**R for Reproducible data wragling**

**What happens before the stats: the power of R Tidyverse for wrangling, cleaning, and exploring your data.**

****

**Rationale:** Most biologists use R for data analysis and visualization. But most still rely on spreadsheets to go from the raw data to data tables suited for specific uses (e.g., a statistical test or plotting). This makes analysis messy, unnecessarily frustrating, and, most importantly, not reproducible. Here we will learn how to do it all within the same workflow using the powerful and intuitive functions of the Tidyverse (includes ggplot!).

**Goal**: reproducible research & higher efficiency in data wrangling (& more fun!)

**Target audience**:

* R novices
* R intermediate users that are not familiar with the Tidyverse

**This workshop will cover**:

* The concept of reproducible research and the importance of always manipulating data through (R) scripts
* Introduction to the powerful and versatile [Tidyverse](https://www.tidyverse.org/)
* Working example of R script (+ associated dummy data set) to simulate a data analysis workflow that covers the most useful functions to explore, check, correct, reshape, and visualize biological data

**This workshop will NOT cover**:

* Download and installation of R and R Studio
* Explanation of the basics of the R coding language.

Therefore, to make the most of the workshop, we recommend that you take care of the following **before** the workshop:

* Install R and R Studio, and the tidyverse package
* Familiarized yourself with the following core concepts of R (no need to become an expert, just don’t be overwhelmed by new terminology):
  + Data structures: vectors, matrices, arrays, data frames, and lists
  + Data types (classes): know that they exist and have a rough understanding of what they mean
  + What are R functions
  + What are R packages

**Structure**

The workshop will last a total of 3 hours (180 minutes), broken down into (roughly):

* 30 min: **presentation** to introduce the concept of reproducible research (and how R fits in it), tidyverse and the tidy data format, and overview of the most useful functions
* 60 min: **live coding example** of a typical data wrangling workflow, to go from raw data to data ready for statistical analysis
* 90 min: **your time to code!** Choice of options between 1) repeating what was shown, on your machine, 2) solving a problem formulated by us (e.g., how to accomplish a specific task using the tools presented); 3) further exploring a topic/function/approach of your choice and ask us for help and tips; or 4) bring your own data and questions and get support from the community (us and the participants)!

**Practically – summary of content and execution**

Slide presentation:

* Very brief intro to the concept of reproducible research and the importance of working in scripts (as opposed to manual edits of spreadsheets)
* Advantages of using R from the very beginning and at each step in which any sort of data manipulation is required (not just for the final stats)
  + Why R when one can do most of this in bash/Unix?
    - “safer” approach: R doesn’t modify data files directly, more wiggle room for trial and error (although disadvantages/additional steps when you want to directly access and modify your data – see it as a complementary tool)
    - Convenience to have “all in one place”: seamlessly proceed with data visualization (ggplot!) and statistical testing
    - Deal with biologists: (unlike bioinformaticians) most biologists usually only know R …
* Intro to the Tidyverse: a collection of packages in R that share a common “philosophy” (fit well together) and rely on the concept of “tidy” data
* Explain tidy data structure (and what is not), and “long” and “wide” data format (link to original paper)
* Brief definition of data wrangling: subset, summarize, merge tables, pivot …
* Brief overview of most useful tidyverse functions and approaches to the (mentioned above) most common needs:
  + Create new variables: mutate()
  + Summarize and check data structure): group\_by() + summarize()
  + Subset: filter(), select()
  + Merge: inner\_join(), left\_join(), outer\_join() …
  + Restructure: pivot\_wider(), pivot\_longer()
  + Change/correct values: rename(), replace()
  + Visualize data (plot): ggplot()
  + Text manipulation (regex): {stringr} stringr::str\_subset(), str\_replace(), separate(), …
* Quick intro to the dummy data set (made up by me with features to exemplify the above functions)