



Versioning of data and code using Git

Introduction to Data Management Practices course **NBIS DM Team**





https://nbisweden.github.io/module-versioning-dm-practices/

Learning Objectives

THIS IS GIT. IT TRACKS COLLABORATIVE WORK ON PROJECTS THROUGH A BEAUTIFUL DISTRIBUTED GRAPH THEORY TREE MODEL.

COOL. HOU DO WE USE IT?

NO IDEA. JUST MEMORIZE THESE SHELL COMMANDS AND TYPE THEM TO SYNC UP. IF YOU GET ERRORS, SAVE YOUR WORK ELSEWHERE, DELETE THE PROJECT, AND DOWNLOAD A FRESH COPY.





- Explain the differences between git and GitHub
- Create your first GitHub project
- Understand how one can collaborate with others using GitHub



Why use versioning for research projects?

Track project history and increase reproducibility

Keeping a record of the changes made ensures that the research findings can be:

- validated
- collaborated upon
- documented accurately







Why use versioning for research projects?



Image from The Turing Way community by Scriberia - <u>CC-BY 4.0</u> DOI: 10.5281/zenodo.3332807

Benefits:

- Store history of changes, **who** changed **what** and **when**
- Go back to a previous version
- Find out when a problem was introduced
- Formalise ways of collaborating
- Useful for many types of files
- Access to same files from different computers lightweight backup





What is version control?



- Automatically creates a robust and rigorous log of changes to a file, without renaming files (v1, v2, *final_copy*)
- **Minimal storage requirements** securely stores only the necessary information to recreate previous versions
- **Prevents user errors** and performs conflict checks
- Coordinate collaborative changes

Image from The Turing Way community by Scriberia - <u>CC-BY 4.0</u> DOI: 10.5281/zenodo.3332807 Slide adapted from: Malvika Sharan. (2020). <u>GitHub-Developing-Collaborative-Document</u>. See DOI: <u>10.5281/zenodo.3835657</u>





Git is one of the most widely used version control systems in the world. It is a free, open source tool you can install locally on your computer

GitHub on the other hand is a **popular website for hosting** and sharing Git repositories remotely. It offers a web interface!

GitHub Desktop is an app on your computer and replaces the use of a terminal window to use git and GitHub



Slide adapted from: Malvika Sharan. (2020). <u>GitHub-Developing-Collaborative-Document</u>. See DOI: <u>10.5281/zenodo.3835657</u>





Make sure public repositories don't contain any secrets or 'sensitive' information.



Image from Mozilla Science Lab: <u>https://mozillascience.github.io/study-group-orientation/3.1-collab-vers-github.html</u> - <u>CC-BY 4.0</u> Slide originates from: Malvika Sharan. (2020). <u>GitHub-Developing-Collaborative-Document</u>. See DOI: <u>10.5281/zenodo.3835657</u> - <u>CC-BY 4.0</u>



-O- Real-time tracking of changes - commits and tags





Image from Mozilla Science Lab: <u>https://mozillascience.github.io/study-group-orientation/3.1-collab-vers-github.html</u> - <u>CC-BY 4.0</u> Slide originates from: Malvika Sharan. (2020). <u>GitHub-Developing-Collaborative-Document</u>. See DOI: <u>10.5281/zenodo.3835657</u> - <u>CC-BY 4.0</u>



Collaborating





Image from Mozilla Science Lab: <u>https://mozillascience.github.io/study-group-orientation/3.1-collab-vers-github.html</u> - <u>CC-BY 4.0</u> Slide originates from: Malvika Sharan. (2020). <u>GitHub-Developing-Collaborative-Document</u>. See DOI: <u>10.5281/zenodo.3835657</u> - <u>CC-BY 4.0</u> -17

وې **Collaborating using branches** main merge chapter_05 Pull Request (PR) Default branch Feature branch 'main' branch 0 0 Merge 'feature' branch into 'main' Create 'feature' branch from 'main' ð Commit changes Discuss proposed changes Submit Pull Request

Image from GitHub Docs https://docs.github.com/en/get-started/start-your-journey/hello-world - CC0 1.0

Navigating the GitHub web interface

- Login to <u>GitHub</u> and explore the <u>nf-core/viralrecon</u> repository
- Change branch from master to dev

 Click on the commits of viralrecon-commits-dev

Code ⊙ Issues (31) 11	Q Type () to sean	Insights	+ • • n @ 👘		☐ (C) nf-core / viralrecon <> Code ⊙ Issues 31 12 Pull requests 4 ⊙ Actions ⊙ Se	Q + • O II A 🛱
🛨 viralrecon Public 🖌	All branches & tags	⊙ Watch 99	*) ♥ Fork 107 * ☆ Star 117 *		Commits	
P dev + P 14 Branches	🛇 11 Tags 🛛 Q. Go to file 🔹	+ <> Code +	About Assembly and intrahost/low-frequency		\$ ² dev -	Rt All users 👻 🗎 All time 👻
This branch is 233 commits ahead of master . Commits		ommits	variant calling for viral samples		-o- Commits on Aug 27, 2024	
😵 Joon-Klaps Merge pull request #439 from nf-c 🚥 🗸 c293787 - 2 weeks ago 🕥 3,021 Commits				Merge pull request #439 from nf-core/bowtie-index-cardinality		
.devcontainer	Template update for nf-core/tools versio	7 months ago	nextflow assembly metagenomics	/ versions	-O- Commits on Aug 26, 2024	🗸 unique
.github	fix input param bowtie2_index name ci	2 weeks ago	amplicon viral ont nf-core		fix bowtie2_index as file	828c524 tD ()
assets	Added header to non filtered BLAST	3 months ago	long-read-sequencing artic covid-19		😵 Joon-Klaps committed 2 weeks ago - 🗸 25 / 25	G- (7
🛅 bin	Merge branch 'dev' into custom_gtf	5 months ago	covid19 sars-cov-2			
conf	Added header to non filtered BLAST	3 months ago	Readme MIT license			



Descrite

Glossary for Git & GitHub



repository/repo: a collection of documents related to your project, contains all data and history (commits, branches, tags)

cloning: copying the whole repository to your laptop - the first time

-O- commit: a saved change/version of the project, gets a unique identifier (commit hash)



tag: a pointer to one commit, to be able to refer to it later. Like a sticky note that you attach to a particular commit (e.g. *phd-printed* or *paper-submitted*)



Glossary for Git & GitHub

branch: a copy of a repo that is contained within the original repo. Branches are used to work on different project features without altering the original or "main" repo.



pull request (PR): a request to merge a commit or collection of commits to a repository



Quiz

Which of the following statements are correct concerning version control?

- 1. Version control serves as a standalone backup system.
- 2. Version control systems makes it easier to collaborate when working with the same file.
- 3. Using a version control system, I don't have to wonder which version of the manuscript was sent to publisher.
- 4. Git is exclusively designed for managing software development and cannot track changes in other types of files like documents or data.
- 5. GitHub Desktop is a version control system.
- 6. Cloning a repository in GitHub copies the repository to your local machine.



Quiz - solution

- No, version control keeps track of different versions, but your work still needs to be backed up. Relying on only the remote repository at GitHub, does not comply with the 3-2-1 backup rule of having at least 3 copies, on 2 different types of digital media, and at least 1 separate location.
- 2. Yes, collaborations, and resolution of conflicts when editing the same file, is much easier with a version control system.
- 3. Well no, all versions are safely kept, but unless you put a (name) tag on a specific version, the system will not be able to help you on which version was sent to publisher.
- 4. No, Git started out as a software versioning control system, but can track a lot of other types of data as well.
- 5. No, GitHub Desktop is an app that provides a graphical user interface to the version control system Git, by communicating with the web service GitHub.
- 6. Yes, the cloned repository is stored on your local computer and connected to the remote repository on GitHub. Changes made locally can then be transferred by pushing them to the remote repository and in the other direction by pulling commits.

