

Sham's intro to Git/GitHub

Disclaimer:

This is non-exhaustive, and I'm not claiming any expertise.
I'm also assuming no one else is an expert.

Git vs GitHub



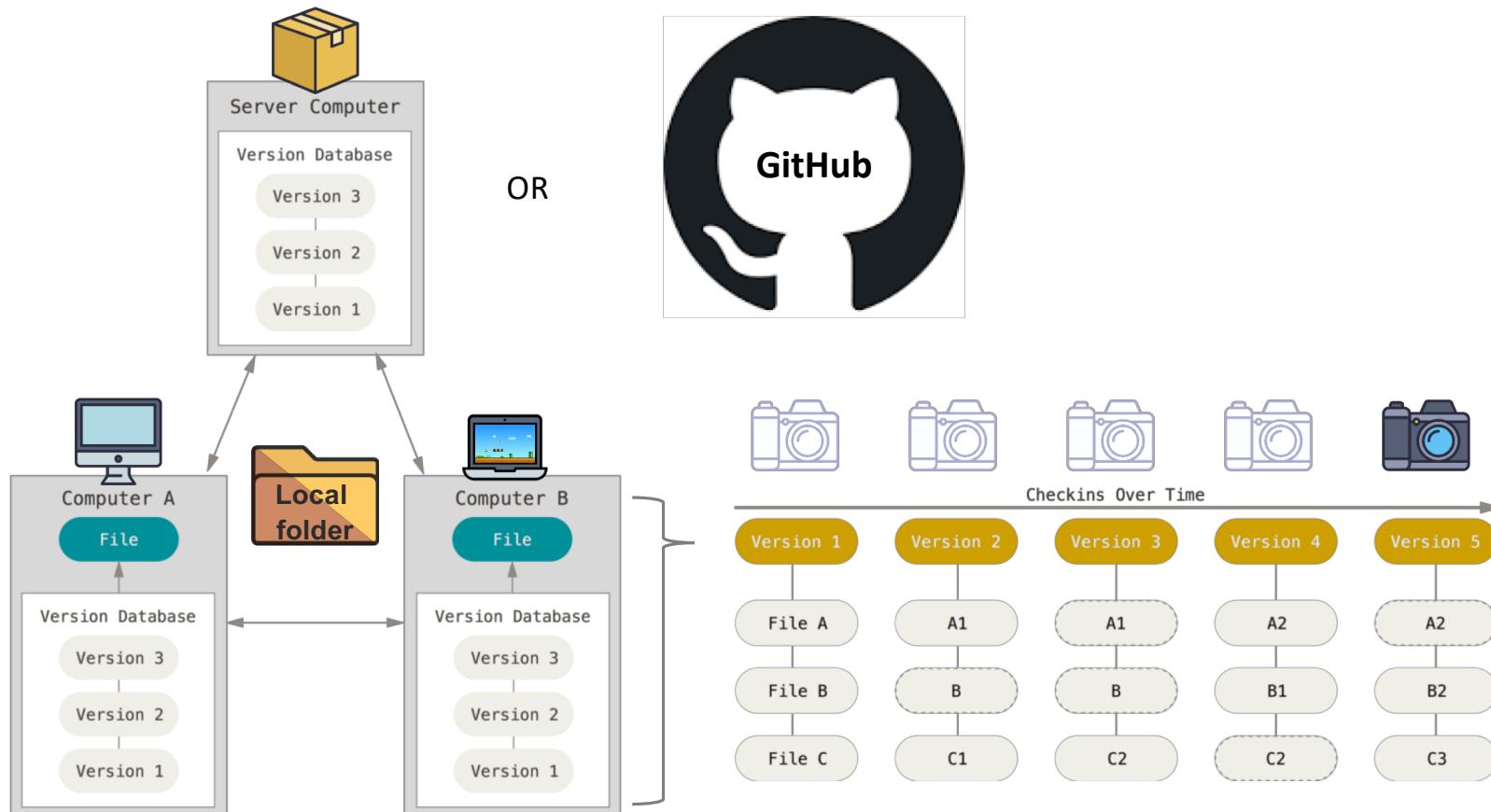
- Version control system for recording changes and coordinating collaborative code
- Parses **text** files for changes – format is important
- Intentional saving – not automated backup
- Requires organization of projects into separate folders
- Macs already have git



- Web host service that uses git version control systems
- Particularly useful for publicly sharing code
- Academic accounts (.edu) can be private

You can use Git without GitHub, but I've been using GitHub

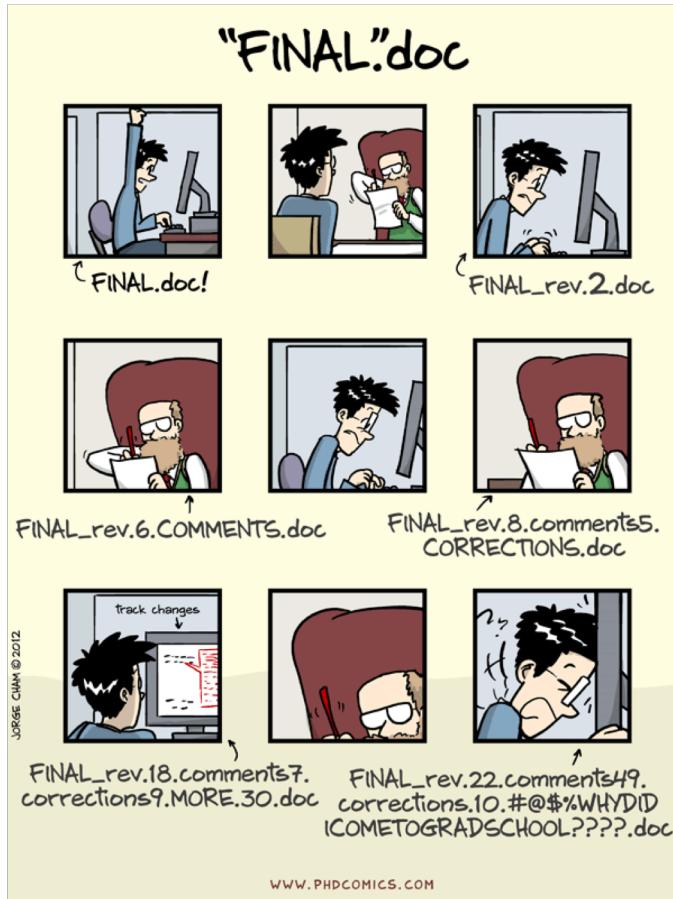
How git stores your files



<https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control>

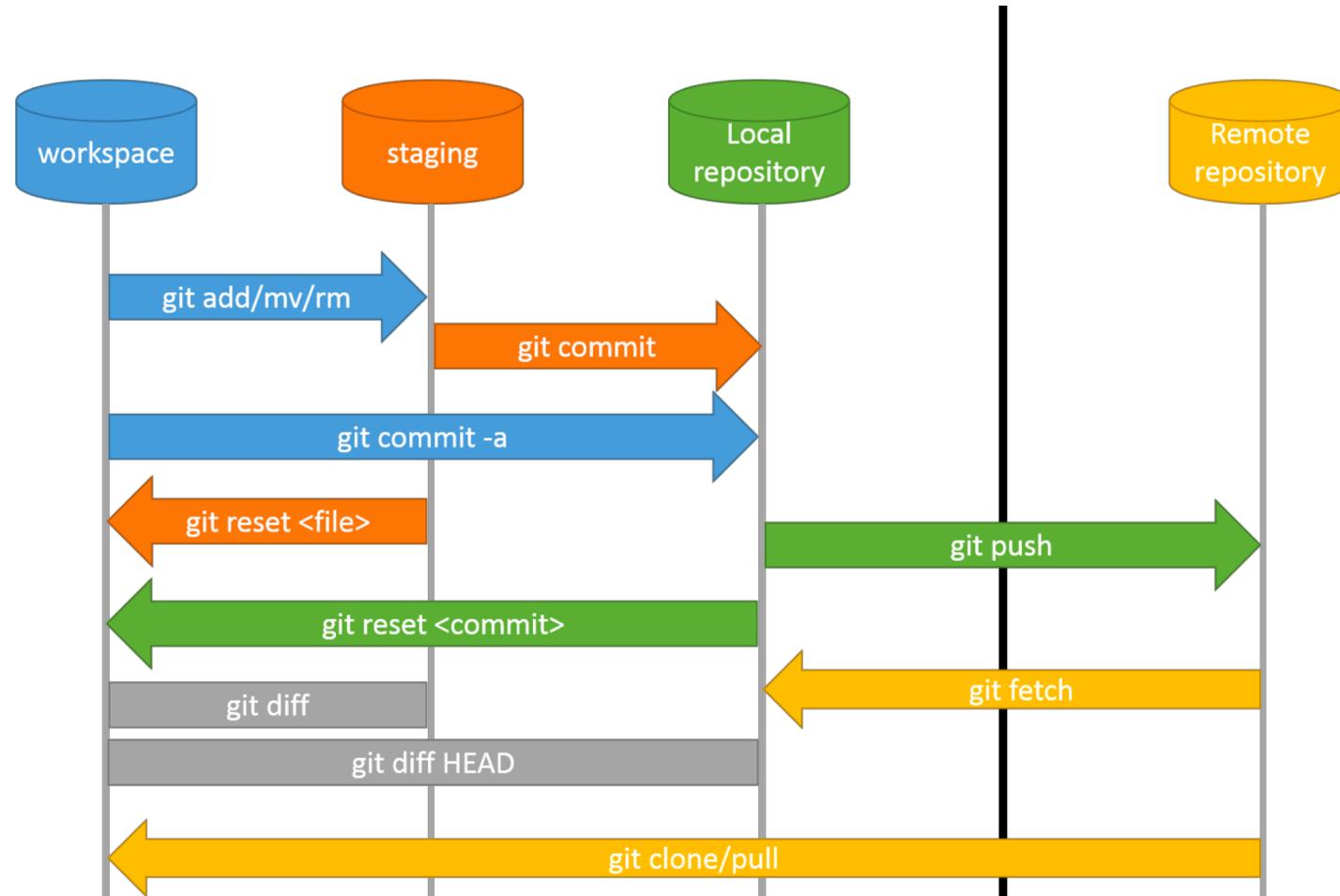
T. K. Phung

Why? To avoid this.....



R	01_18_19_dataset7.R	R Source File
R	04_08_19_TCGA_LCIS.R	R Source File
R	04_23_BCC_splitanalysis.R	R Source File
R	5geneValidation.R	R Source File
R	2018_09_19_BCC2_2_RNAseq.R	R Source File
R	2018_09_20_PAM50_qPCR_BCC2.R	R Source File
R	2018_09_27_deconvolution.R	R Source File
R	2018_10_02_BCC5.R	R Source File
R	2018_10_12_GenesDylan.R	R Source File
R	2019_01_04_PAM50_cluster.R	R Source File
R	2019_01_09_BCC_TCGA_cluster_PAM50.R	R Source File
R	2019_01_10_BCC_hetgenes.R	R Source File
R	2019_01_13_ercc.R	R Source File
R	2019_01_16_ercc.R	R Source File
R	2019_01_16_TCGA_NMF.R	R Source File
R	2019_01_19_BCC_PAM50_resampling.R	R Source File
R	2019_02_28 TPM_Hetgenes.R	R Source File
R	2019_04_11_scde_overdispersion.R	R Source File
R	al4ng-dplyr-hw.R	R Source File
R	al4ng-ggplot2-hw.R	R Source File
R	al4ng-stats-hw.R	R Source File
R	alignment_rate.R	R Source File
R	alignment_rate.R	R Source File

General overview of how it works



Start a web repository:

1. On GitHub's website:

Create a new repository

A repository contains all project files, including the revision history. Already have a project repository elsewhere? [Import a repository](#).

Owner / Repository name * 

Great repository names Your new repository will be created as **pam50**- now about **symmetrical-computing-machine**?

Description (optional)

 **Public**
Anyone can see this repository. You choose who can commit.

 **Private**
You choose who can see and commit to this repository.

Initialize this repository with a **README**
This will let you immediately clone the repository to your computer. Skip this step if you're importing an existing repository.

Add .gitignore: **None** | Add a license: **None** 

Create repository

Initiate local version control workflow



Install/see if you have git `git --version`

Your local project folder
where you store this code



Your local project folder `cd /directory`

Initialize repo `git init`

Set web repository as `git remote add origin
master https://github.com/shamdata/pam50`

Add all files/changing `git add "Code_of_interest.R"`

Commit snapshot `git commit -m "Initial commit"`

Push snapshot to the `git push
master`

**NOTE: you can add EVERY file
in the directory if you want
or the script of interest like in
this example**

History on GitHub:

History for [pam50](#) / [2019_04_08_PAM50_JP.R](#)

- Commits on Apr 25, 2019
 - [tried modifying max filter to range filter to improve outlier behavio...](#) [...](#) [!\[\]\(a3db96e7bf6d55f6fe23107dcf098618_img.jpg\) shamdata committed 4 hours ago](#) [!\[\]\(89d1e09f668245d223896beda39443bd_img.jpg\)](#) [b71b4f1](#) [!\[\]\(5fff40472ad10d456ceb3775edf98340_img.jpg\)](#)
 - [tried a max filter to remove outlier behavior](#) [...](#) [!\[\]\(207d79912909f888f5f071263296f6ac_img.jpg\) shamdata committed 6 hours ago](#) [!\[\]\(a8d88ebf628da168cd51d7459954230e_img.jpg\)](#) [ccda189](#) [!\[\]\(176df70e19313d8dc4b65cd4d681dc4c_img.jpg\)](#)
- Commits on Apr 22, 2019
 - [re-normalized columnwise to allow clustering with TCGA data](#) [...](#) [!\[\]\(04a646b4a04129f475fedd273f7cea2f_img.jpg\) shamdata committed 3 days ago](#) [!\[\]\(f0d47e198adc0c688c8faebdc9bc003d_img.jpg\)](#) [ee97423](#) [!\[\]\(38be0354a9847e686b91de18422d82c2_img.jpg\)](#)
- Commits on Apr 18, 2019
 - [I think Ray's samples are still going to cluster separately, attempte...](#) [...](#) [!\[\]\(43bb564f8e64338559c3fc8e162587a3_img.jpg\) shamdata committed 7 days ago](#) [!\[\]\(e9101f8641b4cf9d570627dddc2e34a2_img.jpg\)](#) [fc3206b](#) [!\[\]\(4683f978cc26d2db4d87aa998af991ec_img.jpg\)](#)
 - [Got TCGA and Ray's data merged on a common set of ~15000 genes that h...](#) [...](#) [!\[\]\(3fcb0854e266c8bf23157da07f5f0d36_img.jpg\) shamdata committed 7 days ago](#) [!\[\]\(48cab3555f72a3f8cc252ceeed6898bc_img.jpg\)](#) [bce51e1](#) [!\[\]\(30c29c7cd27c9d544ed08c8ed3c5f082_img.jpg\)](#)
- Commits on Apr 17, 2019
 - [Lots of TCGA/Ray's data ortholog finagling....](#) [...](#) [!\[\]\(05a3bb401b1d75d8ac535579d8e7616c_img.jpg\) shamdata committed 8 days ago](#) [!\[\]\(9eea409ffe051e787b93cb5e6798f6d1_img.jpg\)](#) [9d14701](#) [!\[\]\(30589c18588cfffba1d708d162cb9ea8_img.jpg\)](#)
- Commits on Apr 16, 2019
 - [Got it running with Ray's data, working on how to get to best fit tra...](#) [...](#) [!\[\]\(746953eeba4585727219973d4cfa92ea_img.jpg\) shamdata committed 9 days ago](#) [!\[\]\(af39c9aa5018af426a880c9cc62a3057_img.jpg\)](#) [4cd498f](#) [!\[\]\(807e82dbbb47295fe1d23f6ee917c140_img.jpg\)](#)
 - [Last minute trying to get Ray's new data PAM50 classifications](#) [...](#) [!\[\]\(3da8c42da2069c76264bb821673cabdf_img.jpg\) shamdata committed 9 days ago](#) [!\[\]\(7fe1b1a049c248e14f51c18ff75dce06_img.jpg\)](#) [e5af9bf](#) [!\[\]\(860123247a7af3e824fe9e21c30da8a6_img.jpg\)](#)

Showing 1 changed file with 26 additions and 8 deletions.

Unified Split

```
v 34 2019_04_08_PAM50_JP.R
@@ -19,38 +19,54 @@ data <- read.delim('gene TPM_all_samples.tsv',sep='\t',stringsAsFactors = F)
19 row.names(data) <- data$Gene_ID
20 ensgtoid <- read.delim('ENSMUSG_ID2Name.txt',sep='\t',header=F,stringsAsFactors = F)
21 row.names(ensgtoid) <- ensgtoid$V1
22 -data.id <- cbind(data,ensgtoid)

23
24 #log2 transform

25 data_norm <- data.id
26 data_norm[,index] <- log2(data_norm[,index]+1)
27

28 #genelist
29 pam50 <- read.csv("/Users/ss4cf/Desktop/Lab/Data/TCGA/PAM50/pam50.csv",header=TRUE,stringsAsFactors = F)
30 row.names(pam50) <- pam50$x

31 orthologs <- read.delim("mart_export.txt",sep='\t',stringsAsFactors = F)
32 pam50mouse <- orthologs[which(orthologs$Gene.name %in% pam50$x),c(3,5)]
33 unique_human <- unique(pam50mouse$Gene.name)
34 unique_mouse <- unique(pam50mouse$Mouse.gene.name)
35 unique_mouse <- unique_mouse[-c(23,25,26,39)]
36 pam50mouse <- data.frame(cbind(unique_human,unique_mouse))
37 row.names(pam50mouse) <- pam50mouse$unique_mouse

38
39 -#NOW, get PAM50 expression in ALL TCGA samples and correct with this median
40 data.pam50 <- data_norm[which(data_norm$V2 %in% pam50mouse$unique_mouse),c(index,15)]
41 rownames(data.pam50) <- data.pam50$V2 #NEED this, was erroring without

19  row.names(data) <- data$Gene_ID
20  ensgtoid <- read.delim('ENSMUSG_ID2Name.txt',sep='\t',header=F,stringsAsFactors = F)
21  row.names(ensgtoid) <- ensgtoid$V1
22 +data.id <- merge(as.data.frame(data), as.data.frame(ensgtoid), by='row.names')
23 +data.id <- data.id[,-1]
24 +#quantile normalized
25 +library(edgeR)
26 +data_mat <- data.matrix(data.id[,index])
27 +normfactors <- calcNormFactors(data_mat,method="upperquartile")
28 +data_mat <- sweep(data_mat,2,normfactors,'/')

29
30 #log2 transform
31 +data.id[,index] <- data_mat
32 data_norm <- data.id
33 data_norm[,index] <- log2(data_norm[,index]+1)
34

35 #genelist
36 pam50 <- read.csv("/Users/ss4cf/Desktop/Lab/Data/TCGA/PAM50/pam50.csv",header=TRUE,stringsAsFactors = F)
37 row.names(pam50) <- pam50$x

38 +
39 +#got a list of human to mouse orthologs from ensbml biomart:
40 orthologs <- read.delim("mart_export.txt",sep='\t',stringsAsFactors = F)
41 pam50mouse <- (orthologs[which(orthologs$Gene.name %in% pam50$x),c(3,5)])
42 unique_human <- unique(pam50mouse$Gene.name)
43 unique_mouse <- unique(pam50mouse$Mouse.gene.name)
44 unique_mouse <- unique_mouse[-c(23,25,26,39)]
45 pam50mouse <- data.frame(cbind(unique_human,unique_mouse))
46 row.names(pam50mouse) <- pam50mouse$unique_mouse
47 +#lots of finagling
48

49 +#NOW, get PAM50 expression in ALL samples and correct with this median
50 data.pam50 <- data_norm[which(data_norm$V2 %in% pam50mouse$unique_mouse),c(index,15)]
51 rownames(data.pam50) <- data.pam50$V2 #NEED this, was erroring without
```

To view changes on your computer: git diff <source_branch> <target_branch>

Collaborative repositories/code

You can add to this today!!

The screenshot shows a GitHub repository page for 'shamdata / janeslab_shared'. The repository is described as 'Repository for janeslab members to share code'. It has 2 commits, 1 branch, and 0 releases. The master branch is selected. There are two files listed: 'rsem2csv_kp1_impos.R' and 'shamdata_janeslab_shared.pdf'. A button to 'Add a README' is visible at the bottom.

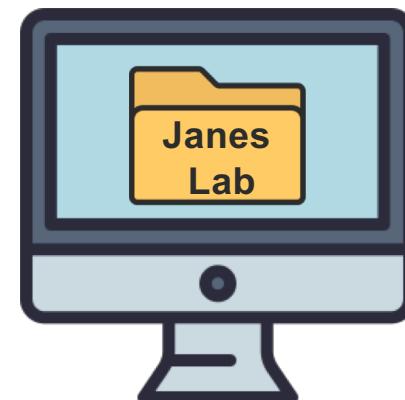
1

Clone repo `git clone https://github.com/shamdata/janeslab_shared`
(optional) Branch code `git checkout -b YOU`

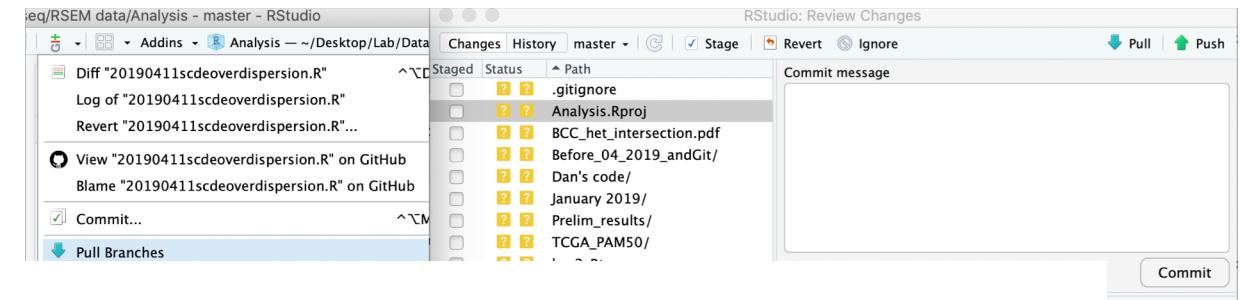
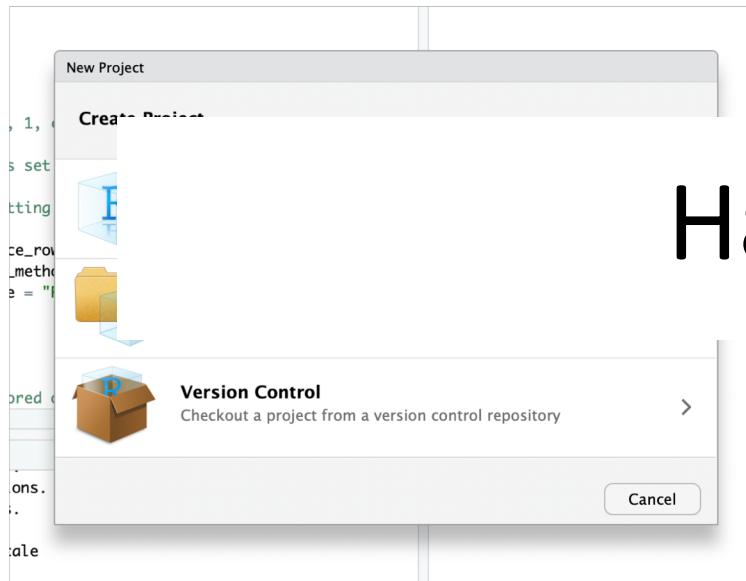
2 You edit or add code in *your folder*

Add all changes
Commit snapshot
Repeat

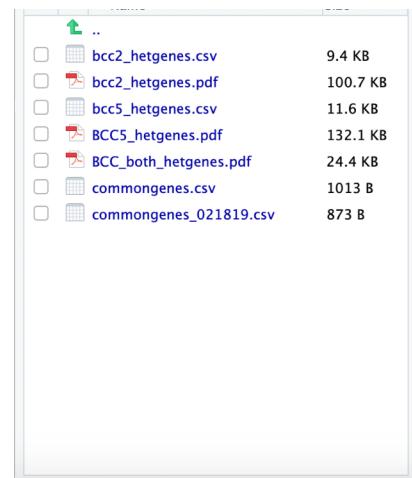
Push changes to server
(if branched)
`git push origin master`
`git push origin YOU`



GIT and RStudio:



Happy Git-ing!



```
7 EnableCodeIndexing: Yes
8 UseSpacesForTab: Yes
9 NumSpacesForTab: 2
10 Encoding: UTF-8
11
12 RnwWeave: Sweave
13 LaTeX: pdfLaTeX
```