R: A Hitchhikers Guide to Reproducible Research

- Everything in it's right place

Brendan Palmer,

Clinical Research Facility - Cork &

School of Public Health

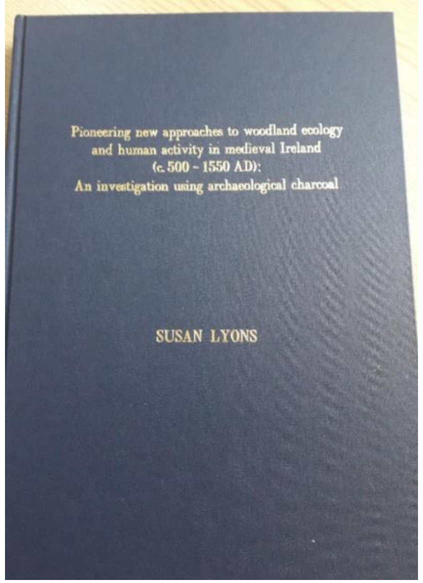
@B_A_Palmer

CRF-C
HRB Clinical Research Facility Cork

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh
School of Public Health

How is research presented?

Theses


Pioneering new approaches to woodland ecology
and human activity in medieval Ireland
(c. 500 - 1550 AD):
An investigation using archaeological charcoal
SUSAN LYONS

Papers

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology

CrossMark

Network Analysis of the Chronic Hepatitis C Virome Defines Hypervariable Region 1 Evolutionary Phenotypes in the Context of Humoral Immune Responses

Brendan A. Palmer,[a] Daniel Schmidt-Martin,[a] Zoya Dimitrova,[b] Pavel Skums,[b] Orla Crosbie,[c] Elizabeth Kenny-Walsh,[c] Liam J. Fanning[a]

Molecular Virology Diagnostic & Research Laboratory, Department of Medicine, University College Cork, Cork, Ireland[a]; Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia, USA[b]; Department of Hepatology, Cork University Hospital, Cork, Ireland[c]

ABSTRACT
Hypervariable region 1 (HVR1) of hepatitis C virus (HCV) comprises the first 27 N-terminal amino acid residues of E2. It is classically seen as the most heterogeneous region of the HCV genome. In this study, we assessed HVR1 evolution by using ultradeep pyrosequencing for a cohort of treatment-naive, chronically infected patients over a short, 16-week period. Organization of the sequence set into connected components that represented single nucleotide substitution events revealed a network dominated by highly connected, centrally positioned master sequences. HVR1 phenotypes were observed to be under strong purifying (stationary) and strong positive (antigenic drift) selection pressures, which were coincident with advancing patient age and cirrhosis of the liver. It followed that stationary viromes were dominated by a single HVR1 variant surrounded by minor variants comprised from conservative single amino acid substitution events. We present evidence to suggest that neutralization antibody efficacy was diminished for stationary-virome HVR1 variants. Our results identify the HVR1 network structure during chronic infection as the preferential dominance of a single variant within a narrow sequence space.

IMPORTANCE
HCV infection is often asymptomatic, and chronic infection is generally well established in advance of initial diagnosis and subsequent treatment. HVR1 can undergo rapid sequence evolution during acute infection, and the variant pool is typically seen to diverge away from ancestral sequences as infection progresses from the acute to the chronic phase. In this report, we describe HVR1 viromes in chronically infected patients that are defined by a dominant epitope located centrally within a narrow variant pool. Our findings suggest that weakened humoral immune activity, as a consequence of persistent chronic infection, allows for the acquisition and maintenance of host-specific adaptive mutations at HVR1 that reflect virus fitness.

Hepatitis C virus (HCV) infection is a global health issue and is recognized as a major etiological agent of liver-related diseases (1). It has been estimated that the current prevalence of HCV represents approximately 2% of the global adult (15 years of age and older) population (2). Following transmission, HCV infection may remain asymptomatic for decades, resulting in the majority of infections initially passing undetected (3). It is estimated that up to 4 million Americans are living with the virus, the majority of whom became infected prior to the isolation and identification of the virus (4, 5). Consequently, the U.S. Centers for Disease Control and Prevention now recommend that Americans born from 1945 to 1965 be screened for the presence of the virus notwithstanding the presence of clinical symptoms (3, 5).

HCV is a single-stranded positive-sense RNA virus of considerable genomic heterogeneity. A recent reclassification defined the HCV global distribution into 7 genotypes and 67 subtypes, with genotypes 1 and 3 accounting for the majority of infections worldwide (6, 7). An error-prone RNA-dependent RNA polymerase, together with an inherent tolerance of defined hypervariable regions (HVR), accounts for much of this variability. Three HVRs are located within the envelope glycoprotein E2. The greatest heterogeneity has been identified at the 27-amino-acid HVR1 (residues 384 to 410 of the H77 reference strain), located at the amino-terminal end of the E2 glycoprotein (8). Recent studies indicated that the central region of E2 (residues 456 to 656) is globular and surprisingly compact, whereas the first 80 amino acids (including HVR1) lack this structural rigidity (9). This observation is consistent with a region that is proposed to shield conserved neutralizing epitopes and to participate in high-density lipoprotein enhancement of infection via scavenger receptor class B type I (SRBI) interactions and is itself targeted by neutralizing antibodies (nAb) (10–16).

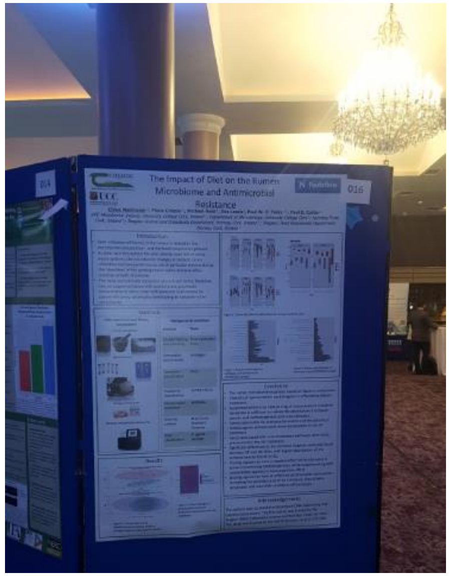
Mutational flexibility at HVR1 was characterized soon after the initial identification of HCV (8, 17). Rapid mutational change of HVR1 has been documented over weeks during the acute phase of infection, where HVR1 evolution is governed predominantly by strong selective pressures, with fixation of beneficial mutations (11, 18, 19). Reports examining samples collected over years to decades have documented the emergence of convergent HVR1

Received 25 November 2015 **Accepted** 22 December 2015
Accepted manuscript posted online 30 December 2015
Citation Palmer BA, Schmidt-Martin D, Dimitrova Z, Skums P, Crosbie O, Kenny-Walsh E, Fanning LJ. 2016. Network analysis of the chronic hepatitis C virome defines hypervariable region 1 evolutionary phenotypes in the context of humoral immune responses. J Virol 90:3318–3329. doi:10.1128/JVI.02995-15.
Editor: M. S. Diamond
Address correspondence to Liam J. Fanning, l.fanning@ucc.ie.
B.A.P. and D.S.-M. contributed equally to this article.

Books

Implementation Guide and Toolkit
for National Clinical Guidelines

Posters



Talks

But what does it really look like?

You were defending, one foot out the door

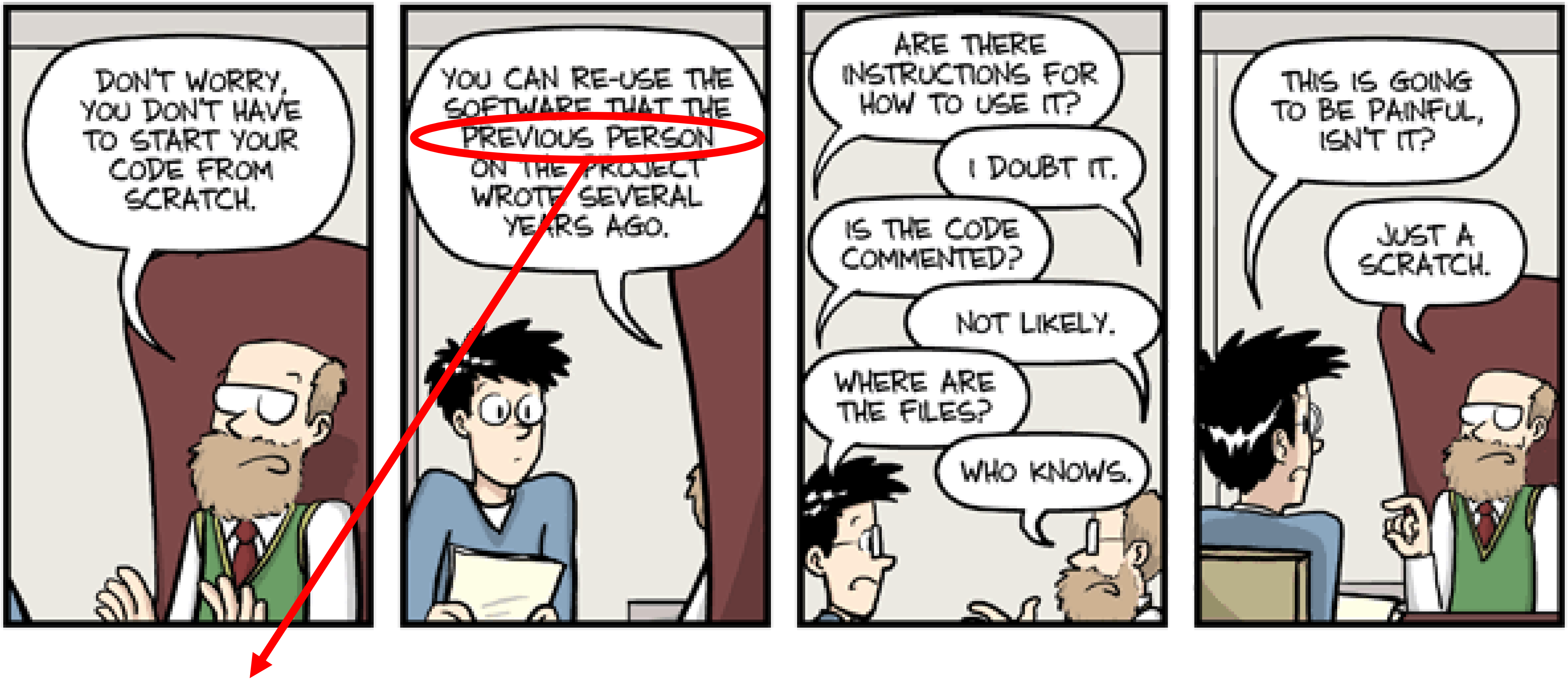
[Zheng Lab - Bad Project (Lady Gaga parody)]

I got your project and its problems galore

Zheng Lab - Bad Project (Lady Gaga parody)

I hate my life,

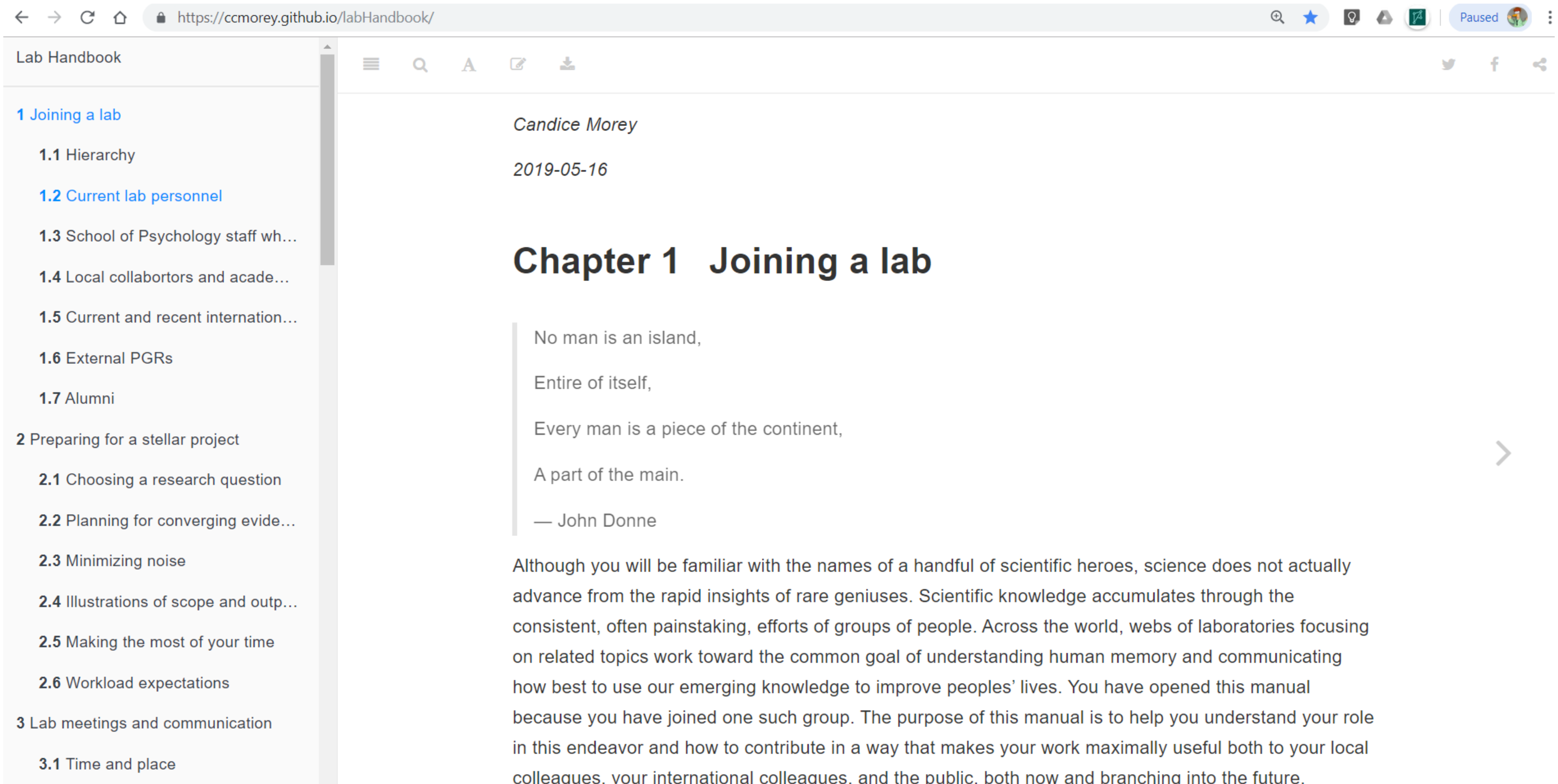
[Zheng Lab - Bad Project (Lady Gaga parody)]


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.

THIS PERSON IS likely to be YOU BTW!!

Jorge Cham | www.phdcomics.com

Reproducibility comes in many forms



Lab Handbook

Candice Morey

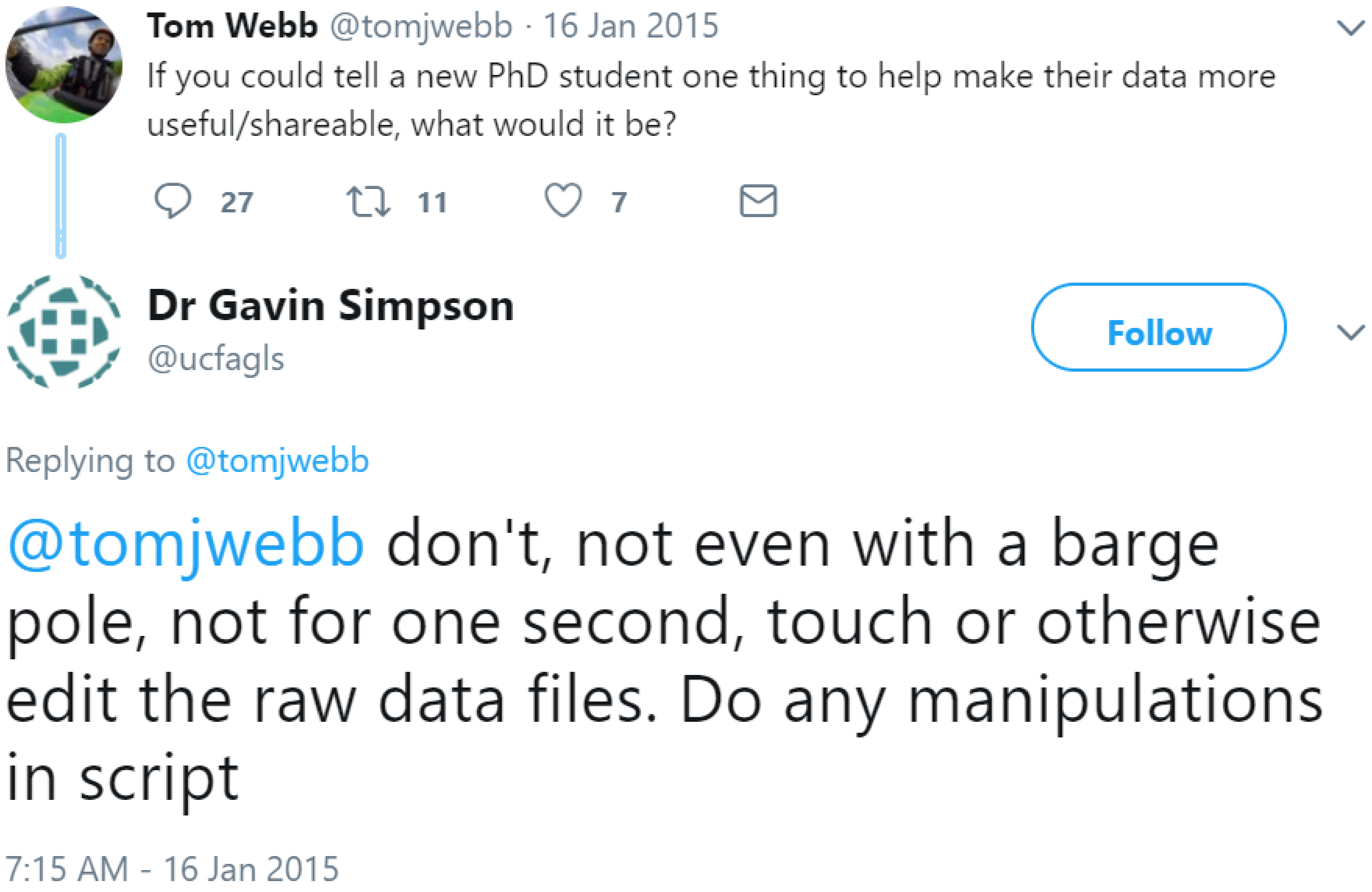
2019-05-16

Chapter 1 Joining a lab

> No man is an island,
>
> Entire of itself,
>
> Every man is a piece of the continent,
>
> A part of the main.
>
> — John Donne

Although you will be familiar with the names of a handful of scientific heroes, science does not actually advance from the rapid insights of rare geniuses. Scientific knowledge accumulates through the consistent, often painstaking, efforts of groups of people. Across the world, webs of laboratories focusing on related topics work toward the common goal of understanding human memory and communicating how best to use our emerging knowledge to improve peoples' lives. You have opened this manual because you have joined one such group. The purpose of this manual is to help you understand your role in this endeavor and how to contribute in a way that makes your work maximally useful both to your local colleagues, your international colleagues, and the public, both now and branching into the future.

Work from the raw data ALWAYS!!



Tom Webb @tomjwebb · 16 Jan 2015

If you could tell a new PhD student one thing to help make their data more useful/shareable, what would it be?

Dr Gavin Simpson
@ucfagls

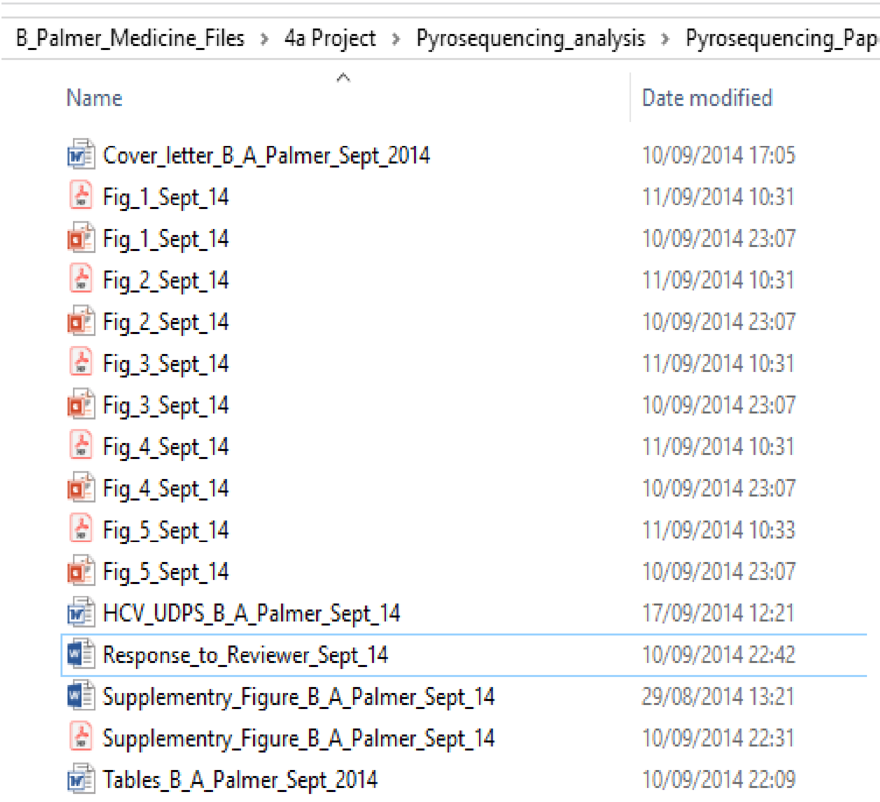
Replying to @tomjwebb

@tomjwebb don't, not even with a barge pole, not for one second, touch or otherwise edit the raw data files. Do any manipulations in script

7:15 AM - 16 Jan 2015

Still haven’t found what I’m looking for

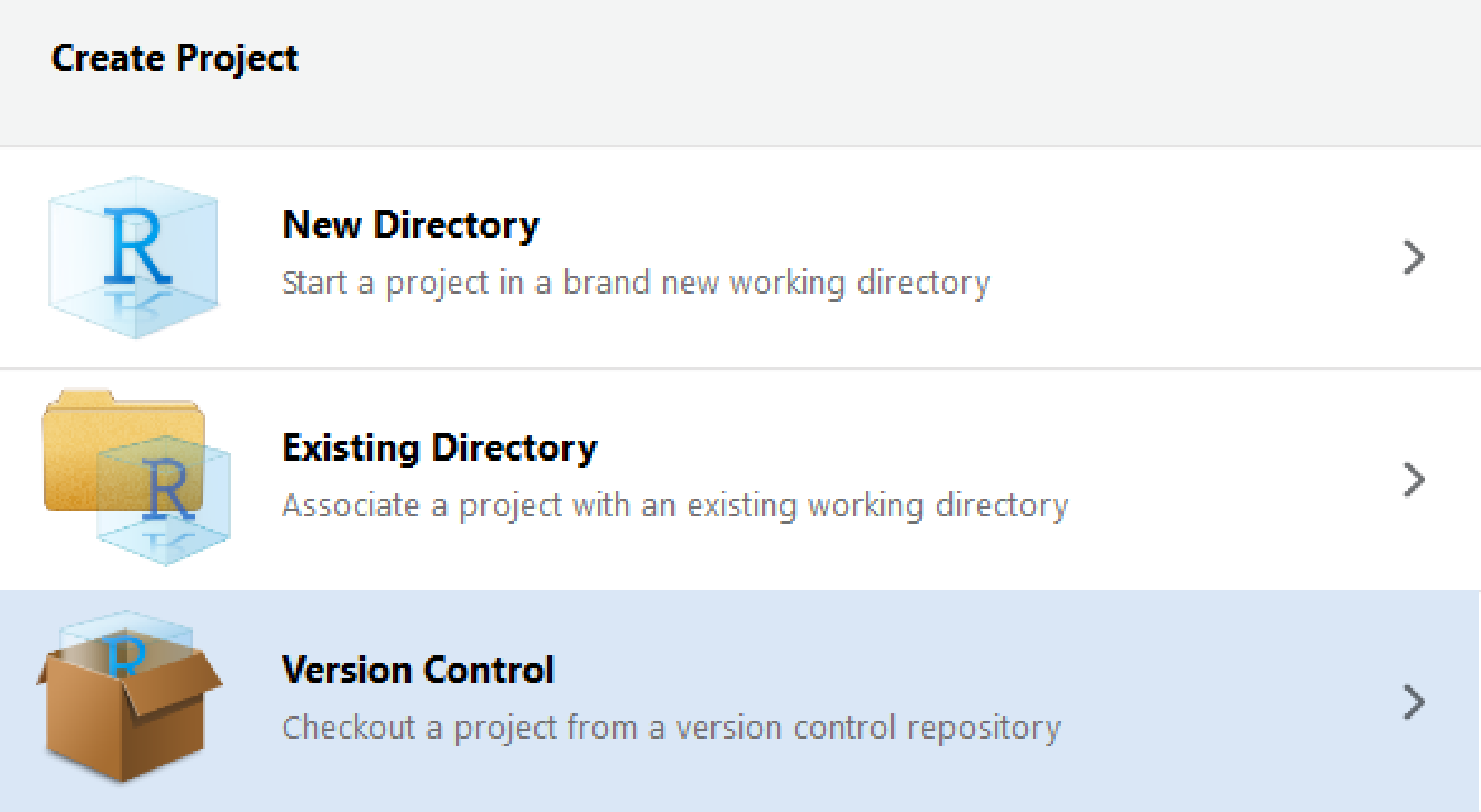
- Help your future-self



B_Palmer_Medicine_Files › 4a Project › Pyrosequencing_analysis › Pyrosequencing_Paper › Draft_Paper_incl_Figs › Submission › JVI_Resubmission › JVI_resubmission_files › Final Final version

Name	Date modified
Cover_letter_B_A_Palmer_Sept_2014	10/09/2014 17:05
Fig_1_Sept_14	11/09/2014 10:31
Fig_1_Sept_14	10/09/2014 23:07
Fig_2_Sept_14	11/09/2014 10:31
Fig_2_Sept_14	10/09/2014 23:07
Fig_3_Sept_14	11/09/2014 10:31
Fig_3_Sept_14	10/09/2014 23:07
Fig_4_Sept_14	11/09/2014 10:31
Fig_4_Sept_14	10/09/2014 23:07
Fig_5_Sept_14	11/09/2014 10:33
Fig_5_Sept_14	10/09/2014 23:07
HCV_UDPS_B_A_Palmer_Sept_14	17/09/2014 12:21
Response_to_Reviewer_Sept_14	10/09/2014 22:42
Supplementry_Figure_B_A_Palmer_Sept_14	29/08/2014 13:21
Supplementry_Figure_B_A_Palmer_Sept_14	10/09/2014 22:31
Tables_B_A_Palmer_Sept_2014	10/09/2014 22:09

R-projects



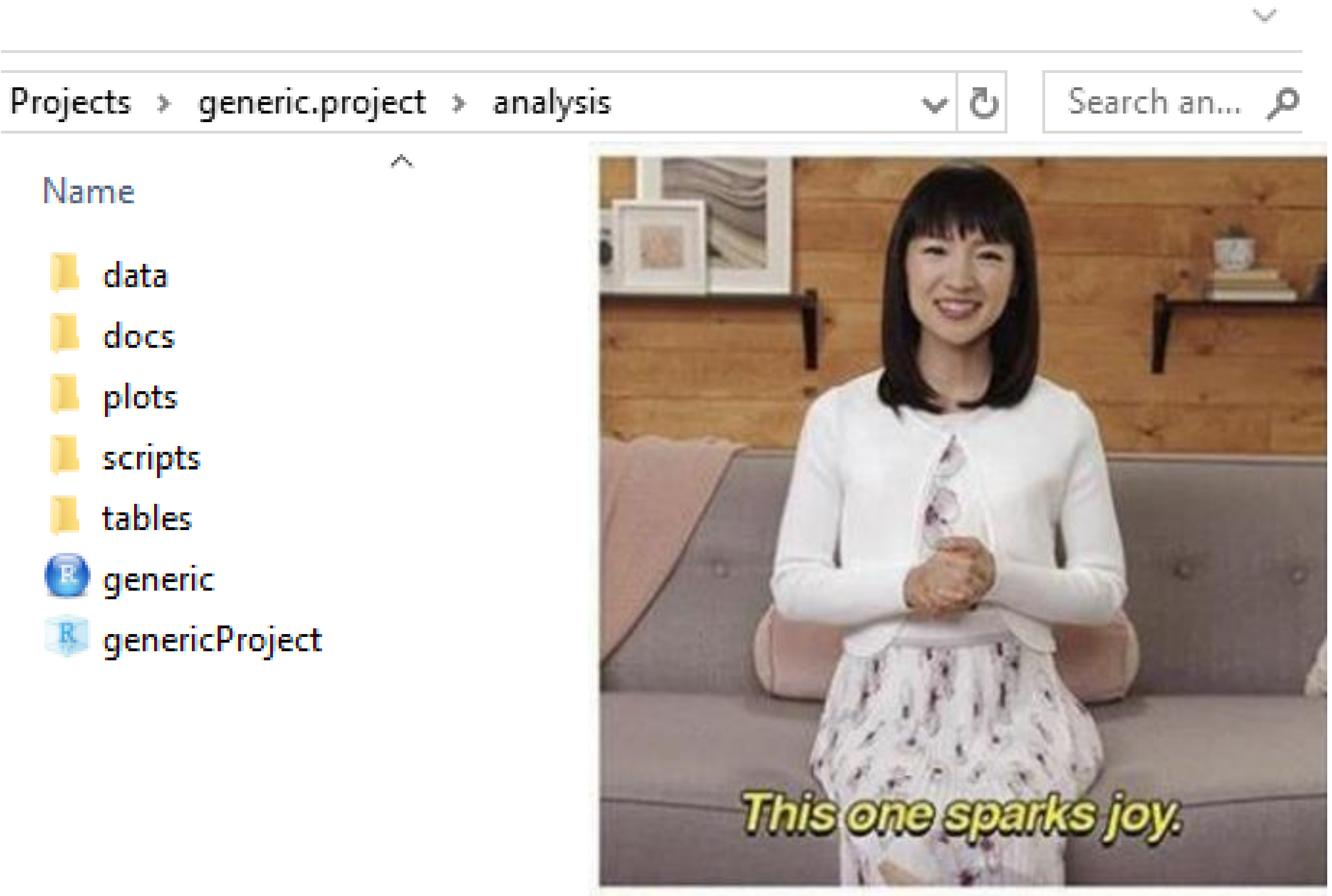
Create Project

New Directory
Start a project in a brand new working directory

Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository

Step 1: Define a generic project structure



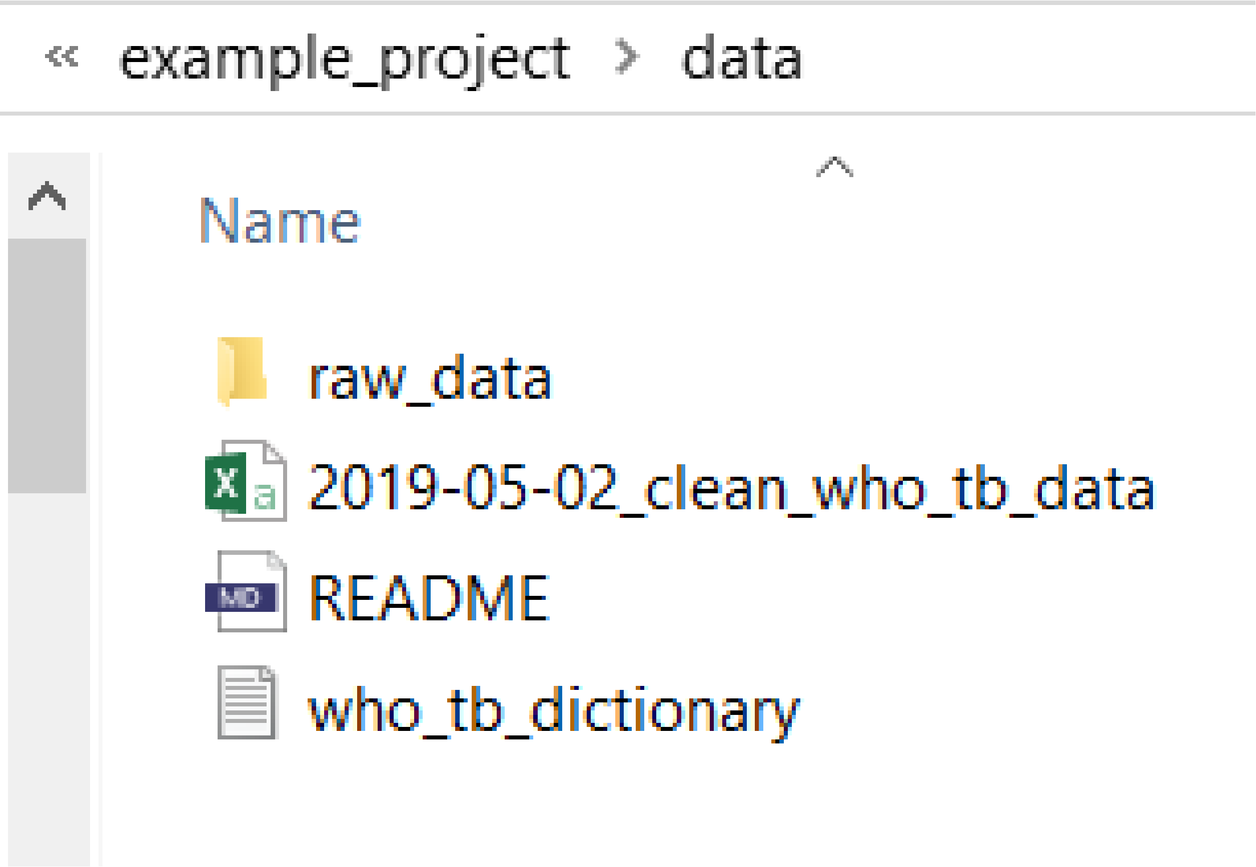
Projects › generic.project › analysis

Search an...

Name

- data
- docs
- plots
- scripts
- tables
- generic
- genericProject

Step 2: Give your files informative names



« example_project › data

Name

raw_data

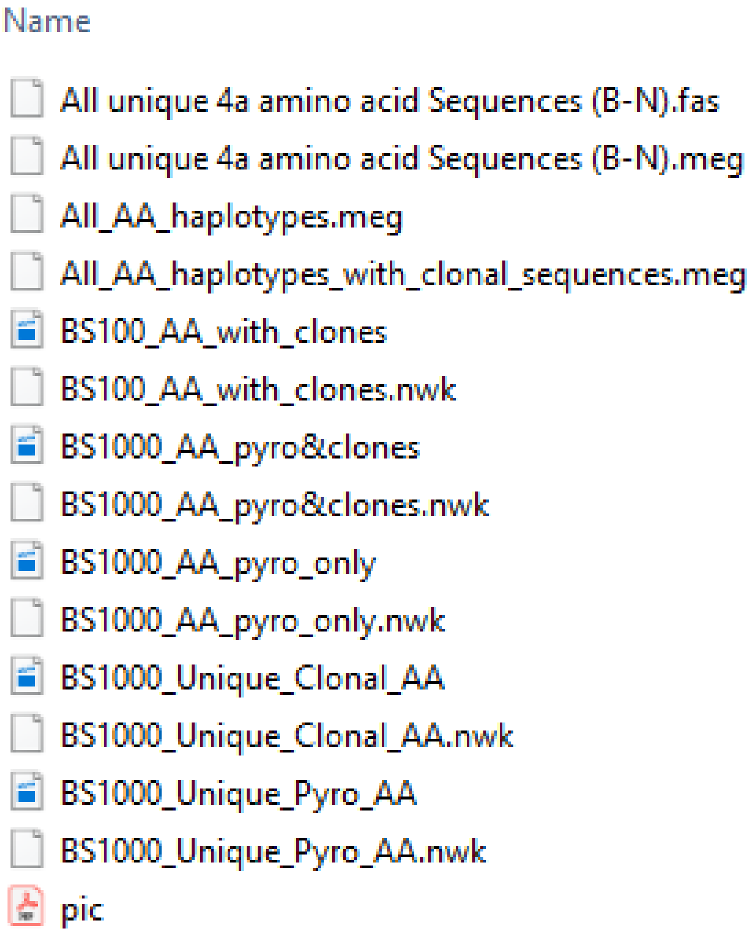
2019-05-02_clean_who_tb_data

README

who_tb_dictionary

Step 3: Make you file names machine readable, human readable and work with default ordering

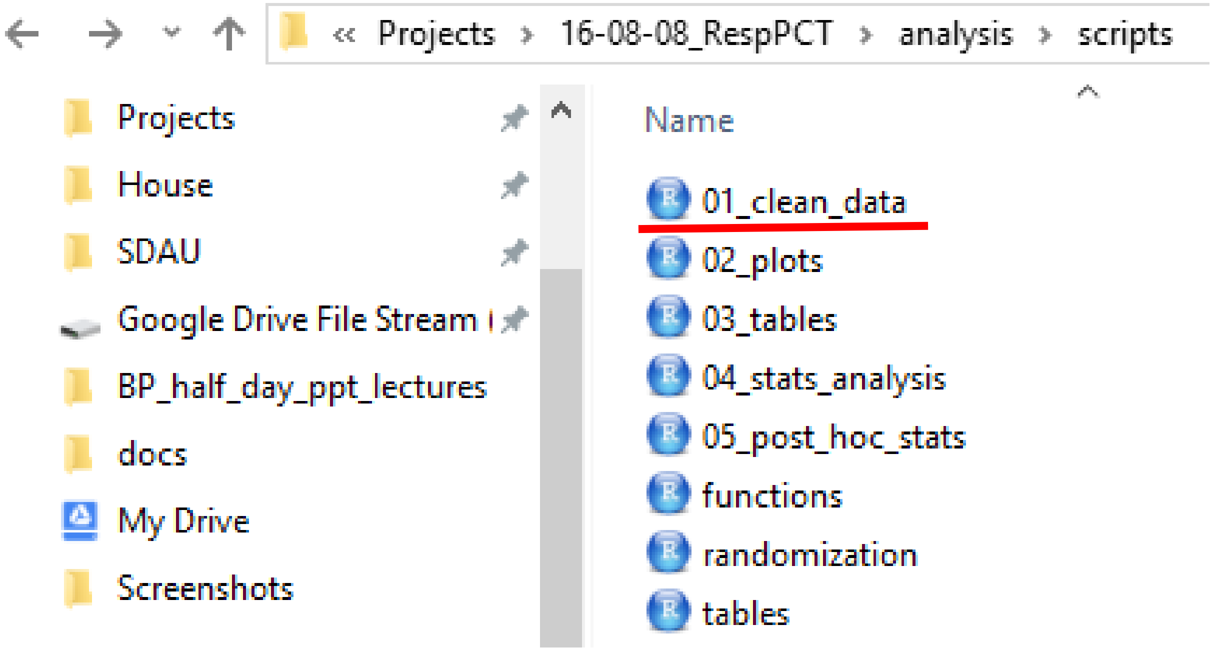
NO



Name

- All unique 4a amino acid Sequences (B-N).fas
- All unique 4a amino acid Sequences (B-N).meg
- All_AA_haplotypes.meg
- All_AA_haplotypes_with_clonal_sequences.meg
- BS100_AA_with_clones
- BS100_AA_with_clones.nwk
- BS1000_AA_pyro&clones
- BS1000_AA_pyro&clones.nwk
- BS1000_AA_pyro_only
- BS1000_AA_pyro_only.nwk
- BS1000_Unique_Clonal_AA
- BS1000_Unique_Clonal_AA.nwk
- BS1000_Unique_Pyro_AA
- BS1000_Unique_Pyro_AA.nwk
- pic

Yes



« Projects › 16-08-08_RespPCT › analysis › scripts

- Projects
- House
- SDAU
- Google Drive File Stream (
- BP_half_day_ppt_lectures
- docs
- My Drive
- Screenshots

Name

- 01_clean_data
- 02_plots
- 03_tables
- 04_stats_analysis
- 05_post_hoc_stats
- functions
- randomization
- tables

Step 4: Outline a file naming convention

Machine readable:

- Inherent order
- Avoid spaces
- Avoid punctuation
- Remove case-sensitivity

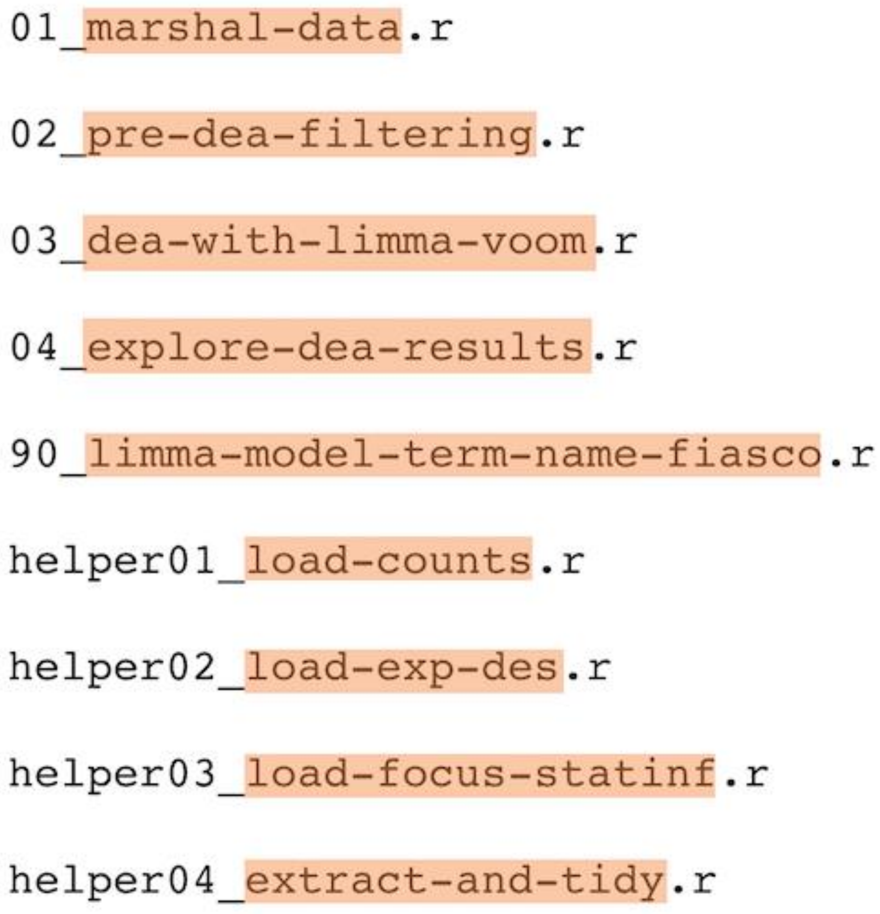
Human readable:

- Contains info on content
- Avoid spaces
- Avoid punctuation
- Remove case sensitivity

Metadata:

Separate with underscores (“_”)

- Avoid punctuation
- Remove case-sensitivity

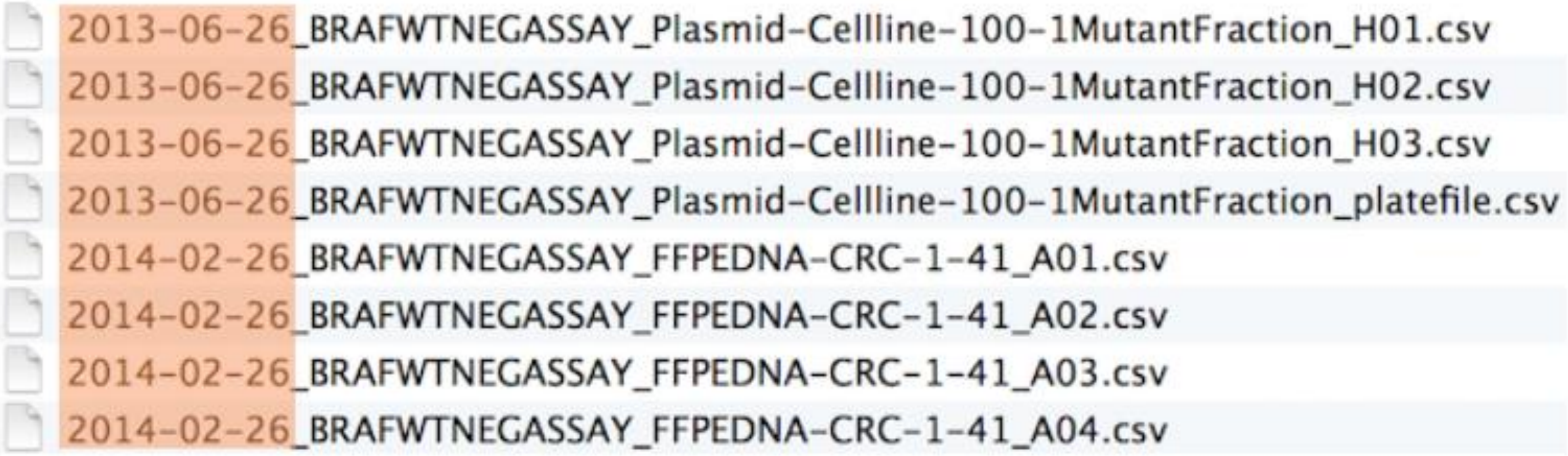


```
01_marshal-data.r

02_pre-dea-filtering.r

03_dea-with-limma-voom.r

04_explore-dea-results.r

90_limma-model-term-name-fiasco.r

helper01_load-counts.r

helper02_load-exp-des.r

helper03_load-focus-statinf.r

helper04_extract-and-tidy.r
```

Data carpentry – File naming

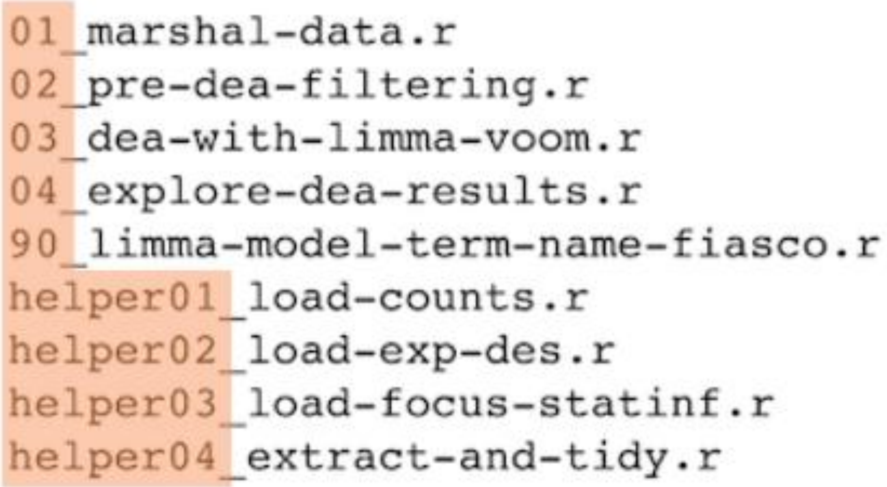
Step 4: Outline a file naming convention

Chronological order:



```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv
```

Logical order:

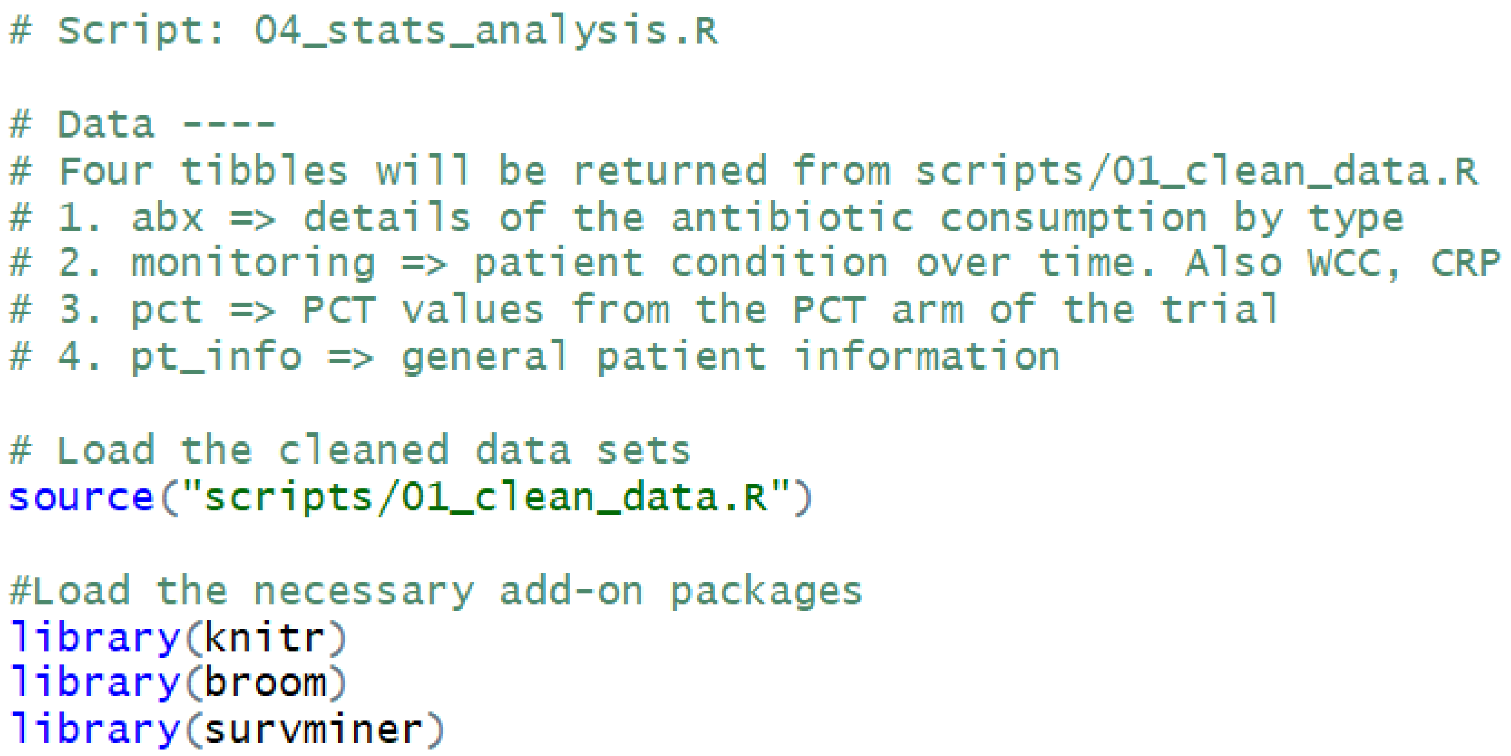


```
01_marshal-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

Data carpentry – File naming

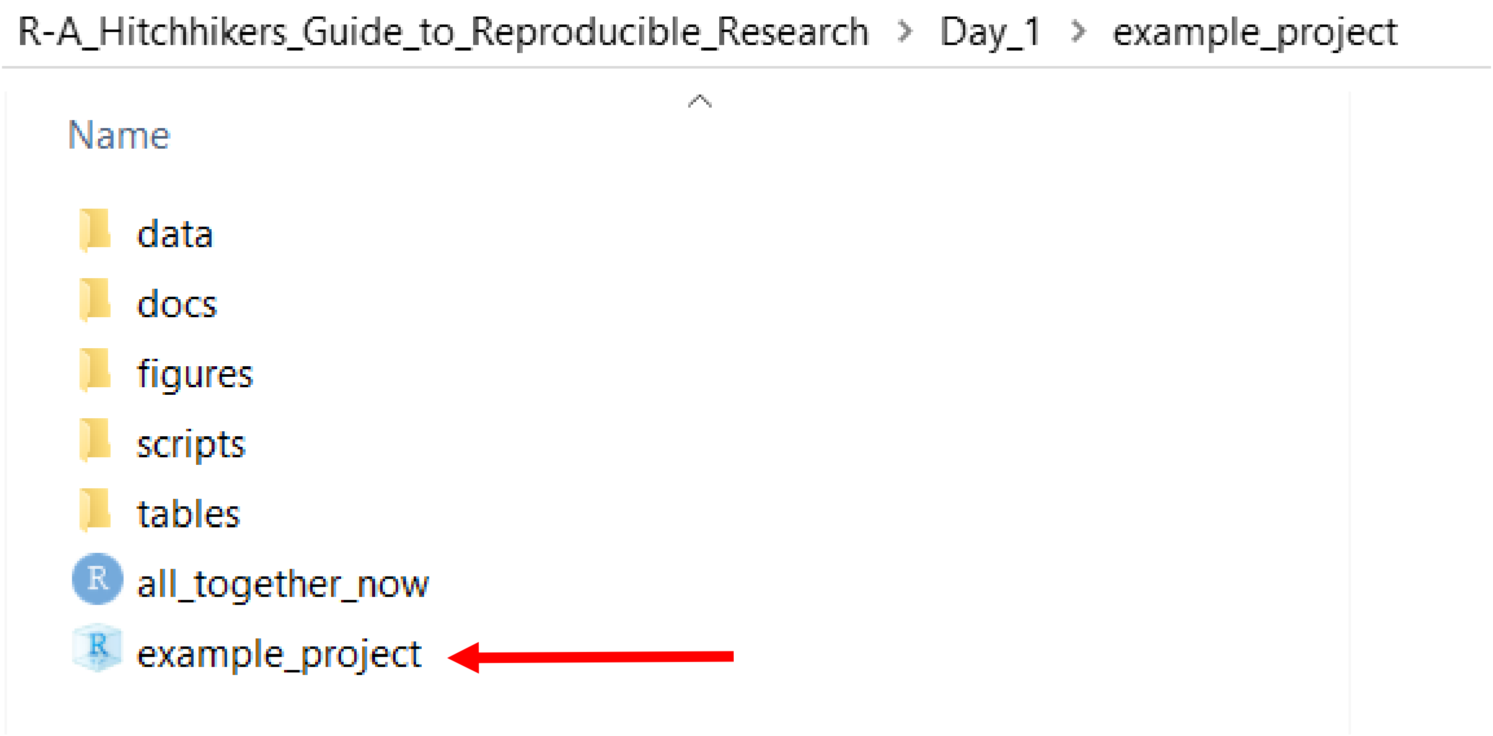
Step 5: Joined up thinking

- The R scripts you generate should be human readable
    - Annotate the code
    - Break up the scripts into dedicated tasks
    - Interlink with other within project scripts



```
# Script: 04_stats_analysis.R

# Data ----
# Four tibbles will be returned from scripts/01_clean_data.R
# 1. abx => details of the antibiotic consumption by type
# 2. monitoring => patient condition over time. Also WCC, CRP
# 3. pct => PCT values from the PCT arm of the trial
# 4. pt_info => general patient information

# Load the cleaned data sets
source("scripts/01_clean_data.R")

#Load the necessary add-on packages
library(knitr)
library(broom)
library(survminer)
```

Step 6: R-projects



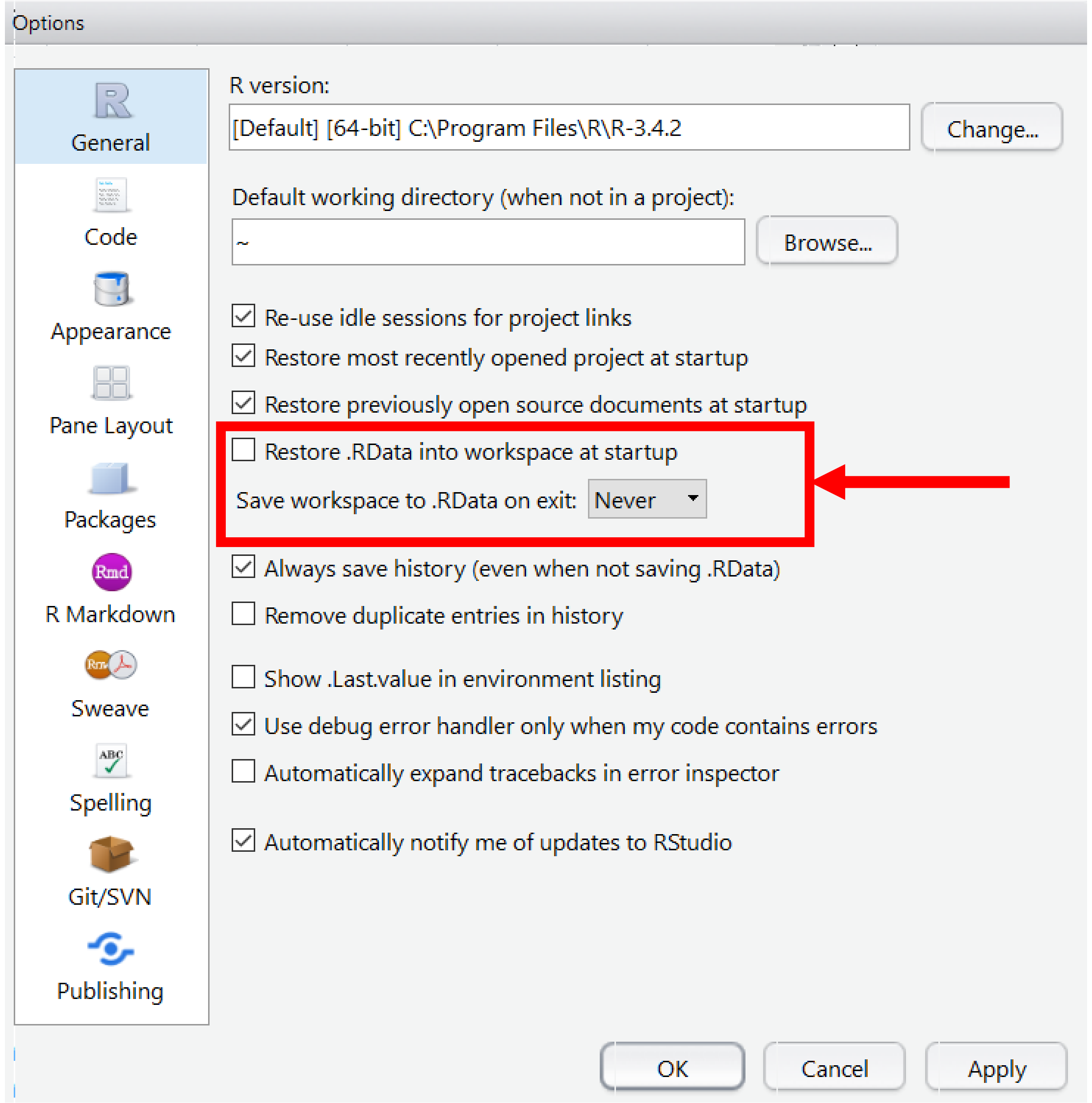
R-A_Hitchhikers_Guide_to_Reproducible_Research › Day_1 › example_project

Name

- data
- docs
- figures
- scripts
- tables
- all_together_now
- example_project

- Switch to the R-project file… Day_1/example_project/example_project.Rproj
- Open the scripts 01_eg_clean_data.R, 02_eg_figures.R and 03_eg_analysis.R

Other points to note



R version:

[Default] [64-bit] C:\Program Files\R\R-3.4.2

Default working directory (when not in a project):

~

☑ Re-use idle sessions for project links

☑ Restore most recently opened project at startup

☑ Restore previously open source documents at startup

□ Restore .RData into workspace at startup

Save workspace to .RData on exit: Never

☑ Always save history (even when not saving .RData)

□ Remove duplicate entries in history

□ Show .Last.value in environment listing

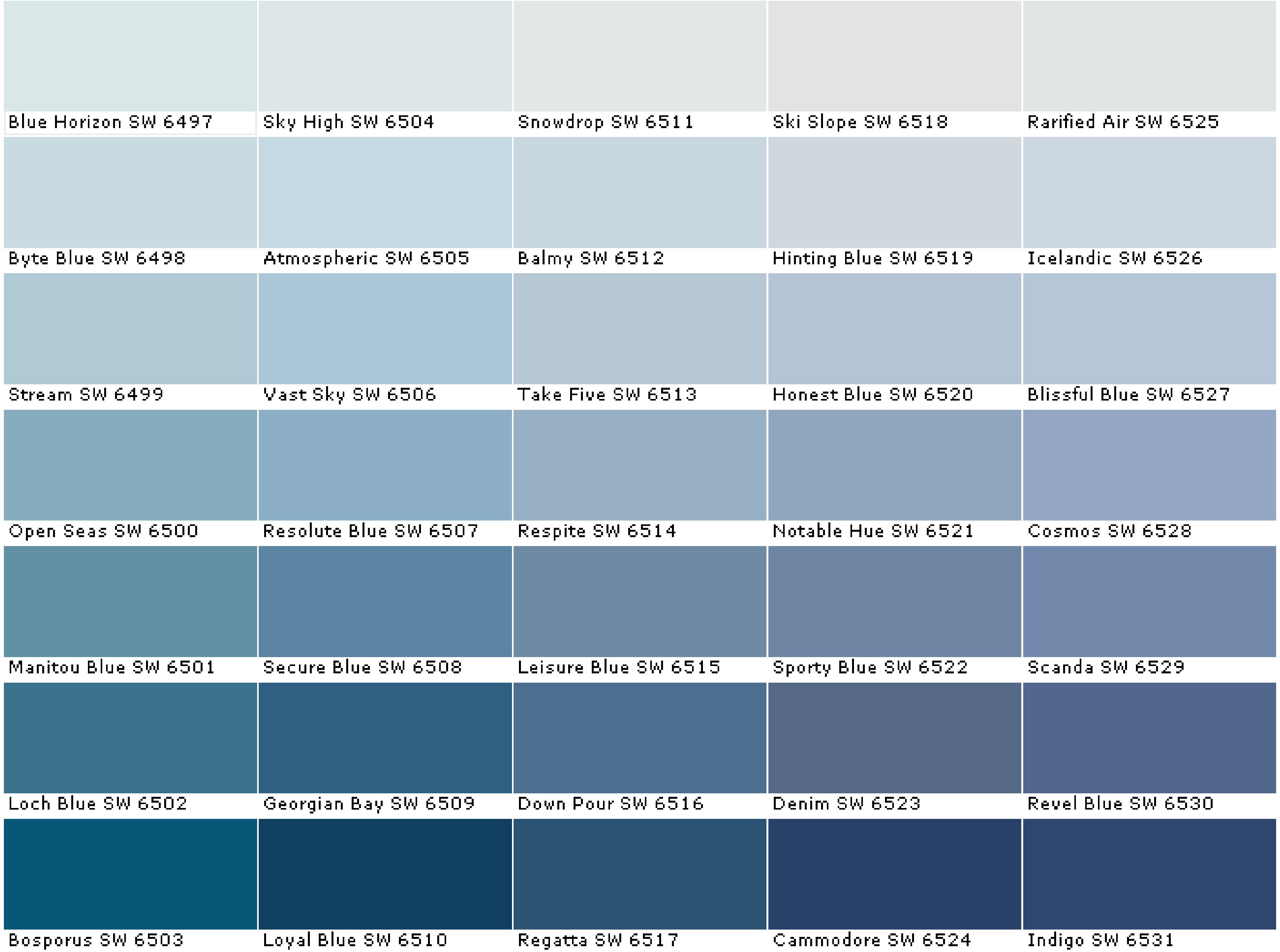
☑ Use debug error handler only when my code contains errors

□ Automatically expand tracebacks in error inspector

☑ Automatically notify me of updates to RStudio

- You might consider your environment as "real"
- If you continue to use R, it is better for you to consider your R scripts as "real", as these should recreate the environment
- You may suffer short term pain
- This will prevent long term agony

Is too much choice good or bad?



Blue Horizon SW 6497	Sky High SW 6504	Snowdrop SW 6511	Ski Slope SW 6518	Rarified Air SW 6525
Byte Blue SW 6498	Atmospheric SW 6505	Balmy SW 6512	Hinting Blue SW 6519	Icelandic SW 6526
Stream SW 6499	Vast Sky SW 6506	Take Five SW 6513	Honest Blue SW 6520	Blissful Blue SW 6527
Open Seas SW 6500	Resolute Blue SW 6507	Respite SW 6514	Notable Hue SW 6521	Cosmos SW 6528
Manitou Blue SW 6501	Secure Blue SW 6508	Leisure Blue SW 6515	Sporty Blue SW 6522	Scanda SW 6529
Loch Blue SW 6502	Georgian Bay SW 6509	Down Pour SW 6516	Denim SW 6523	Revel Blue SW 6530
Bosporus SW 6503	Loyal Blue SW 6510	Regatta SW 6517	Cammodore SW 6524	Indigo SW 6531

Inconsistent function names, inconsistent syntax

- R is a very versatile language
  - Sometimes it can be too versatile
  - Do you want to use…

    row.names or rownames
    rowSums or rowsum
    Sys.time, system.time

  - Should it be written as…

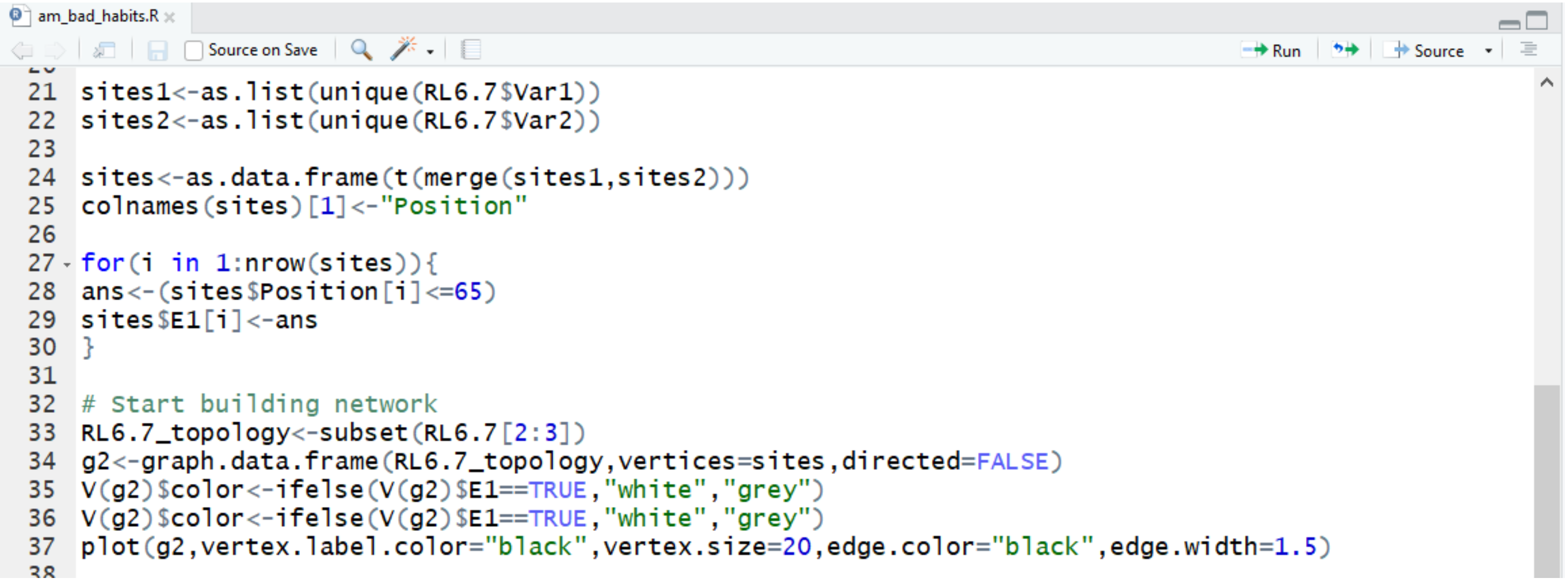
    newobject or new.Object
    x = 5 or x <- 5
    mapping=aes(x,y) or mapping = aes(x, y)

Variable selection

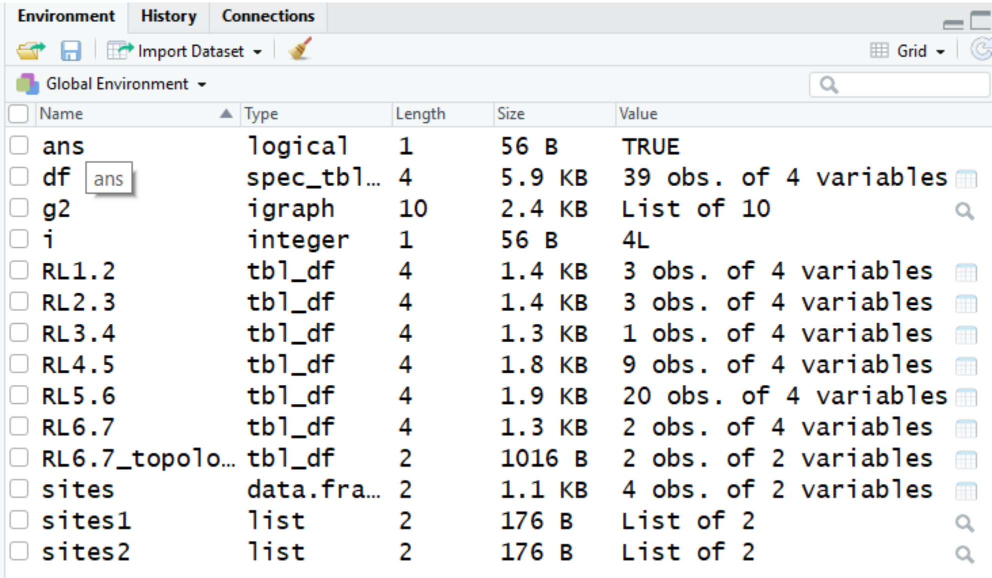
```
summary(starwars$name)

summary(starwars$"name")

summary(starwars["name"])

summary(starwars[ , "name"])

summary(starwars[1])

summary(starwars[ , 1])

summary(starwars[[1]])
```

- Open the script 04_too_much.choice.R

Motivation to move on from poorly written code



```
sites1<-as.list(unique(RL6.7$Var1))
sites2<-as.list(unique(RL6.7$Var2))

sites<-as.data.frame(t(merge(sites1,sites2)))
colnames(sites)[1]<-"Position"

for(i in 1:nrow(sites)){
ans<-(sites$Position[i]<=65)
sites$E1[i]<-ans
}

# Start building network
RL6.7_topology<-subset(RL6.7[2:3])
g2<-graph.data.frame(RL6.7_topology,vertices=sites,directed=FALSE)
V(g2)$color<-ifelse(V(g2)$E1==TRUE,"white","grey")
V(g2)$color<-ifelse(V(g2)$E1==TRUE,"white","grey")
plot(g2,vertex.label.color="black",vertex.size=20,edge.color="black",edge.width=1.5)
```



Name	Type	Length	Size	Value
ans	logical	1	56 B	TRUE
df	spec_tbl…	4	5.9 KB	39 obs. of 4 variables
g2	igraph	10	2.4 KB	List of 10
i	integer	1	56 B	4L
RL1.2	tbl_df	4	1.4 KB	3 obs. of 4 variables
RL2.3	tbl_df	4	1.4 KB	3 obs. of 4 variables
RL3.4	tbl_df	4	1.3 KB	1 obs. of 4 variables
RL4.5	tbl_df	4	1.8 KB	9 obs. of 4 variables
RL5.6	tbl_df	4	1.9 KB	20 obs. of 4 variables
RL6.7	tbl_df	4	1.3 KB	2 obs. of 4 variables
RL6.7_topolo…	tbl_df	2	1016 B	2 obs. of 2 variables
sites	data.fra…	2	1.1 KB	4 obs. of 2 variables
sites1	list	2	176 B	List of 2
sites2	list	2	176 B	List of 2

Lack of annotation
Poor naming conventions
Poor readability
Spacing absent

Cluttered environment
Intermediate objects

- Open the script 05_bad_habits.R

Writing clearer code

- Annotation

- Object names
  - should use only lowercase letters, numbers, and “_”

- Spacing
  - Put a space before and after =
  - Put a space after a ,
  - Operators should be surrounded by spaces e.g. ==, <-, +

- For a more complete list visit
  - http://style.tidyverse.org/syntax.html

- Open the script 06_good_habits.R