only done once:

- git clone
  - We clone the repository to our computer when we want to start using that repository.
- git add
  - We add a new file to git control.
  - Once it has been added, git will subsequently keep track of all changes to that file.

git - best practices

pull - work - commit - pull - push

- `git pull`
- Make changes
- `git commit` your changes to your local repository
- `git pull` the latest remote changes to your local repository
- `git push` your changes.

Pay attention to any error messages.

GitHub - best practices

- After committing and pushing, it may be a good idea to check the online view of the repository at github.com to verify that your commit was successful.
- Engagement/Feedback questions can be edited within the online github.com interface.

The End

- What questions do you have?