Making your Workflow work for you

Chloe Yap

MD-PhD candidate

Cognitive Health Genomics (Gratten)

MRI-UQ, PCTG-IMB

I begin today by acknowledging the Traditional Custodians of the land on which we meet, and pay respects to their elders past, present and emerging.

About me


Gyarados LV.55
HP 130
STAGE 1 Evolves from Magikarp
Pokémon Rumble
Thrash Tackle 80
Gyarados does 20 damage to itself.
Pokémon Rumble
weakness ×2
resistance -20
retreat cost
©2009 Pokémon/Nintendo
6/16

About me

Background

Undergraduate: BSc (BiomedSc), spent more time in lab than in lectures …

Postgraduate: MD-PhD, PhD in autism omics (boils down to data analysis)

R experience

Been "in the game" since end-2014

Organic, on-the-job learning

Eternally indebted to my "teachers" …

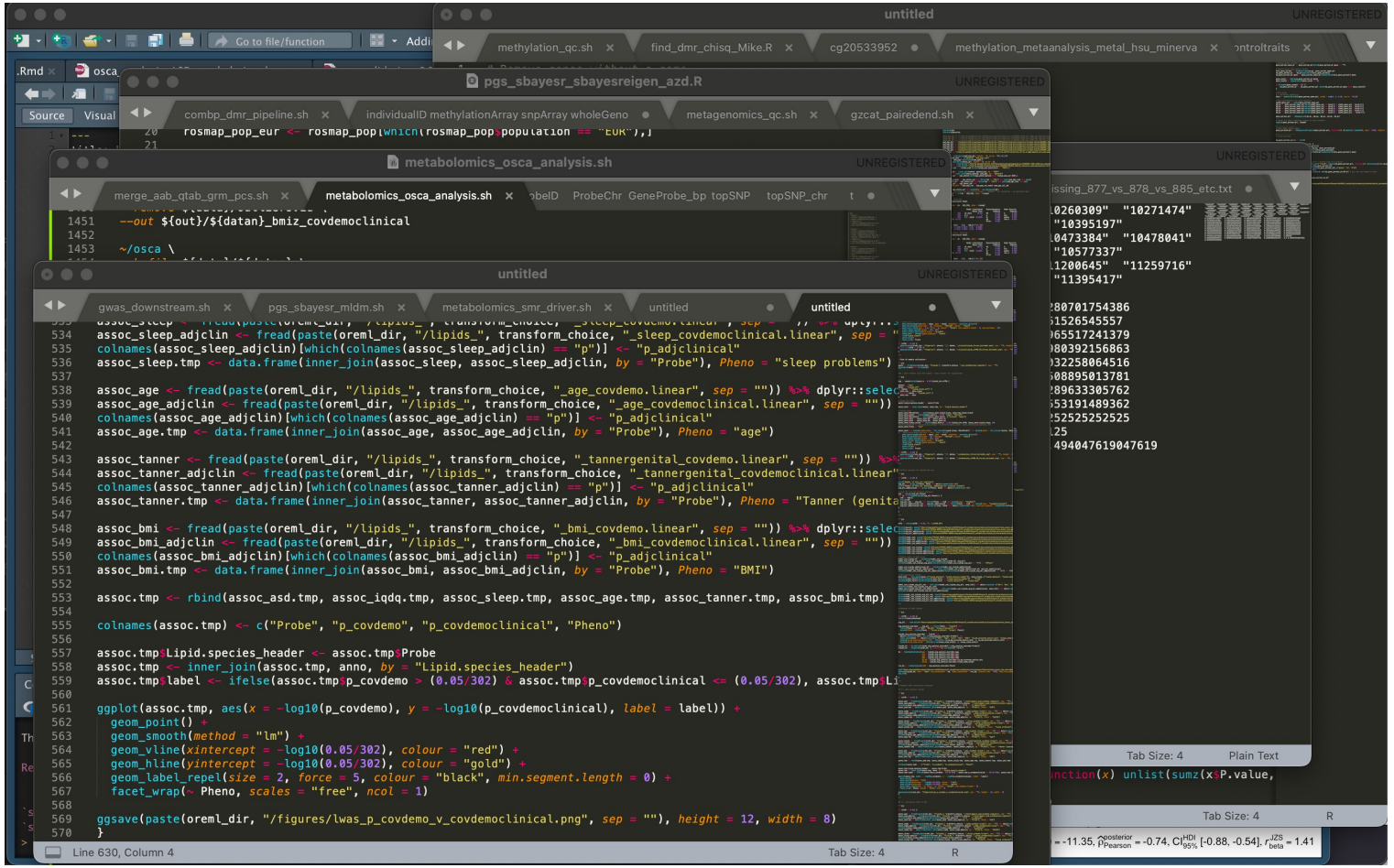
Priorities born of necessity

Multiple projects, distractors, on/off work → readability, organised directories, "centralised" results, working with manuscript output in mind

Running many iterations of similar analyses → automate as much as possible

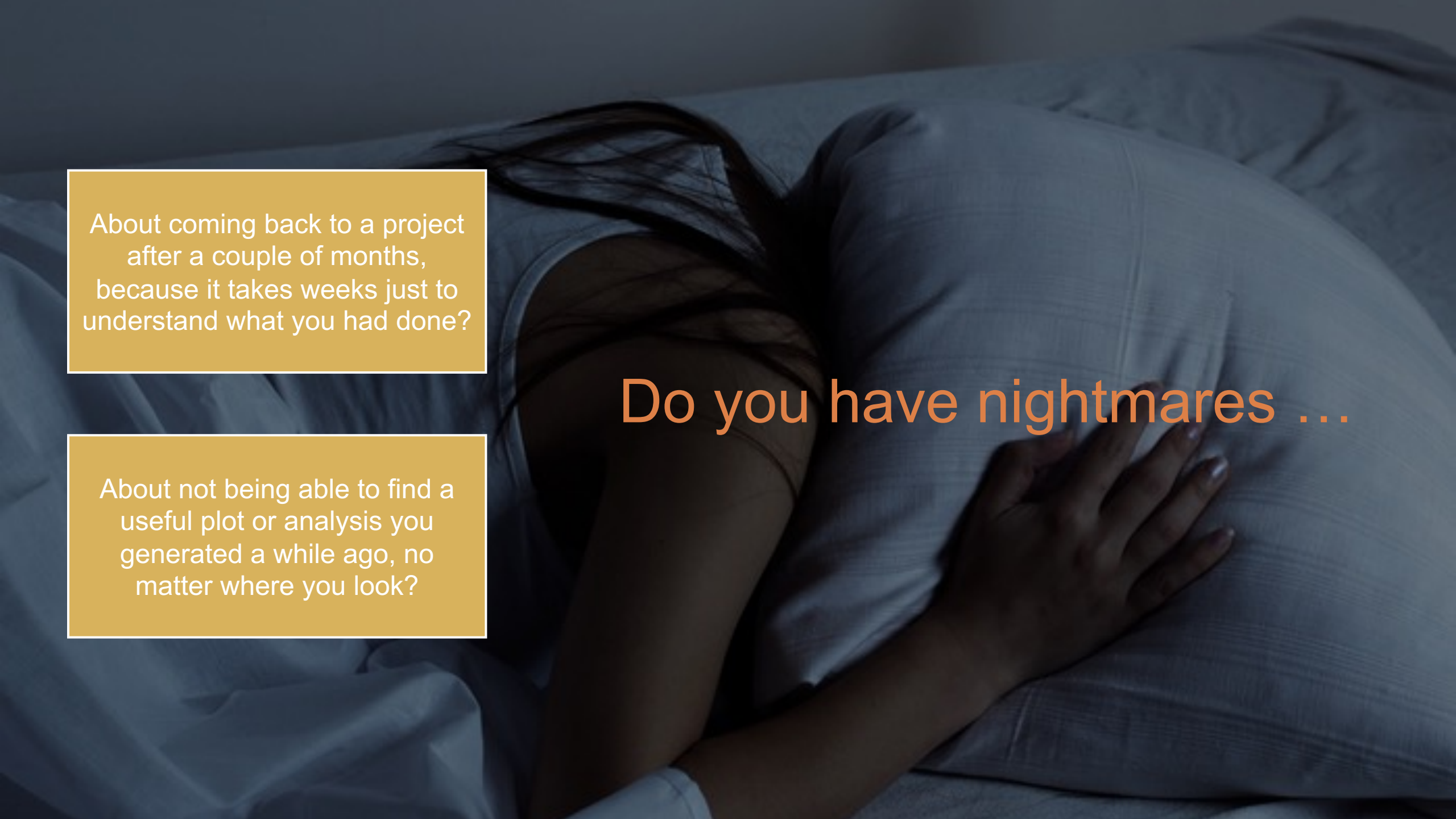
Aims of this talk

- I *think* I'm relatively efficient(?) → some insights into how I operate
- Disclaimer: many things I do are pretty hacky and not best practice! But they're "good enough"…



Outline

- Directory set-up, readability
- ~Automating repetitive analyses
- Results output: RMarkdown
- Keeping manuscript in mind
  - Plots



About coming back to a project after a couple of months, because it takes weeks just to understand what you had done?

About not being able to find a useful plot or analysis you generated a while ago, no matter where you look?

Do you have nightmares …

Directory set-up: Principles

- Why?
  - Readability, much easier to start back where you let off
  - Helps to trace outputs to the data and scripts used
- Principles
  - Modular organisation, 1 folder per analysis
    - ***If it will be a new figure or results section → new folder!***
    - ***Try to keep a driver script for an analysis (.sh script with qsub or all .Rmd)***
    - ***If you've put enough effort in typing code, it's probably worth filing away***
  - Consistency
  - Readability (?hackable?) → make and update README files
- See: Noble 2009 *PLOS Computational Biology* **A Quick Guide to Organizing Computational Biology Projects**
  https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

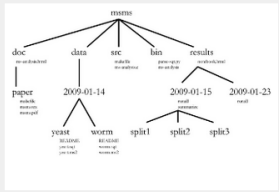
EDUCATION

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble

Published: July 31, 2009 • https://doi.org/10.1371/journal.pcbi.1000424

Figures



Citation: Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. https://doi.org/10.1371/journal.pcbi.1000424

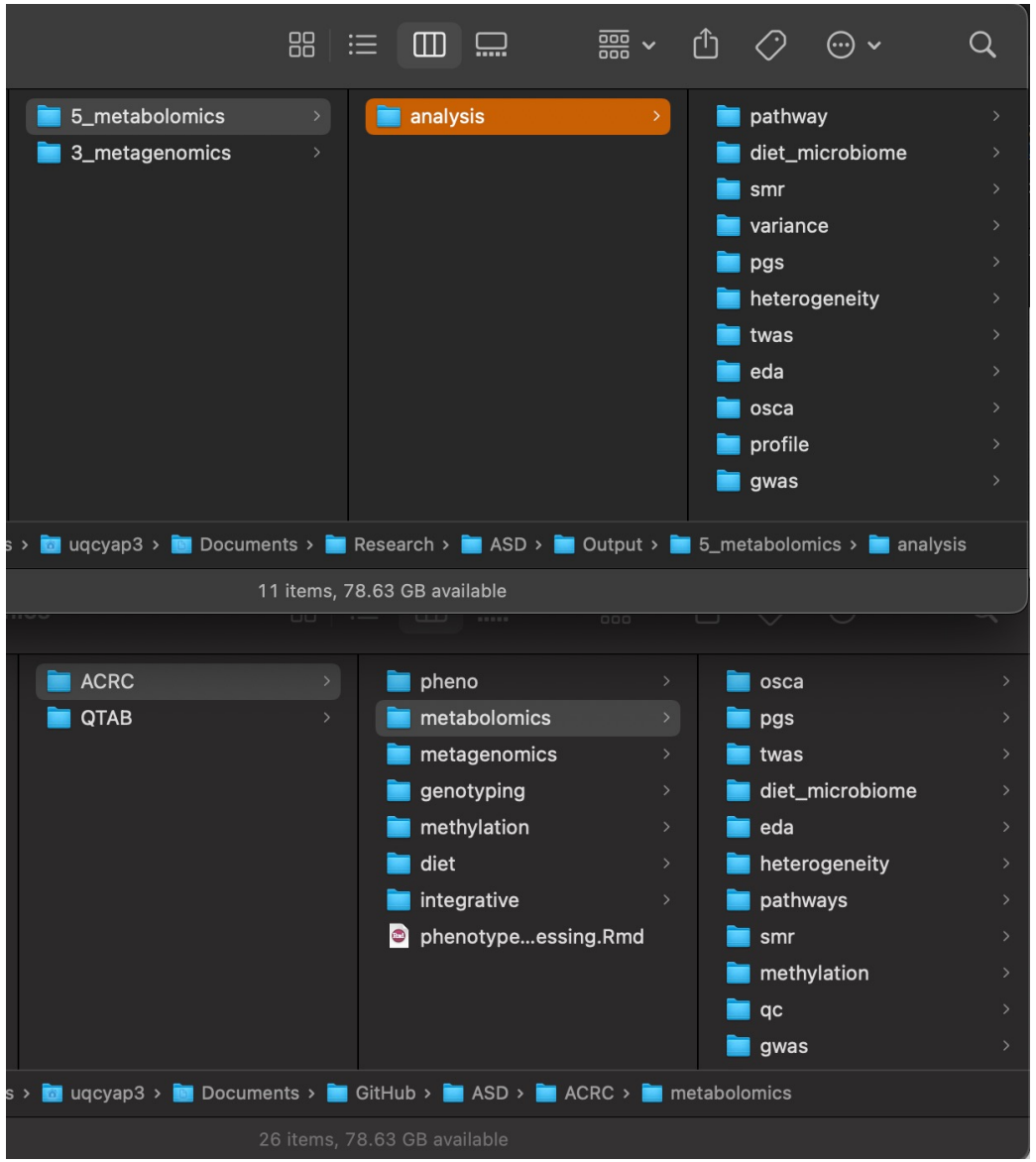
Editor: Fran Lewitter, Whitehead Institute, United States of America

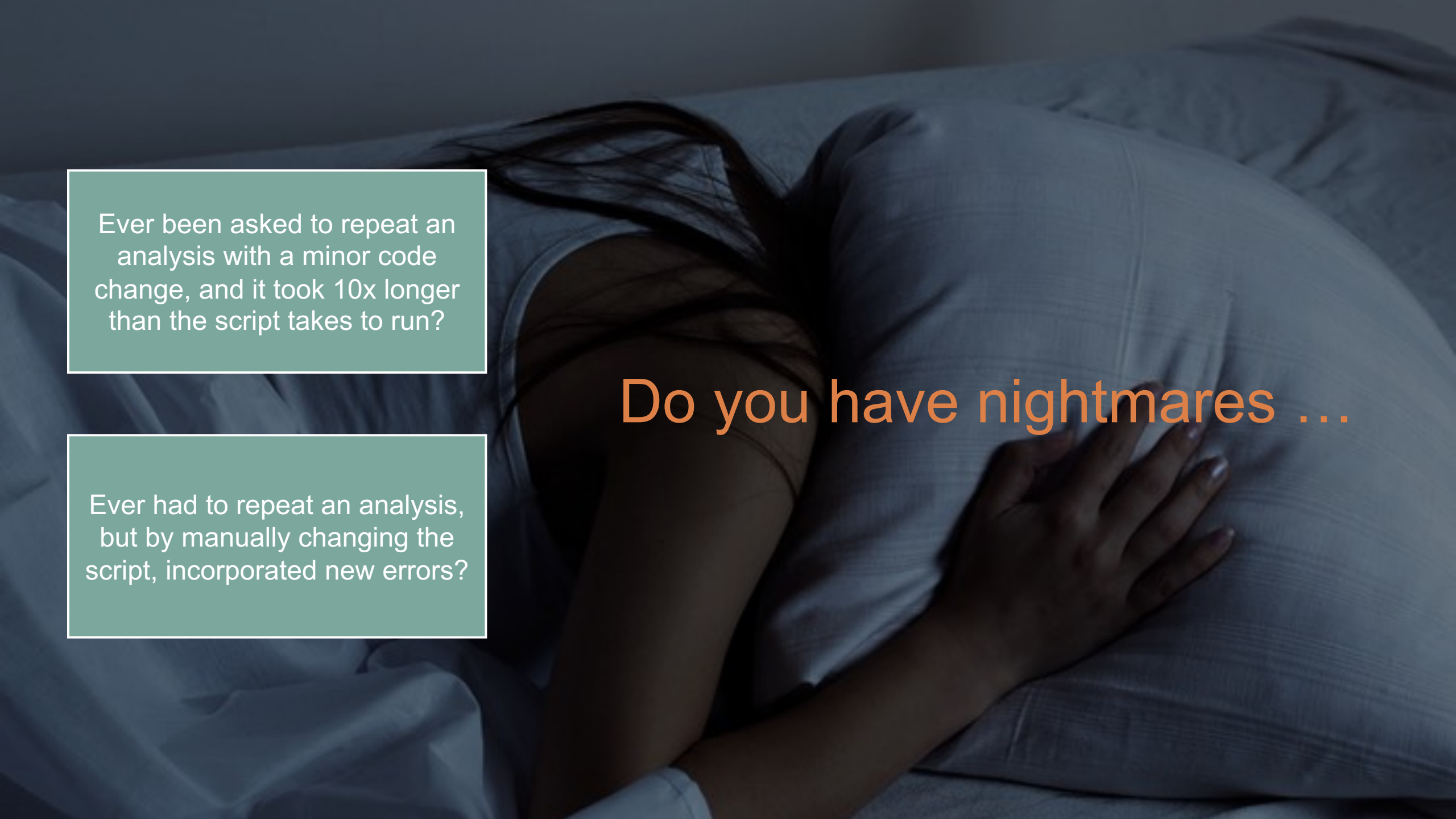
Published: July 31, 2009

An example:

For each project …

- Data/ Output/ Scripts/
  - Within each, aim for ~identical folder structure eg.
    - Data/gwas
    - Output/gwas
    - Scripts/gwas
- Write-up/
  - Manuscripts/
  - Presentations/
  - Applications/, etc.
- File naming
  - Logical
  - Date: YYMMDD so auto-sorts!





Ever been asked to repeat an analysis with a minor code change, and it took 10x longer than the script takes to run?

Ever had to repeat an analysis, but by manually changing the script, incorporated new errors?

Do you have nightmares …

Automating repetitive analyses

- Why?
  - No matter how well you do it the first time, there's a 99% chance* you'll have to rerun the analysis …
  - Massive time-saver (and hair-saver!)
  - Reduces errors
- Principles
  - Generalisability
  - Flexibility

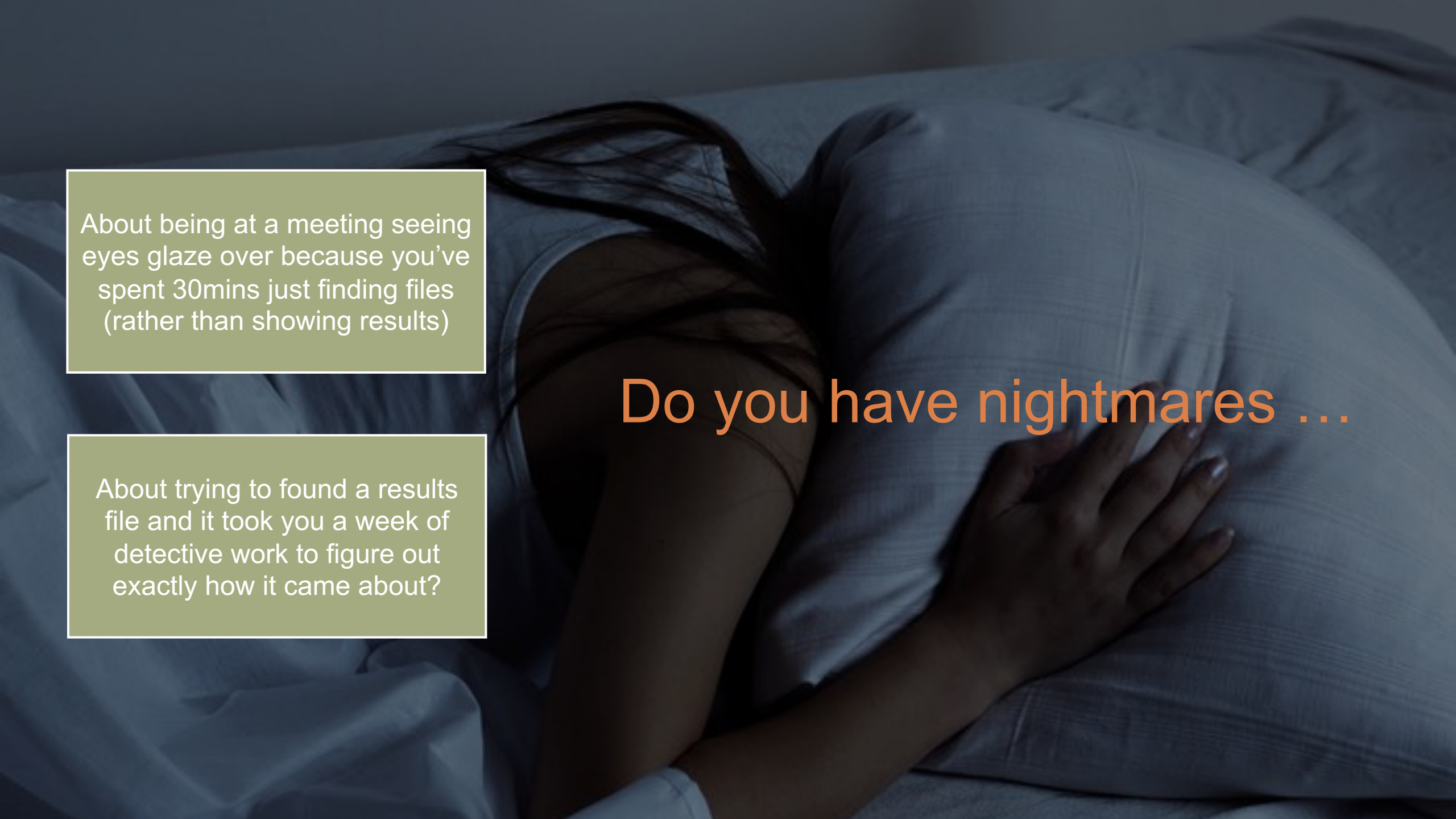
*not backed by empirical data

Automating repetitive analyses

- Create “settings”
  - Generate separate .Rmd with different settings
  - Settings reflected in output file name
  - Eg. “filen” variable
- source() scripts at the start of analysis
  - If there is common code at the start of multiple scripts
  - If input file is being updated
- On HPC → “pipeline” script in bash
  - .sh wrappers of .R scripts

Miscellaneous fun things in analysis

- group_by()
  - %>% summarise()
  - %>% pivot_wider()
  - Performing operations in parallel eg. lm()



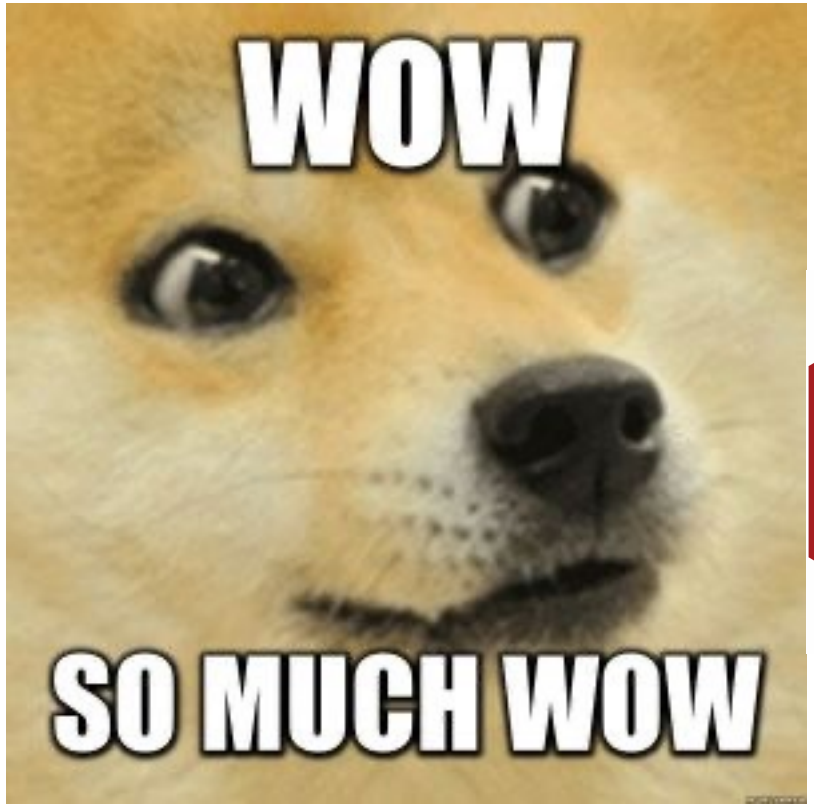
About being at a meeting seeing eyes glaze over because you've spent 30mins just finding files (rather than showing results)

About trying to found a results file and it took you a week of detective work to figure out exactly how it came about?

Do you have nightmares …

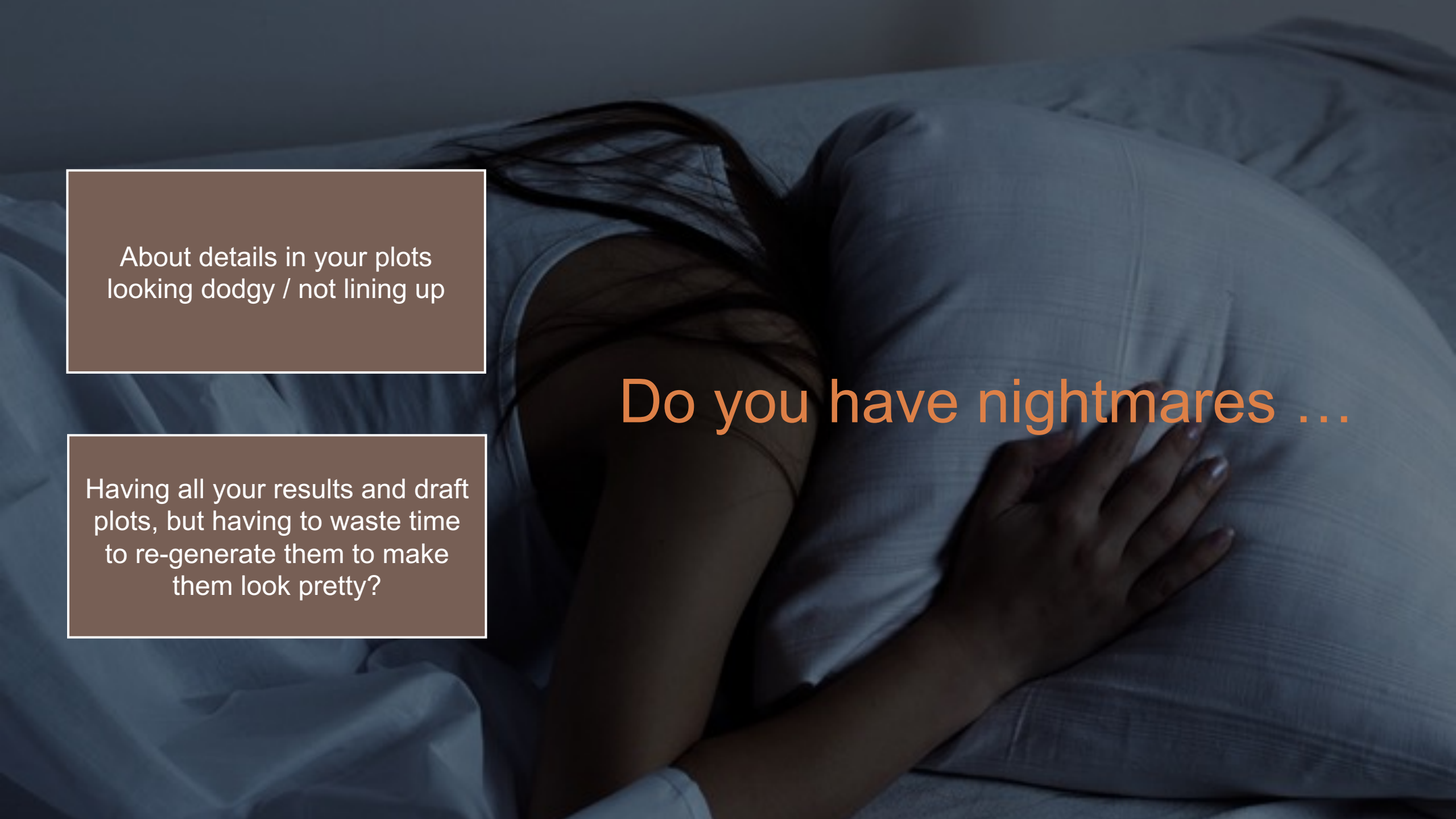
Results output: RMarkdown

- Why?
  - Reproducible research (runnable)
  - Re-run analysis with a single click
  - Easier to present results
  - “Centralised” results → easier to pull into manuscript
  - Much wow, little effort
  - … or amazing things with more effort
    - Supplementary Materials
    - “Website” of results
    - Interactivity, etc.



Results output: RMarkdown

- “Keeping manuscript in mind”
  - Set up
    - QC .Rmd → exploratory data analysis, separate so doesn’t clutter results
    - analysis .Rmd → results that are more geared towards the manuscript
- HPC users
  - I can never get RStudio to reliably work on HPC …
  - Solutions:
    - GitHub on HPC + local → update HPC version
    - Write code locally as an .Rmd and test interactively on HPC → generate .Rmd html by copying across .Rmd from local → HPC



Do you have nightmares …

About details in your plots looking dodgy / not lining up

Having all your results and draft plots, but having to waste time to re-generate them to make them look pretty?

Ironically ugly slide?!?

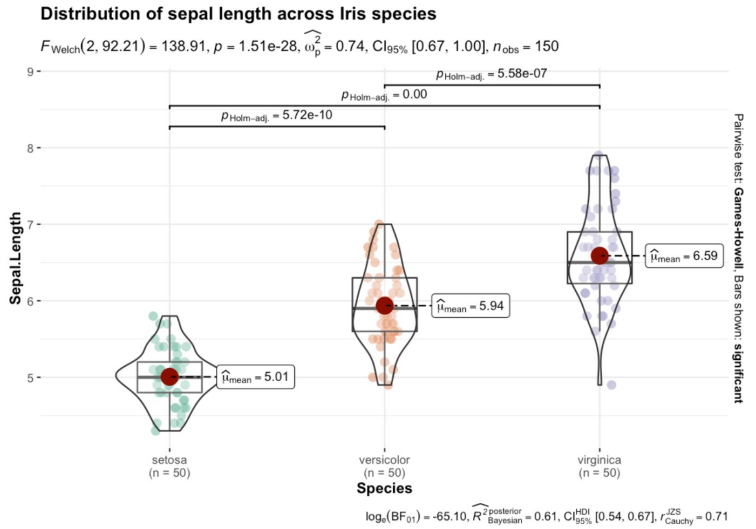
Plots

- Preferred packages
  - ggplot
  - ggstatsplot
  - ggrepel for labels
- Design preferences
  - Colour palette: Dark2, various wesanderson palettes
    - Choose early (so you don't have to go back and spend time re-making plots)
    - Keep it simple
    - Keep colour scheme consistent across figures
  - Remove default grey plot background
  - Export as .svg → fiddly editing on Inkscape (or Illustrator)

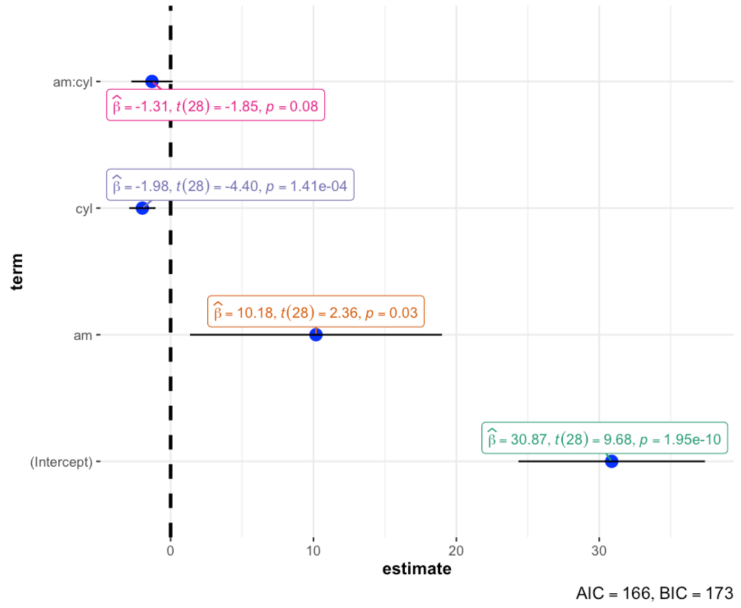
ggstatsplot

https://indrajeetpatil.github.io/ggstatsplot/

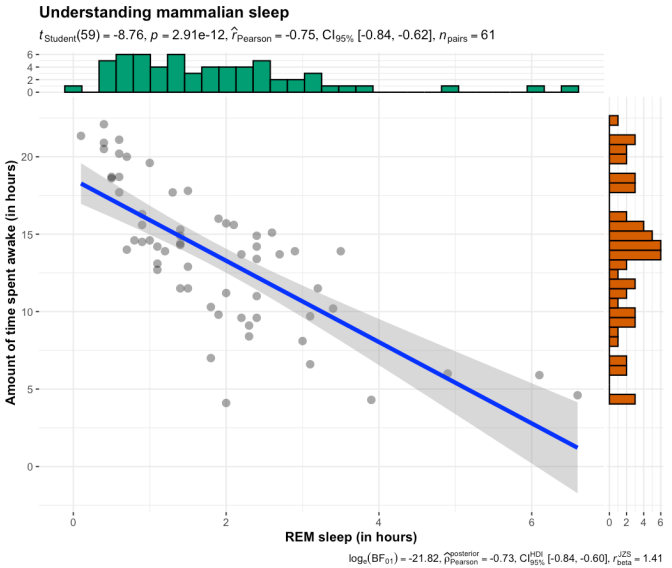

```
ggbetweenstats(
  data  = iris,
  x     = Species,
  y     = Sepal.Length,
  title = "Distribution of sepal length across Iris species"
)
```



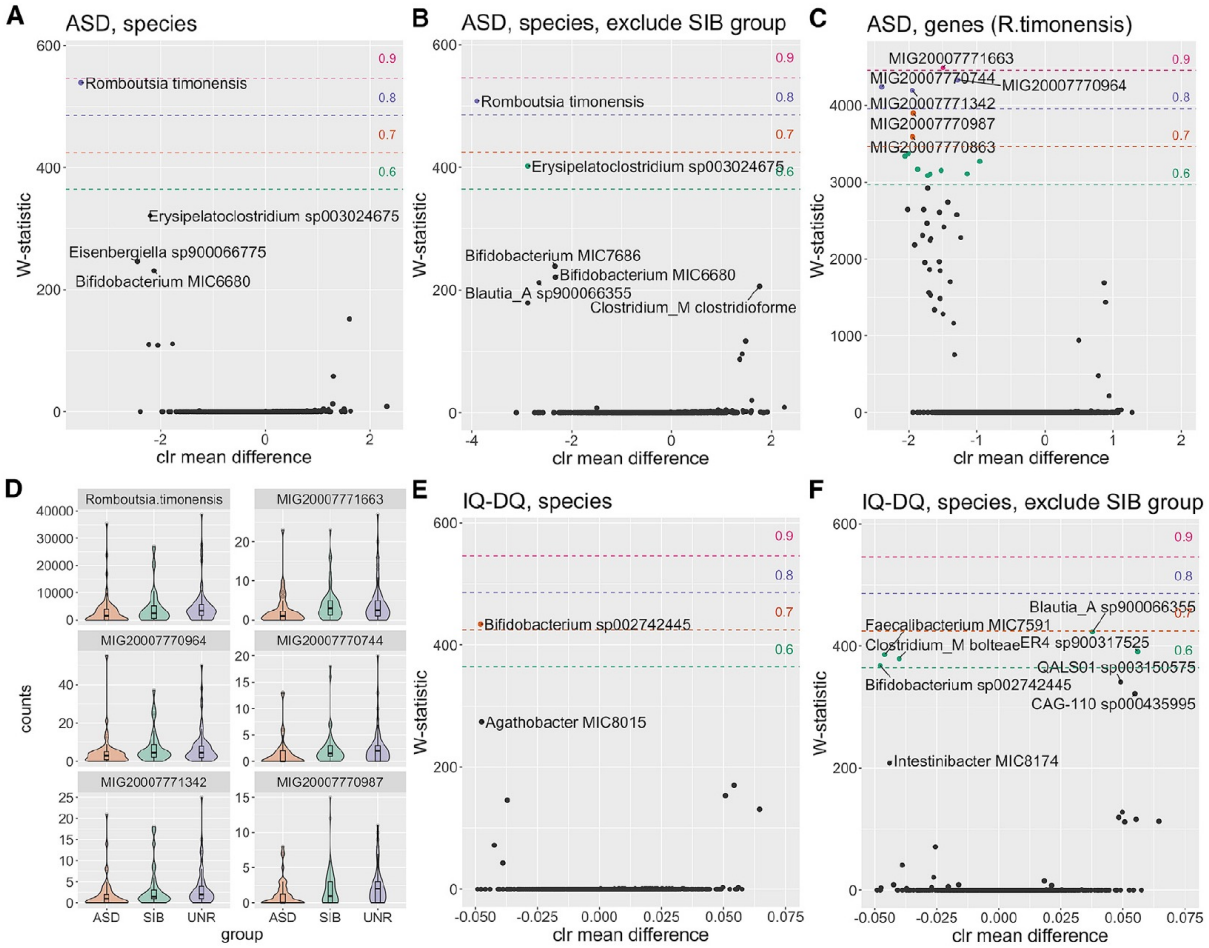

```
## model
mod <- stats::lm(formula = mpg ~ am * cyl, data = mtcars)

ggcoefstats(mod)
```

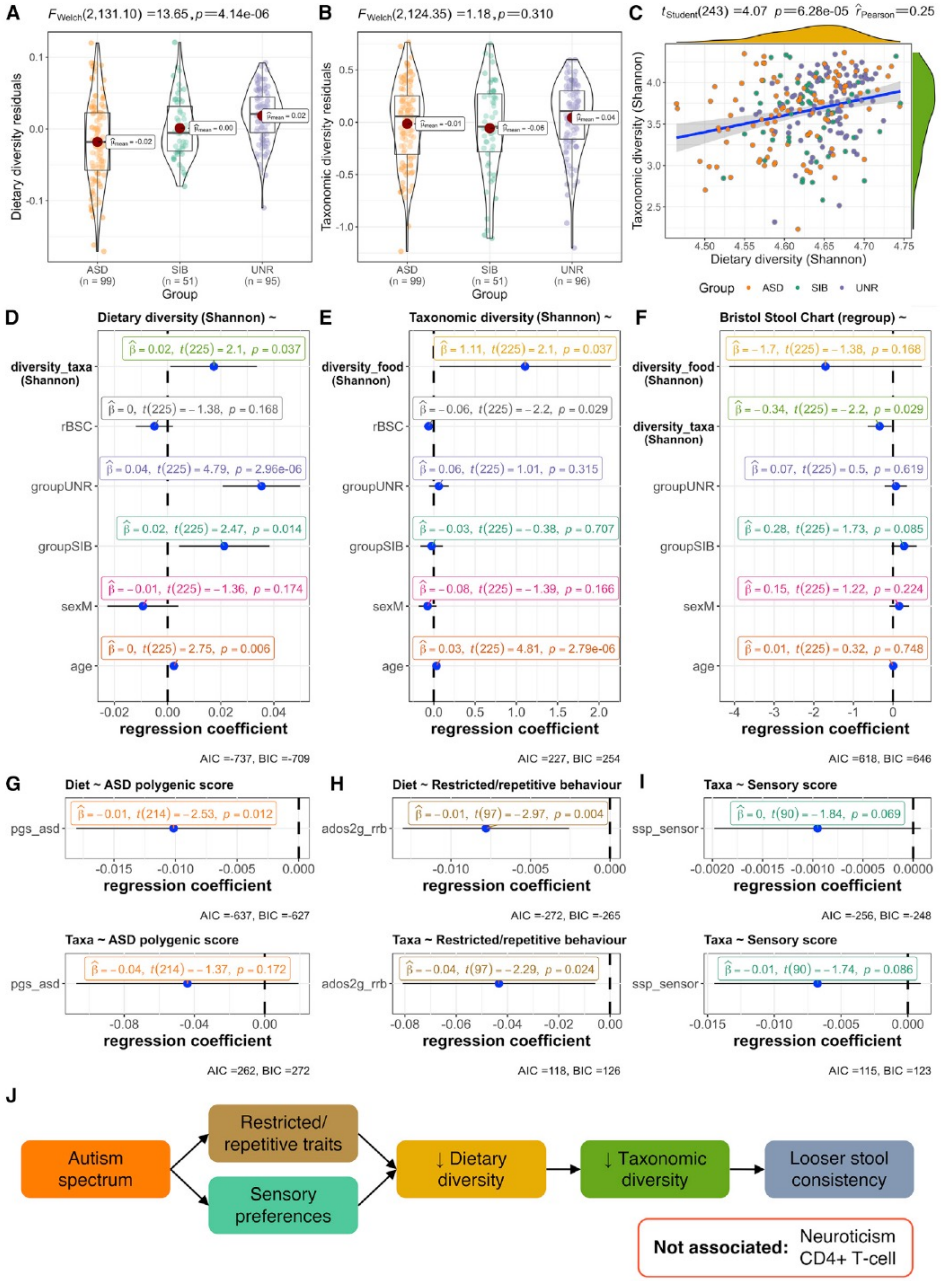



```
ggscatterstats(
  data  = ggplot2::msleep,
  x     = sleep_rem,
  y     = awake,
  xlab  = "REM sleep (in hours)",
  ylab  = "Amount of time spent awake (in hours)",
  title = "Understanding mammalian sleep"
)
```



Colour coding





.svg → Inkscape

What is Inkscape?

- Vector graphics software, like Illustrator
  - “infinite zoom”
- Free and open-source! (unlike Illustrator)

Use

- Complicated figures
  - Eg. R figures + drawings/illustrations
- Figure editing that is fiddly/annoying in R
  - Eg. overlapping labels, annotations

Handy resources

- StackOverflow
- R Bloggers
- STHDA
- R Graph Gallery
- R Memes for Statistical Fiends (light content)
- swirl, etc.
- Noble 2009 *PLOS Computational Biology* **A Quick Guide to Organizing Computational Biology Projects**
  https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424