R-eproducible workflows

1-day workshop
Morning practical session

Brendan Palmer,
Statistics & Data Analysis Unit,
Clinical Research Facility - Cork
@B_A_Palmer

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

CRF-C
HRB Clinical Research Facility Cork

HRB
Health Research Board

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 R-project file reproducible-workflows_2019.Rproj
- Open the script am_too_much.choice.R

Motivation to move on from poorly written code

am_bad_habits.R

Source on Save | Run | Source

```r
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)
```

Environment | History | Connections

Import Dataset | Grid

Global Environment

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 am_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. ==, <-, +

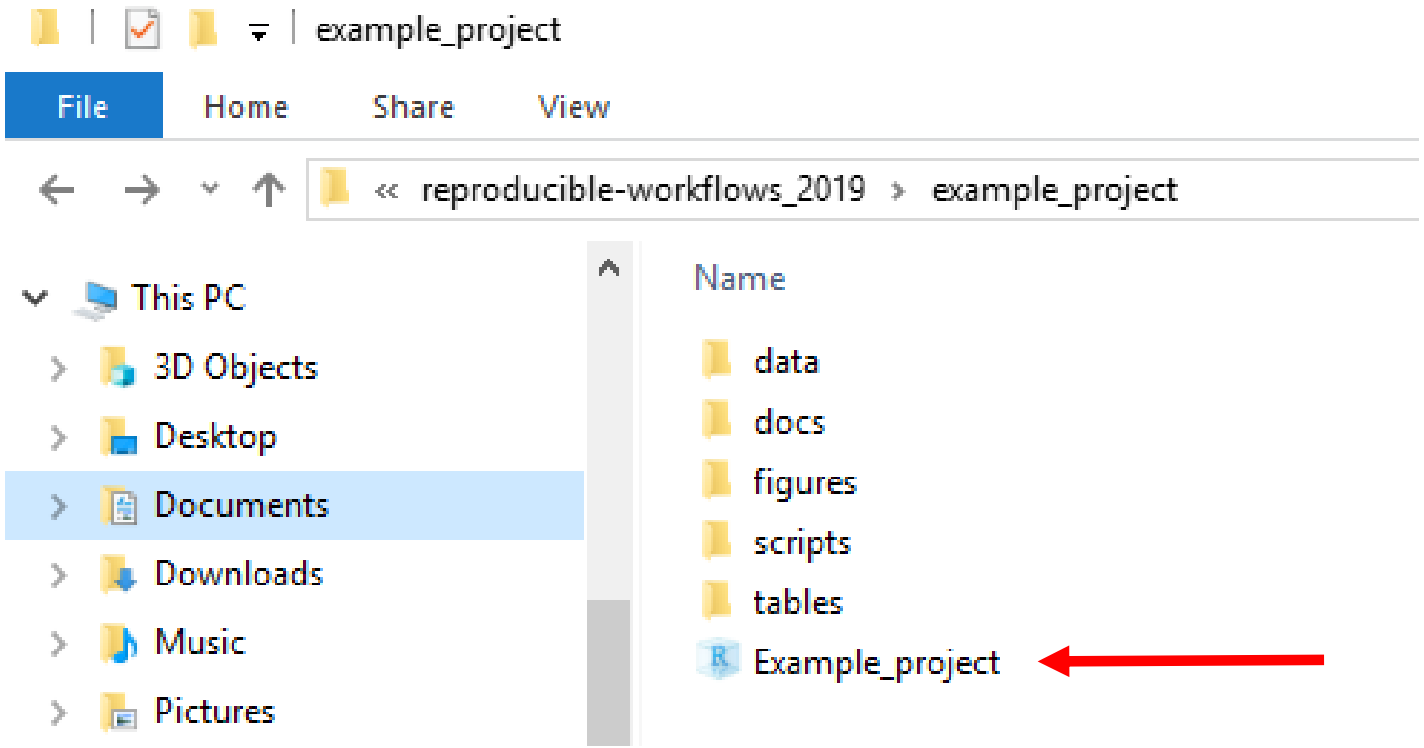
- For a more complete list visit
  - http://style.tidyverse.org/syntax.html

- Open the script am_good_habits.R

Navigating RStudio – some useful tips

- Open the script am_rstudio_ide_tricks.R

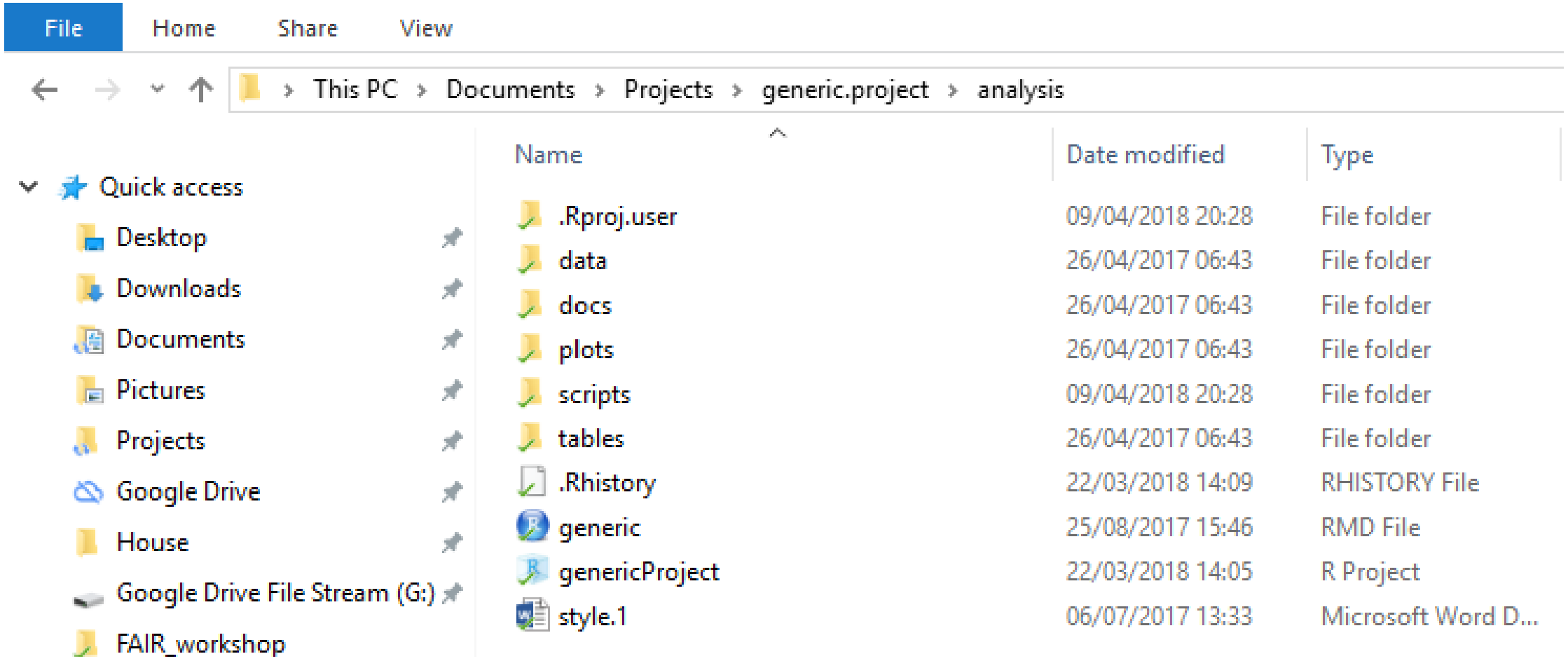
B: R-projects - everything in its right place



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

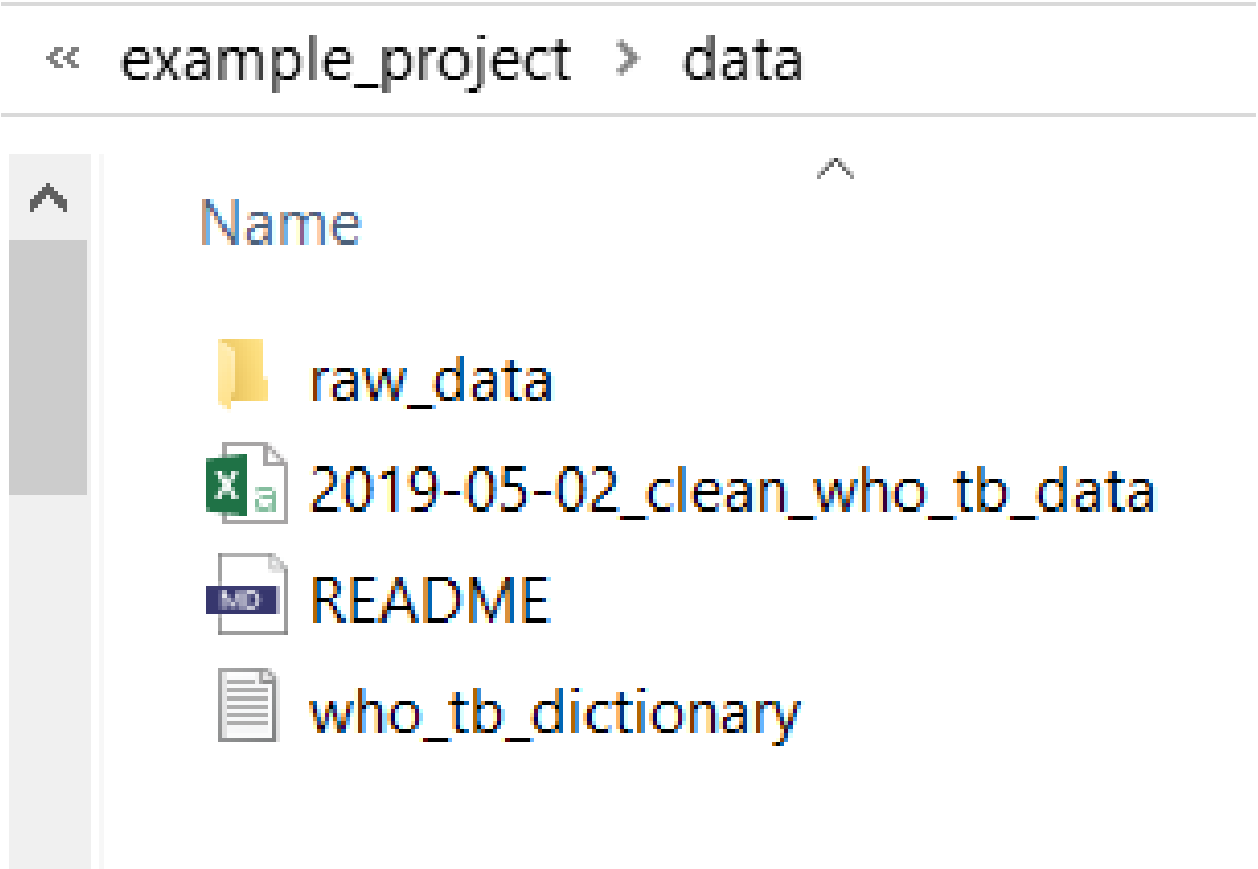
Define a generic project structure

- STEP 1: Give your research projects a shared structure



Give your files informative names

- STEP 2: Include metadata in the file names



Come back to what you know

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

NO

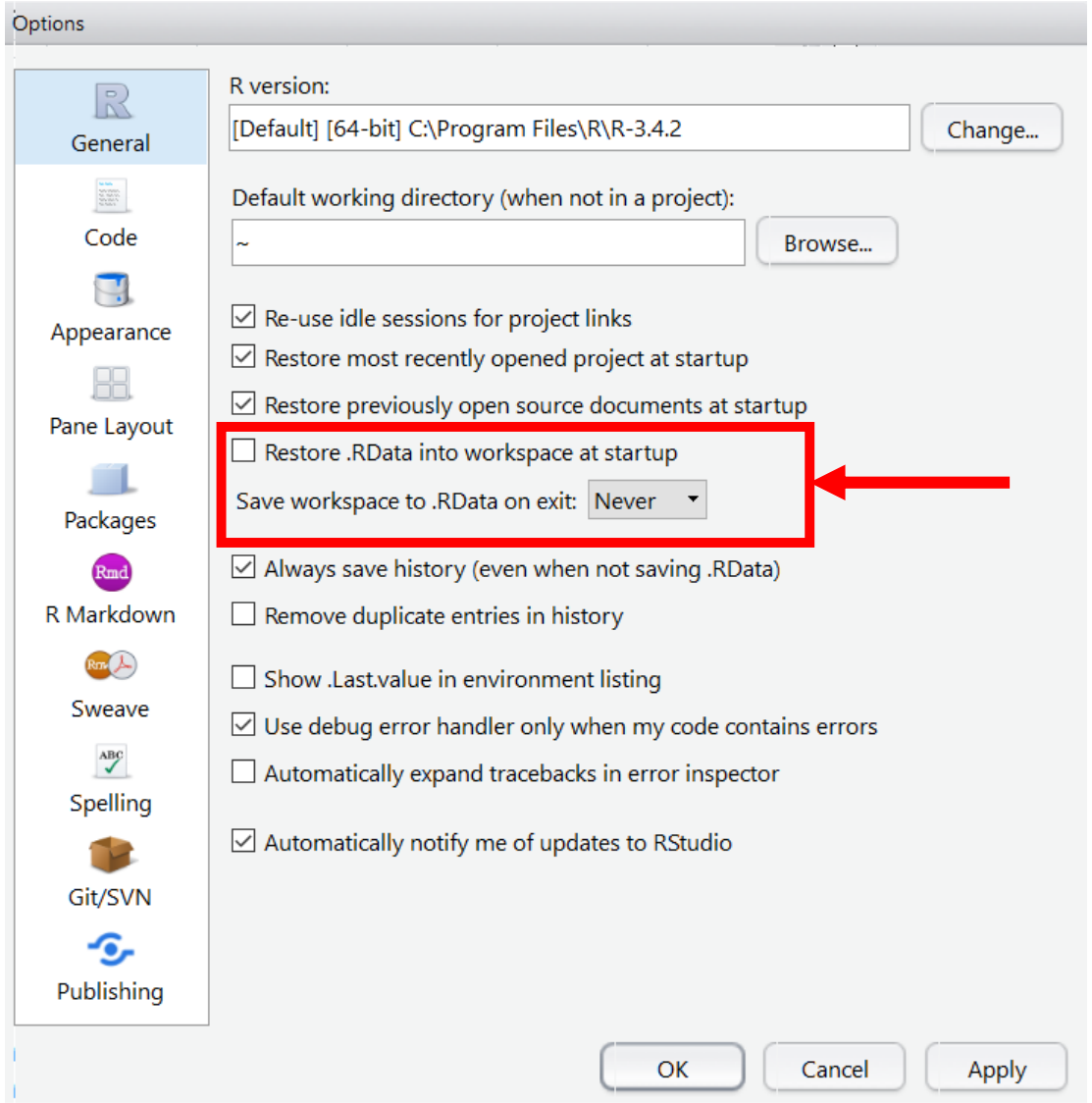
- Epistatic_change
- Epistatic_change_match_discovery
- Epistatic_change_match_discovery_fig_2_point_1
- Epistatic_change_v2
- epistatic_codon _change_tracking
- Epistatic_connection_network
- Heatmap_for_epistatic_syn
- Heatmap1_for_epi_site_co-change
- Heatmap2_for_epi_fdr_adjusted_p-value
- Heatmap2_for_epi_p-value

Yes

Documents › Projects › 18.04.27-WP3_Feeding_Trial › analysis › scripts

Name	Date modified
01_data_import_and_tidying_master_file	02/10/2018 18:51
02_data_import_and_tidying_nutritics_grouped	19/10/2018 19:47
03_figures	17/10/2018 16:40
04_tables	22/05/2018 12:26
05_study_overview	19/10/2018 23:06
functions	13/05/2018 23:13

Other points to note



- 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