R-eproducible workflows

1-day workshop
Afternoon overview

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

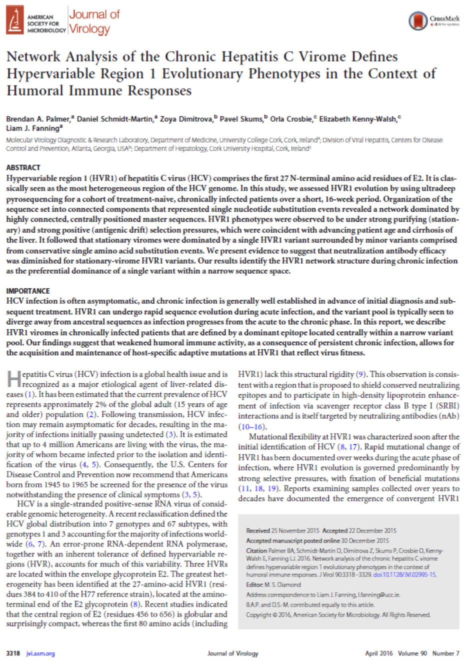
UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

CRF-C
HRB Clinical Research Facility Cork

HRB
Health
Research
Board

How is research presented?

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).

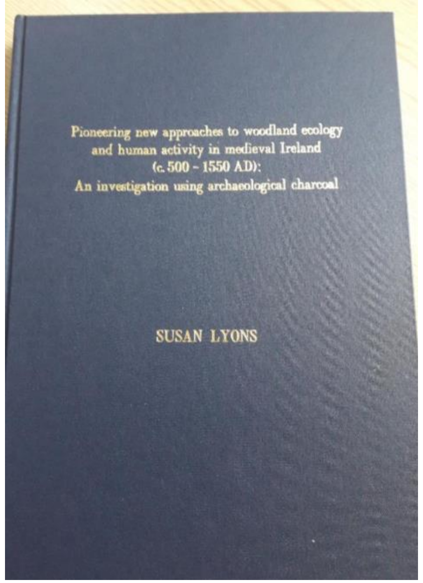
Mutational flexibility at HVR1 was characterized soon after the initial identification of HCV (8, 17). Rapid mutational change of HVR1 has been documented over weeks during the acute phase of infection, where HVR1 evolution is governed predominantly by strong selective pressures, with fixation of beneficial 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

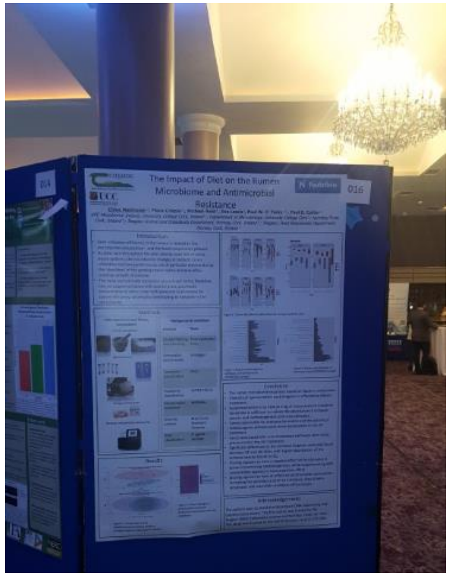
Implementation Guide and Toolkit
for National Clinical Guidelines

Theses


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

Talks

Posters



But what does it look like under the bonnet?

You were defending, one foot out the door

www.youtube.com/watch?v=Fl4L4M8m4d0

I got your project and its problems galore

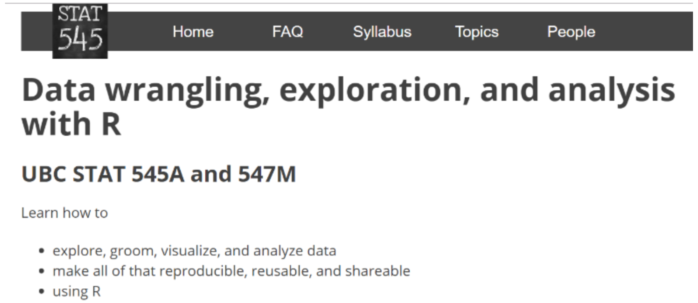
www.youtube.com/watch?v=Fl4L4M8m4d0

I hate my life,

www.youtube.com/watch?v=Fl4L4M8m4d0

Disclaimer

- Jenny Bryan is the go-to resource for most of the content in this session



Data wrangling, exploration, and analysis with R

UBC STAT 545A and 547M

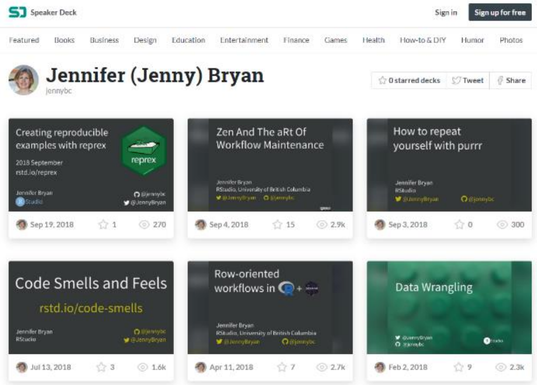
Learn how to

- explore, groom, visualize, and analyze data
- make all of that reproducible, reusable, and shareable
- using R

Jenny Bryan

@JennyBryan

Software engineer @rstudio, humane #rstats, adjunct prof @UBC where I created @STAT545, part of @ropensci



Jennifer (Jenny) Bryan

jennybc

Creating reproducible examples with reprex
2018 September
rstd.io/reprex
Jennifer Bryan
RStudio
@jennybc
@JennyBryan
Sep 19, 2018 · 1 · 270

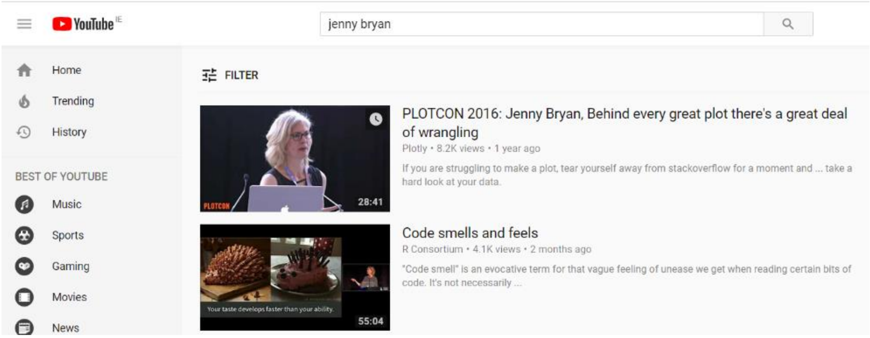
Zen And The aRt Of Workflow Maintenance
Jennifer Bryan
RStudio, University of British Columbia
@JennyBryan @jennybc
Sep 4, 2018 · 15 · 2.9k

How to repeat yourself with purrr
Jennifer Bryan
RStudio
@JennyBryan @jennybc
Sep 3, 2018 · 0 · 300

Code Smells and Feels
rstd.io/code-smells
Jennifer Bryan
RStudio
@jennybc
@JennyBryan
Jul 13, 2018 · 3 · 1.6k

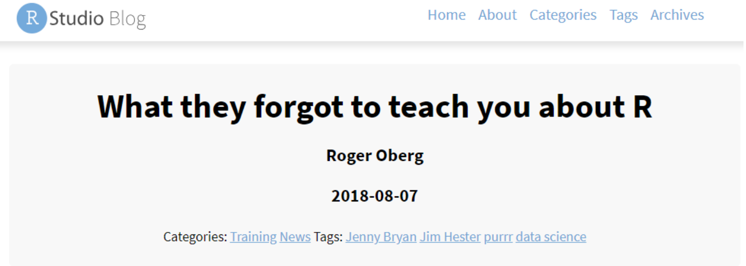
Row-oriented workflows in R +
Jennifer Bryan
RStudio, University of British Columbia
@JennyBryan @jennybc
Apr 11, 2018 · 7 · 2.7k

Data Wrangling
@JennyBryan
@jennybc
Feb 2, 2018 · 9 · 2.3k



PLOTCON 2016: Jenny Bryan, Behind every great plot there's a great deal of wrangling
Plotly • 8.2K views • 1 year ago
If you are struggling to make a plot, tear yourself away from stackoverflow for a moment and ... take a hard look at your data.

Code smells and feels
R Consortium • 4.1K views • 2 months ago
"Code smell" is an evocative term for that vague feeling of unease we get when reading certain bits of code. It's not necessarily ...



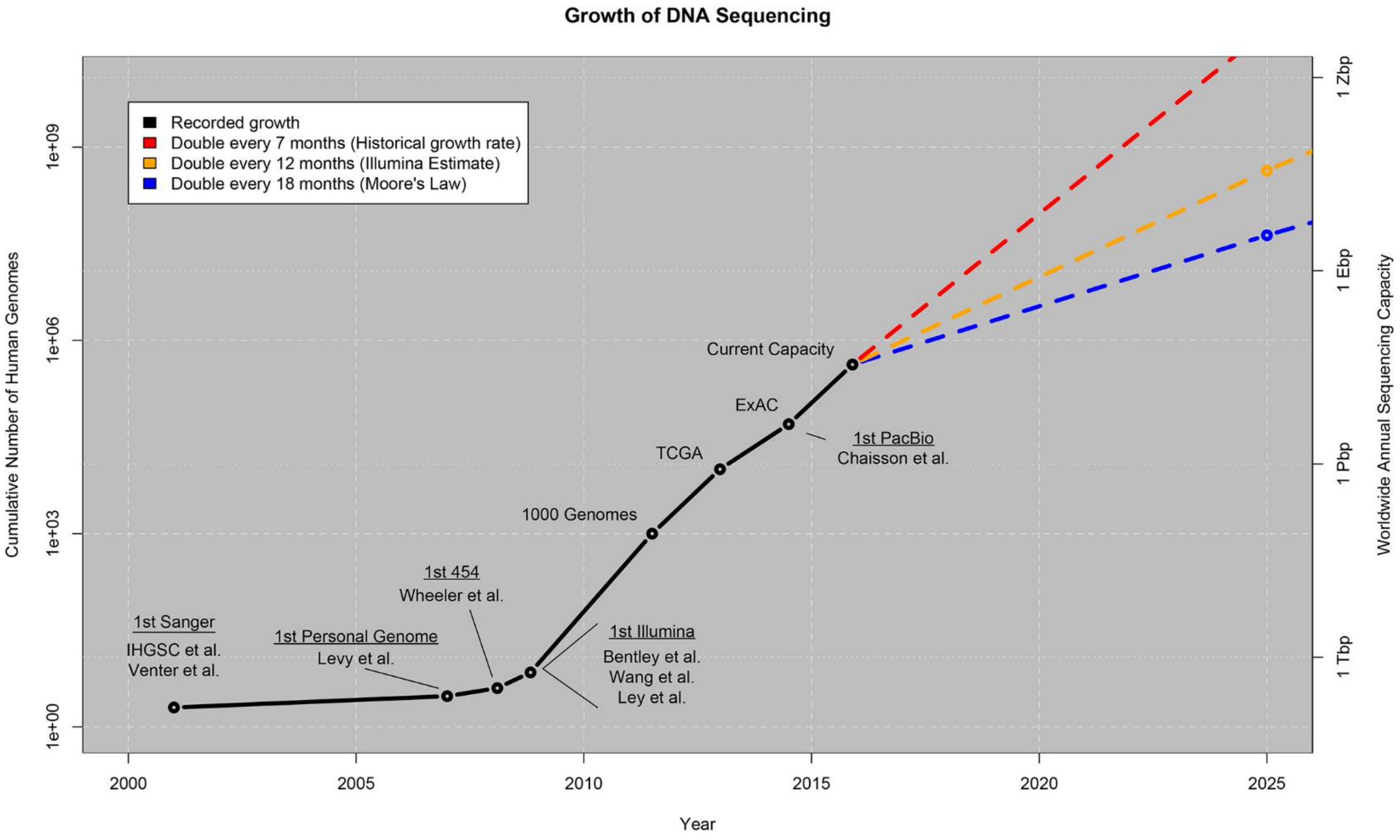
What they forgot to teach you about R

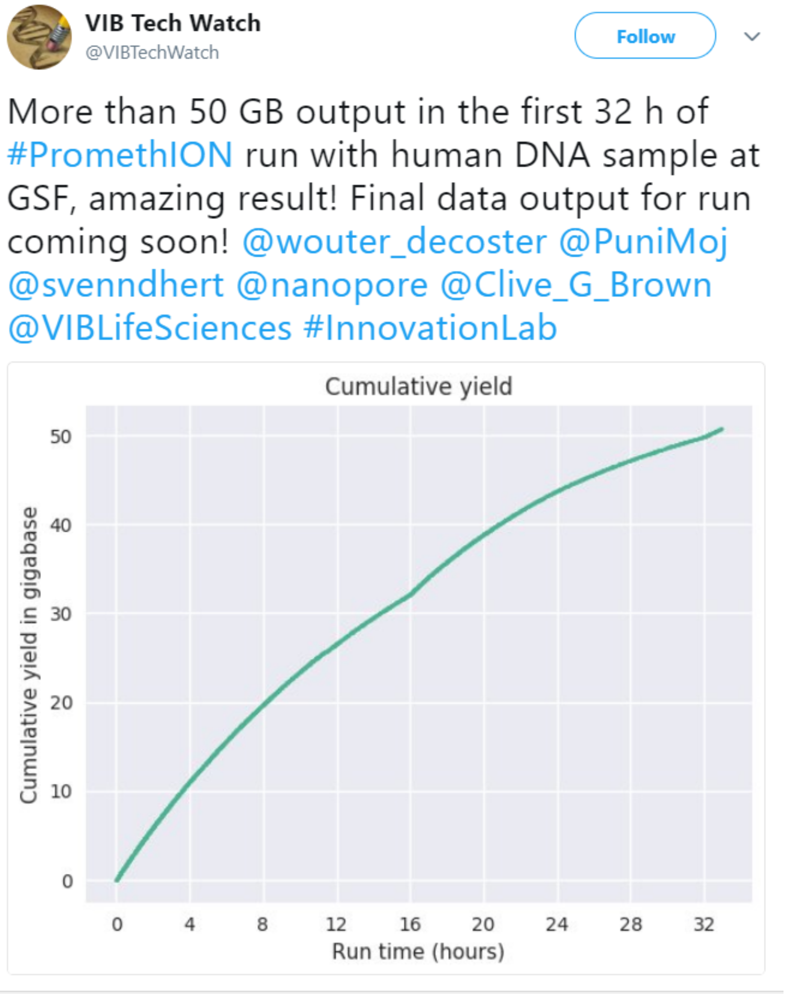
Roger Oberg

2018-08-07

Categories: Training News Tags: Jenny Bryan Jim Hester purrr data science

The explosion of data in the life sciences



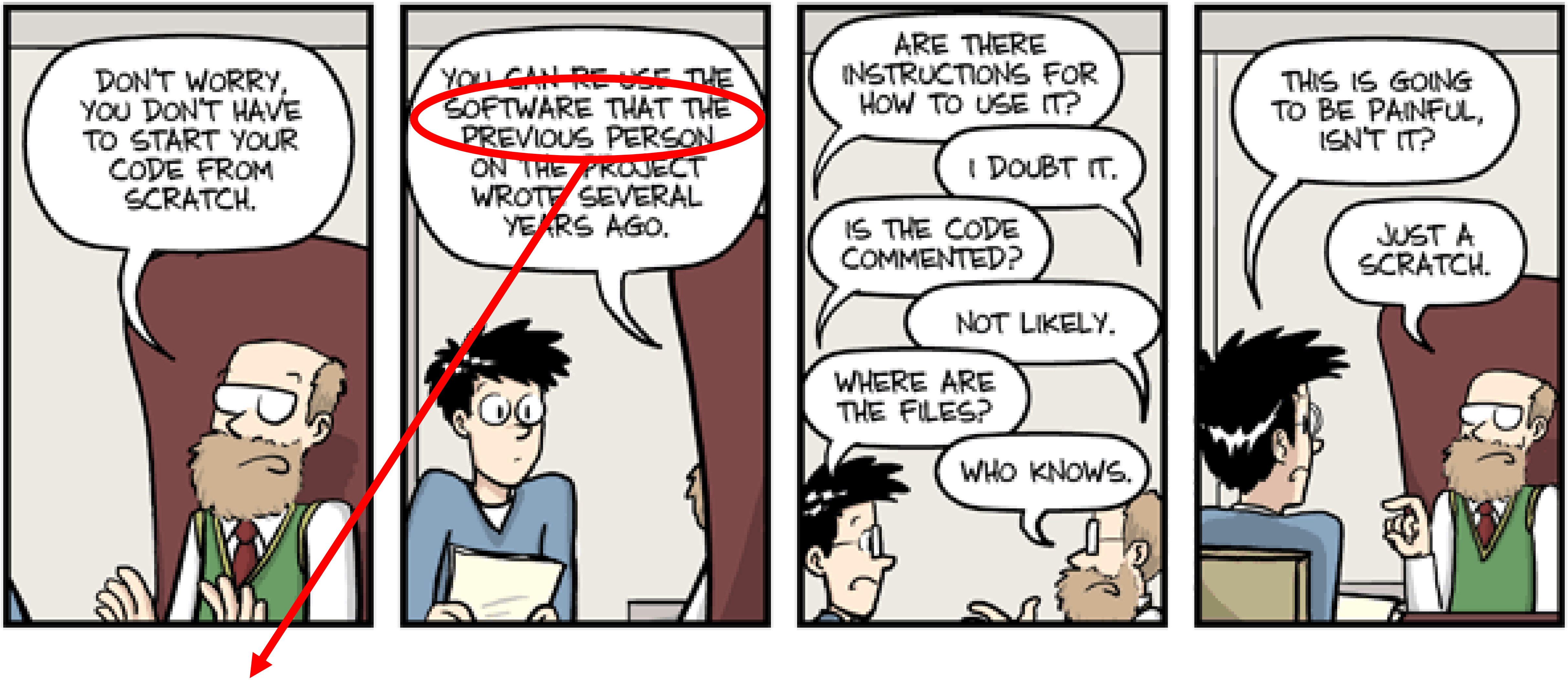


More than 50 GB output in the first 32 h of #PromethION run with human DNA sample at GSF, amazing result! Final data output for run coming soon! @wouter_decoster @PuniMoj @svenndhert @nanopore @Clive_G_Brown @VIBLifeSciences #InnovationLab

Stevens et. al., 2015, Plos Biology

This session

- Project structure
    - Naming conventions
        - Scripted workflows
            - RMarkdown reports
                - Reproducible research

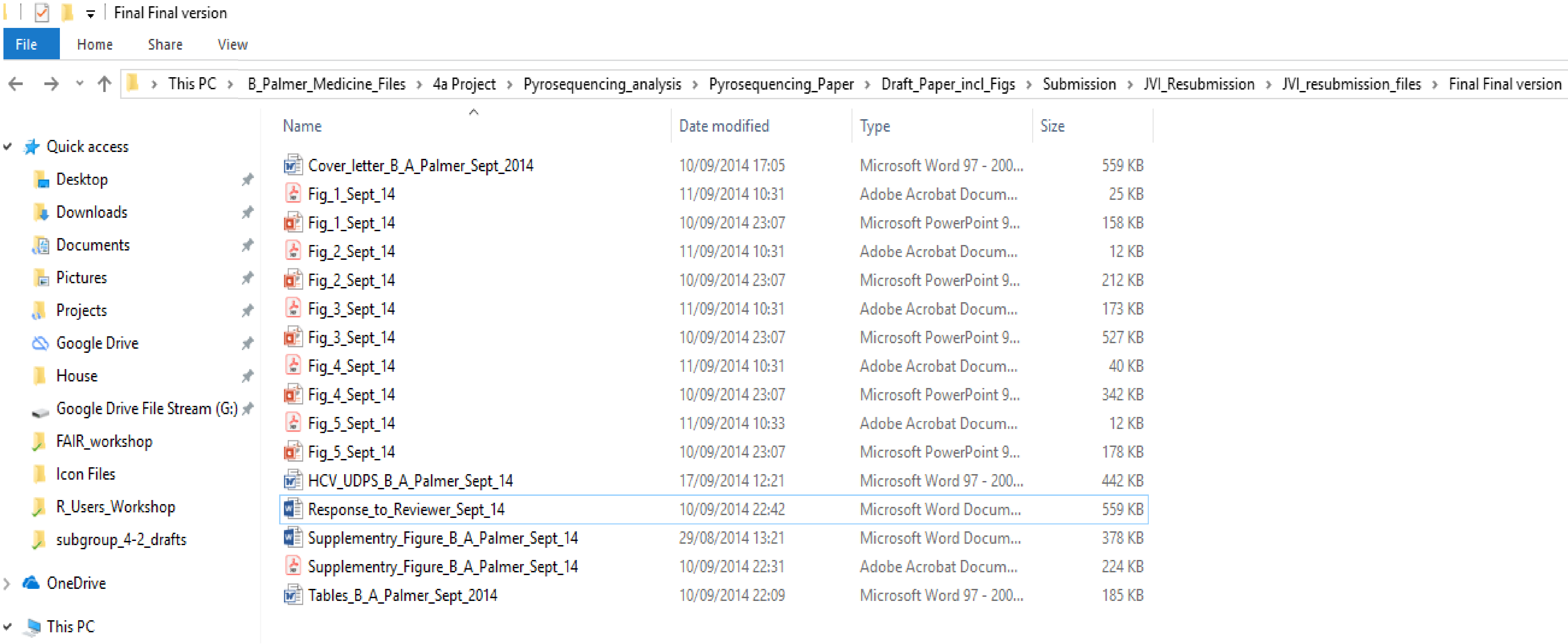

DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
THIS IS GOING TO BE PAINFUL,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

THIS PERSON IS likely to be YOU BTW!!

Jorge Cham | www.phdcomics.com

Still haven’t found what I’m looking for

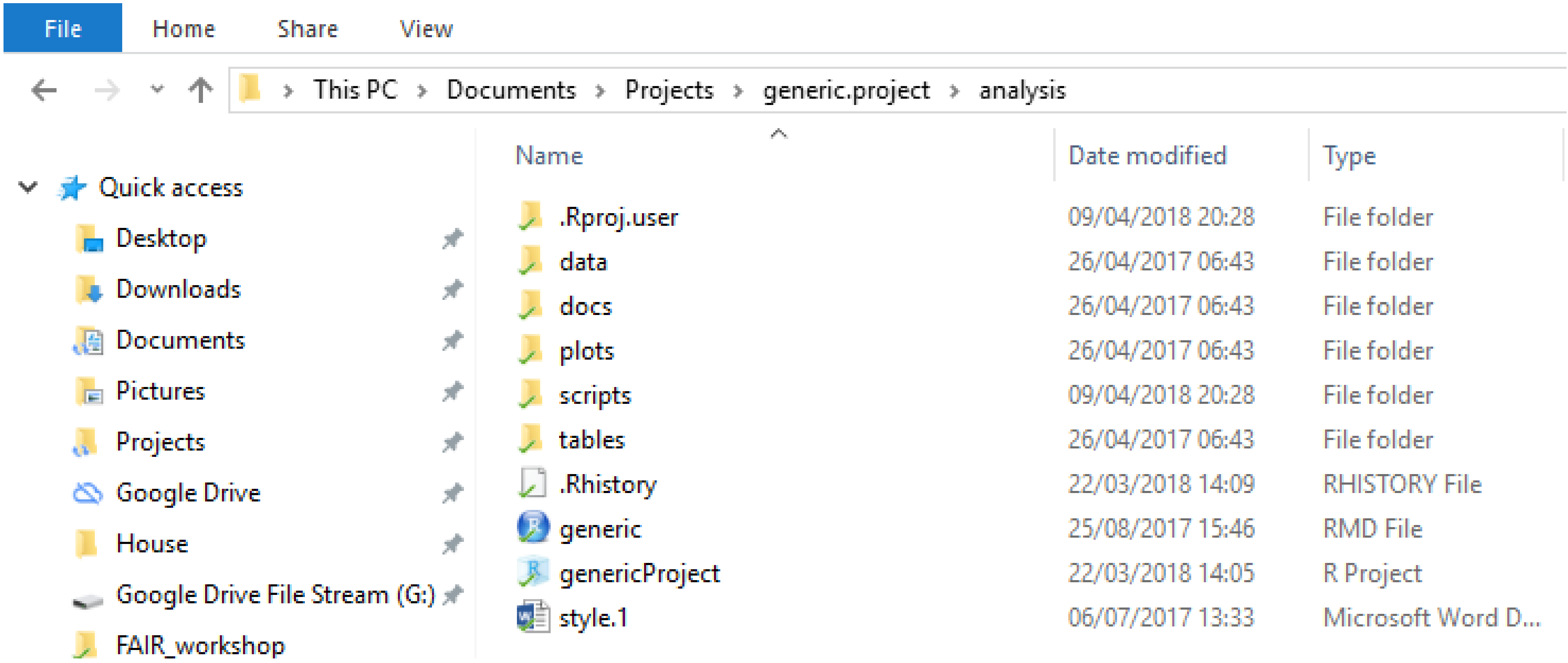
- **Help your future-self**


Final Final version
File
Home
Share
View
This PC › B_Palmer_Medicine_Files › 4a Project › Pyrosequencing_analysis › Pyrosequencing_Paper › Draft_Paper_incl_Figs › Submission › JVI_Resubmission › JVI_resubmission_files › Final Final version
Quick access
Desktop
Downloads
Documents
Pictures
Projects
Google Drive
House
Google Drive File Stream (G:)
FAIR_workshop
Icon Files
R_Users_Workshop
subgroup_4-2_drafts
OneDrive
This PC

Name	Date modified	Type	Size
Cover_letter_B_A_Palmer_Sept_2014	10/09/2014 17:05	Microsoft Word 97 - 200...	559 KB
Fig_1_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	25 KB
Fig_1_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	158 KB
Fig_2_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	12 KB
Fig_2_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	212 KB
Fig_3_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	173 KB
Fig_3_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	527 KB
Fig_4_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	40 KB
Fig_4_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	342 KB
Fig_5_Sept_14	11/09/2014 10:33	Adobe Acrobat Docum...	12 KB
Fig_5_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	178 KB
HCV_UDPS_B_A_Palmer_Sept_14	17/09/2014 12:21	Microsoft Word 97 - 200...	442 KB
Response_to_Reviewer_Sept_14	10/09/2014 22:42	Microsoft Word Docum...	559 KB
Supplementry_Figure_B_A_Palmer_Sept_14	29/08/2014 13:21	Microsoft Word Docum...	378 KB
Supplementry_Figure_B_A_Palmer_Sept_14	10/09/2014 22:31	Adobe Acrobat Docum...	224 KB
Tables_B_A_Palmer_Sept_2014	10/09/2014 22:09	Microsoft Word 97 - 200...	185 KB

Define a generic project structure

- STEP 1: Give your research projects a shared structure

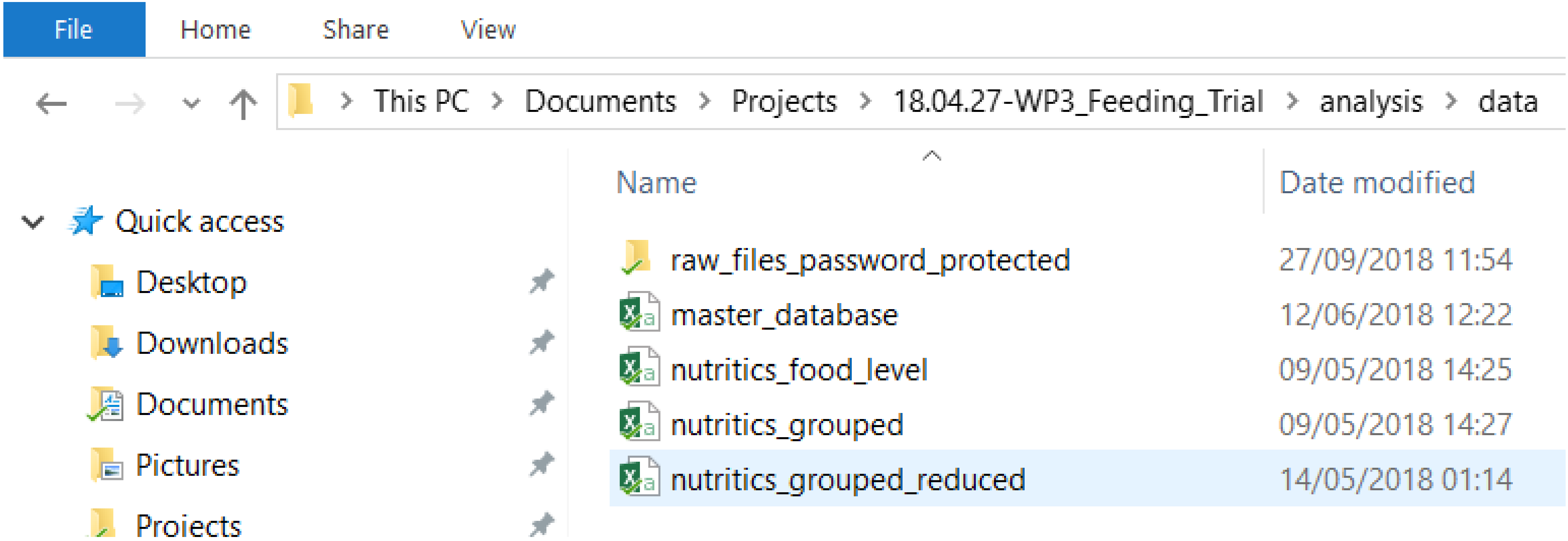


This PC › Documents › Projects › generic.project › analysis

Name	Date modified	Type
.Rproj.user	09/04/2018 20:28	File folder
data	26/04/2017 06:43	File folder
docs	26/04/2017 06:43	File folder
plots	26/04/2017 06:43	File folder
scripts	09/04/2018 20:28	File folder
tables	26/04/2017 06:43	File folder
.Rhistory	22/03/2018 14:09	RHISTORY File
generic	25/08/2017 15:46	RMD File
genericProject	22/03/2018 14:05	R Project
style.1	06/07/2017 13:33	Microsoft Word D...

Give your files informative names

- STEP 1: Give your research projects a shared structure



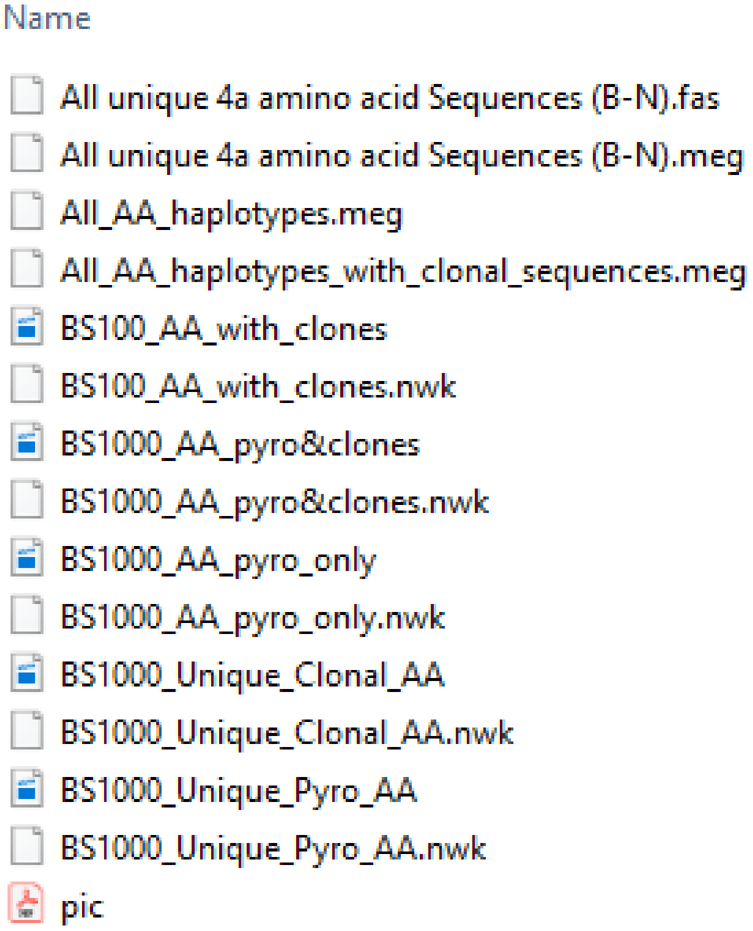
This PC › Documents › Projects › 18.04.27-WP3_Feeding_Trial › analysis › data

Name	Date modified
raw_files_password_protected	27/09/2018 11:54
master_database	12/06/2018 12:22
nutritics_food_level	09/05/2018 14:25
nutritics_grouped	09/05/2018 14:27
nutritics_grouped_reduced	14/05/2018 01:14

Everything in its right place

- STEP 2: Make you file names machine readable, human readable and work with default ordering

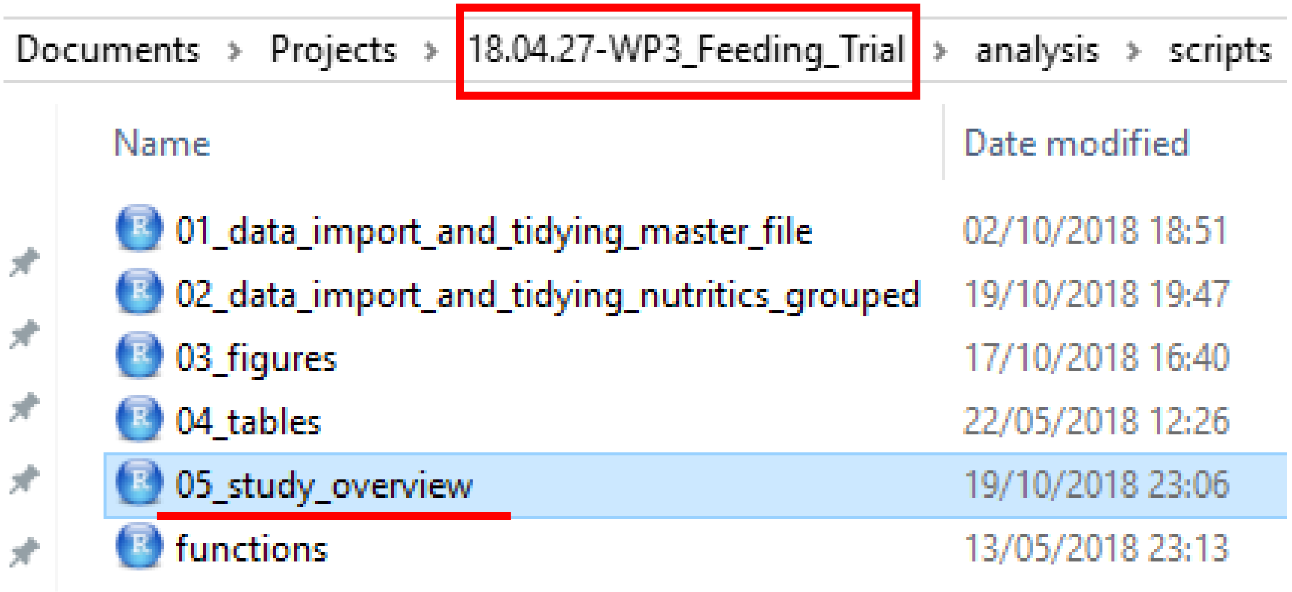
NO



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



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

Outline a file naming convention

Machine readable:

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

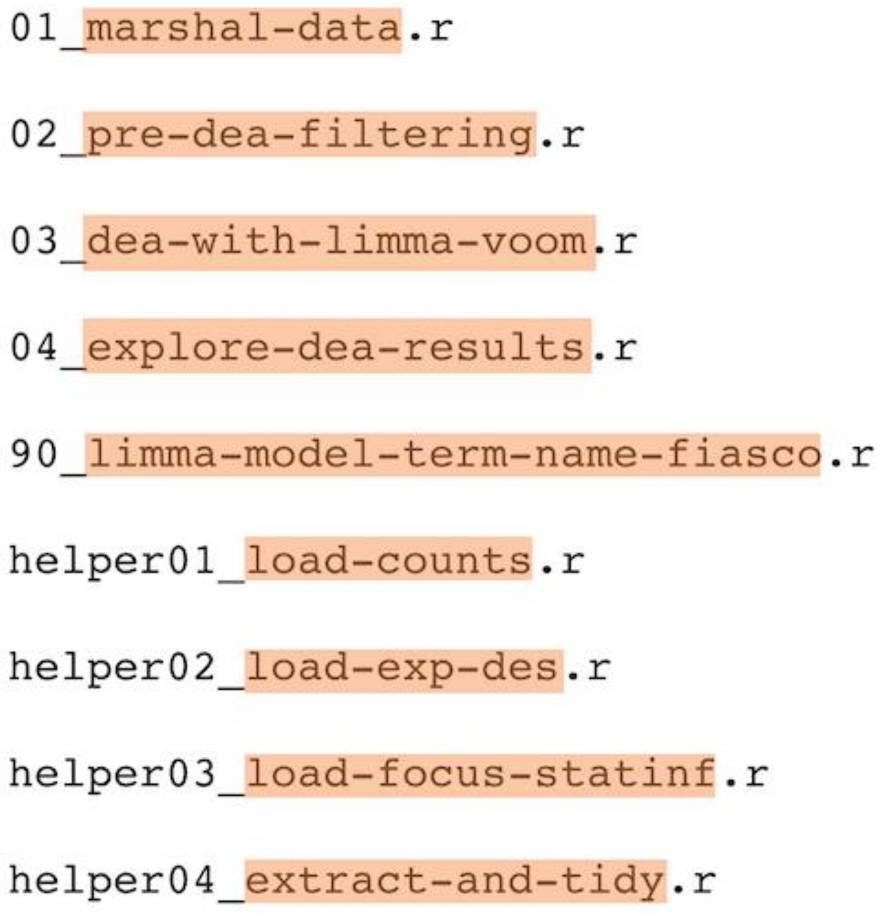
Human readable:

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

Metadata:

Separate with underscores ("_")

- Avoid punctuation
- Remove case-sensitivity

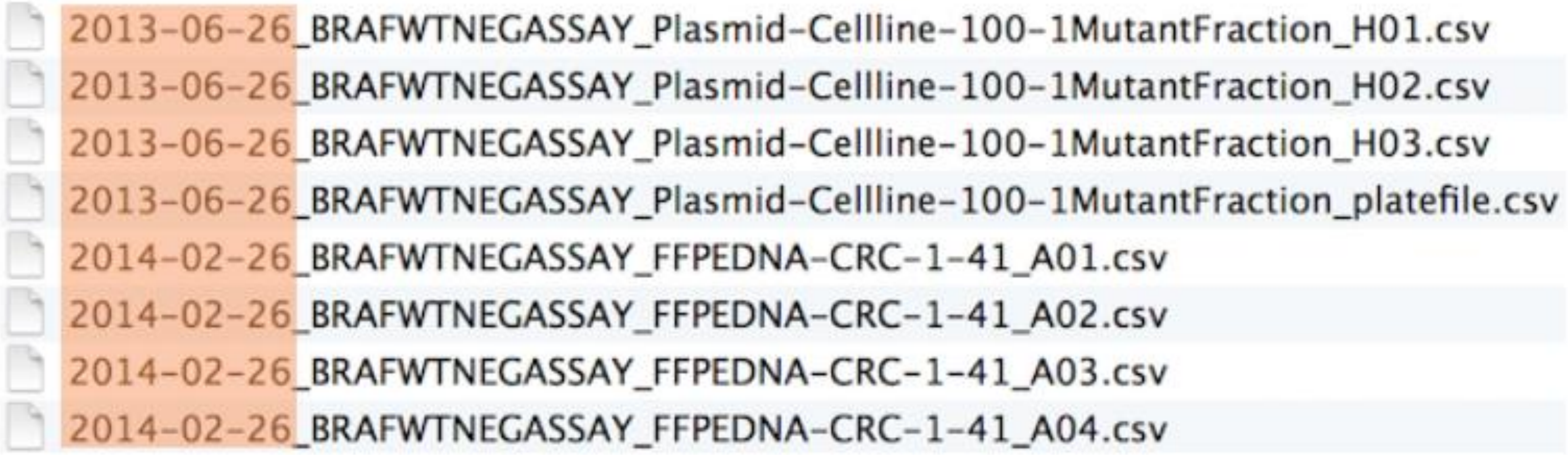


```
01_marshal-data.r

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
```

datacarpentry.org/rr-organization1/01-file-naming/

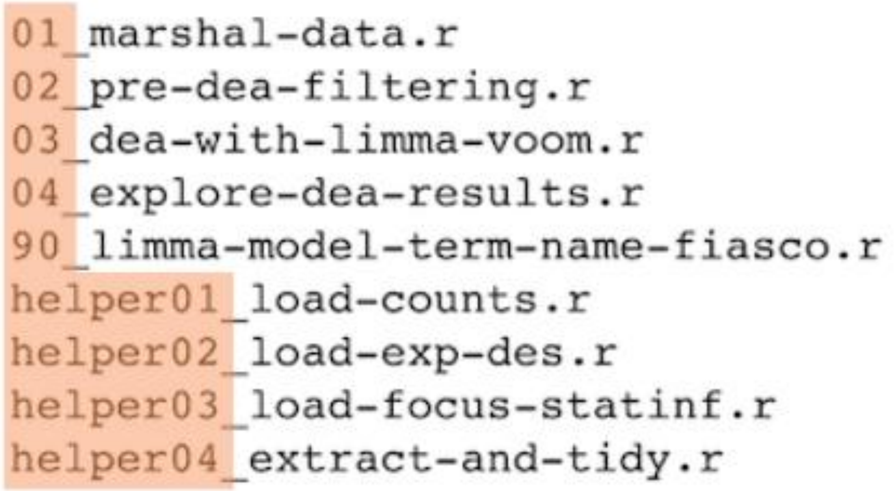
Outline a file naming convention

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:



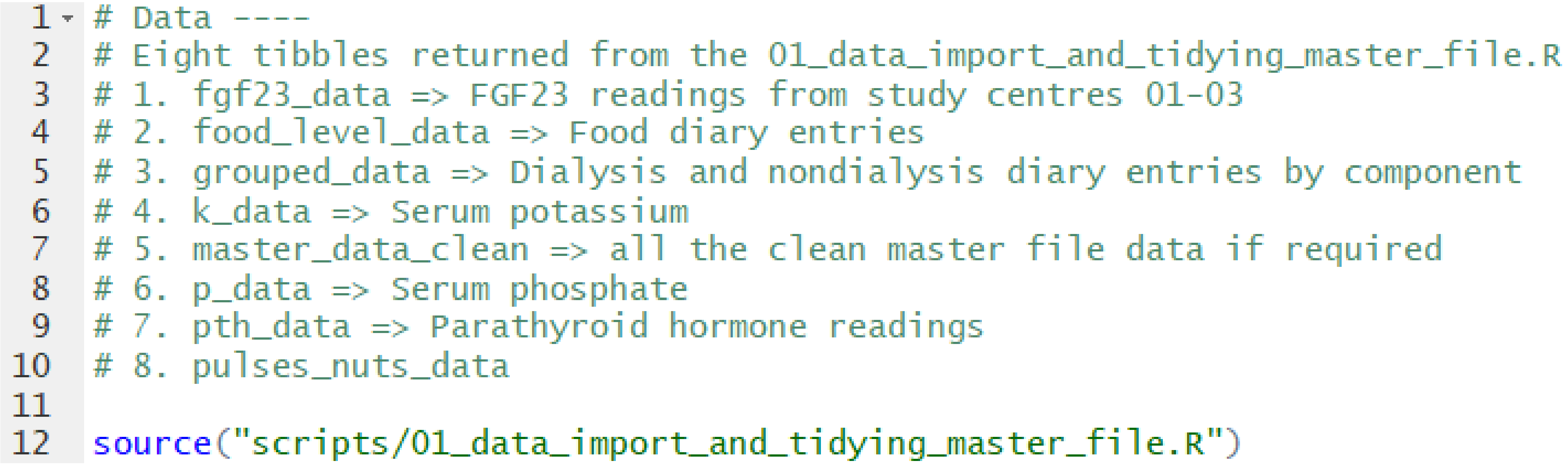
```
01_marshal-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

- Reopen ~/morning_session/data
  - Anything we can improve on here?

datacarpentry.org/rr-organization1/01-file-naming/

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



```r
# Data ----
# Eight tibbles returned from the 01_data_import_and_tidying_master_file.R
# 1. fgf23_data => FGF23 readings from study centres 01-03
# 2. food_level_data => Food diary entries
# 3. grouped_data => Dialysis and nondialysis diary entries by component
# 4. k_data => Serum potassium
# 5. master_data_clean => all the clean master file data if required
# 6. p_data => Serum phosphate
# 7. pth_data => Parathyroid hormone readings
# 8. pulses_nuts_data

source("scripts/01_data_import_and_tidying_master_file.R")
```

Don’t Do What Donny Don’t Does!!

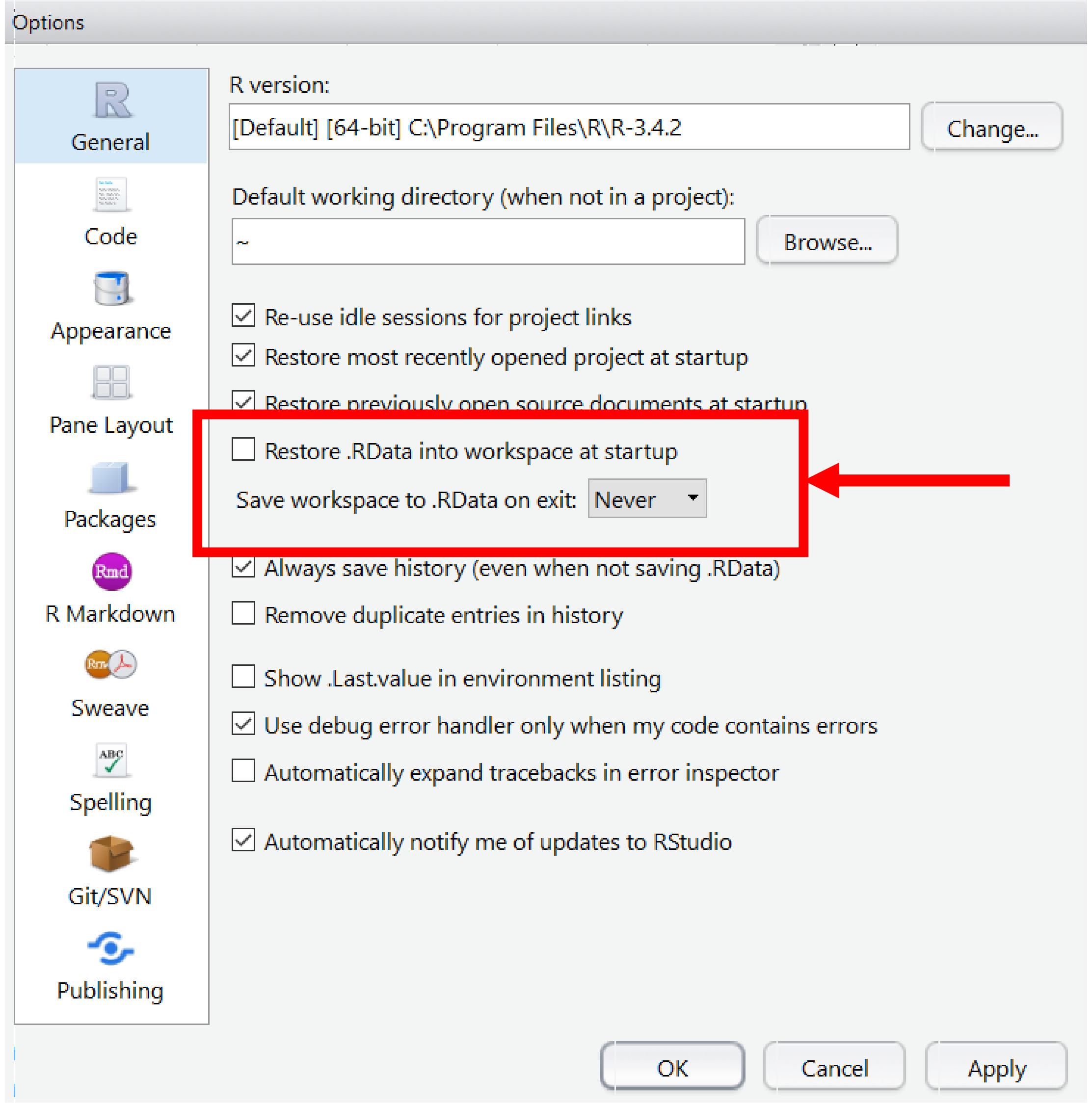
Donny Don’t:

- Start your script with… `setwd()`

Donny Don’t:

- Start your script with… `rm(list = ls())`

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

R Markdown

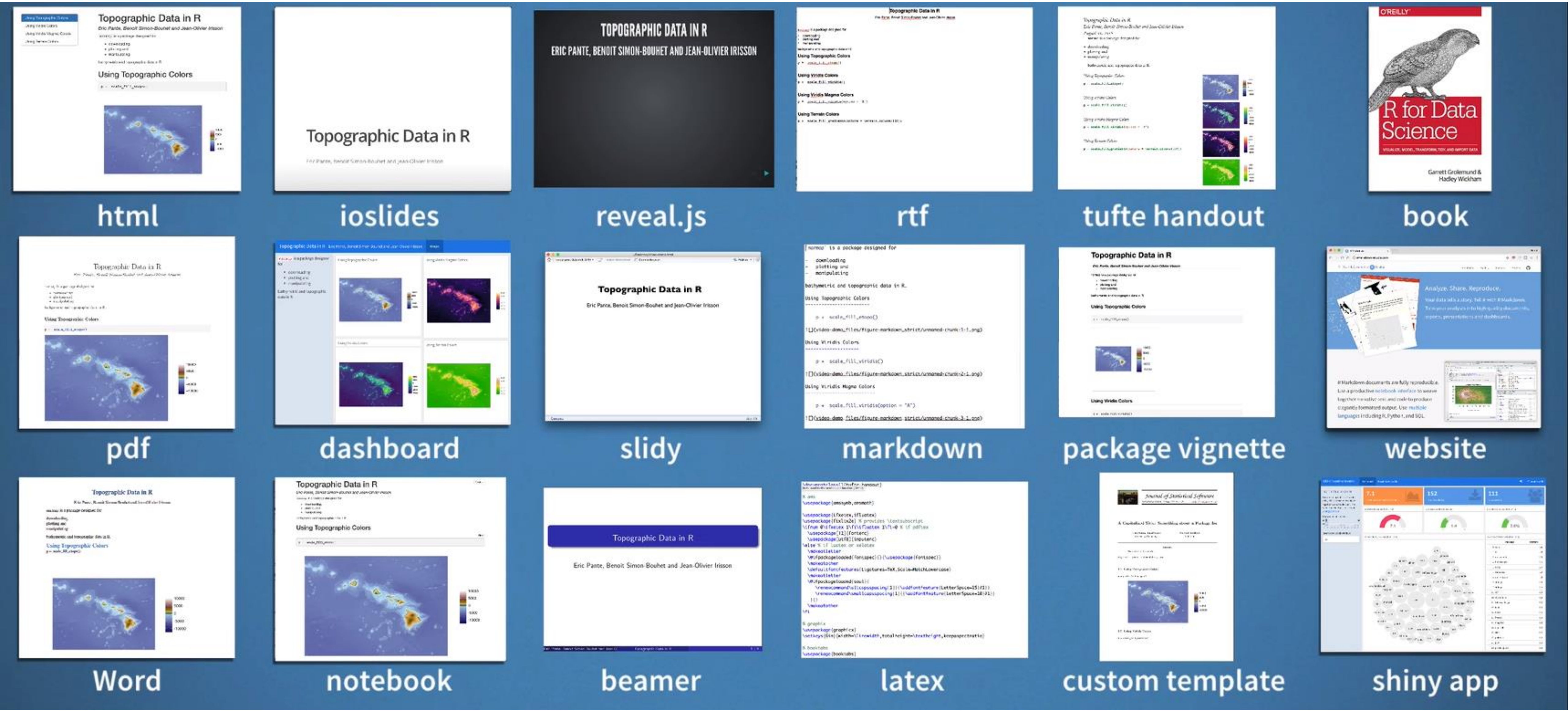
- R Markdown combines the code you wrote, the output produced and you own comments

- You can view it as a digital lab notebook, where you are both recording what you're doing, and what you were thinking while you were thinking it!

- R Markdown outputs can take many forms
    - Word documents, PDFs, slideshows etc.

- Once created the .Rmd file get sent to knitr, which executes the chunks of code and creates a new markdown document (.md)
    - this is then processed by pandoc which creates the finished file
        - knitr and pandoc are external websites

What has R Markdown ever done for us?



R Markdown

YAML header

~~~
---
title: "Diamond sizes"
date: 2016-08-25
output: html_document
---
~~~

Chunks of code

~~~
```{r setup, include = FALSE}
library(ggplot2)
library(dplyr)
smaller <- diamonds %>%
filter(carat <= 2.5)
```
~~~

Plain text with integrated outputs from R

~~~
We have data about `r nrow(diamonds)`
diamonds. Only
`r nrow(diamonds) - nrow(smaller)` are
larger than
2.5 carats. The distribution of the
remainder is shown below:
~~~

Chunks of code

~~~
```{r, echo = FALSE}
smaller %>%
ggplot(aes(carat)) +
geom_freqpoly(binwidth = 0.01)
```
~~~