Tidyverse

Oxford Biomedical Data Science Training Programme
University of Oxford
2020-05-27

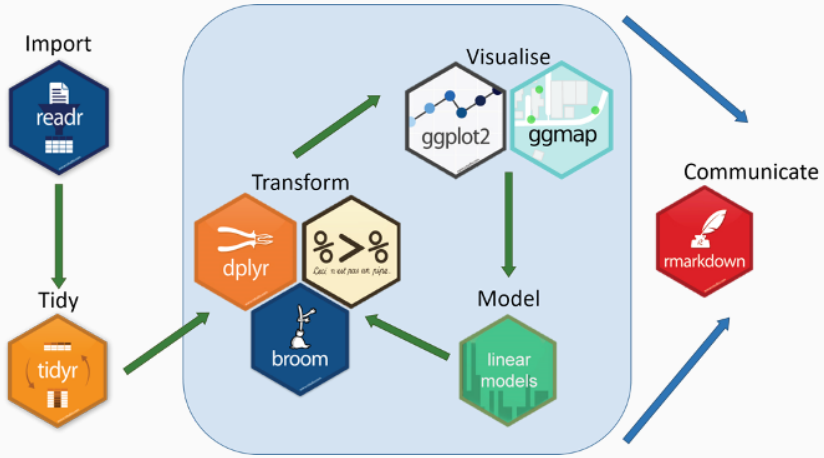
Overview

- Base R vs. Tidyverse
- Tidy data
  - How Tidyverse helps
- Exercises using the Tidyverse - RNA-seq data exploration

General idea of Tidyverse

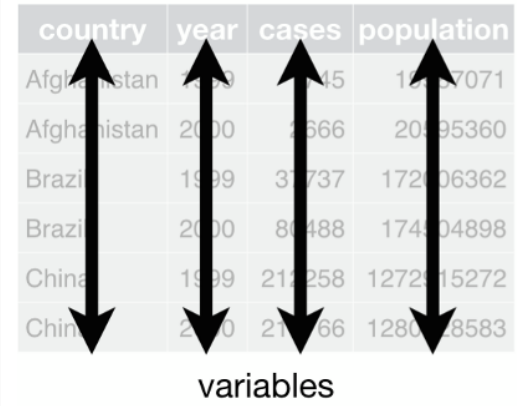
- Collection of R packages for data science that reimplement basic R functionality
  - Developed by Hadley Wickham
  - Provide functionality to model, transform and visualise data
  - Human readable code

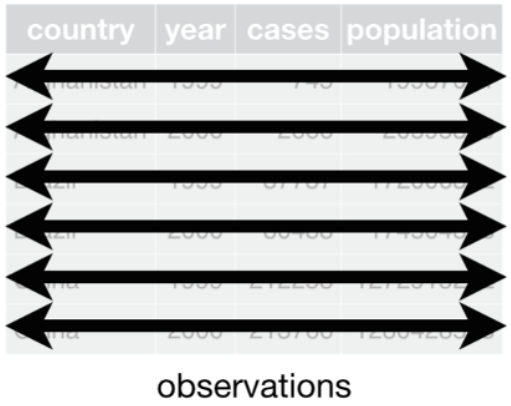
- RStudio cheat sheets available:
  - https://rstudio.com/resources/cheatsheets/
  - Nice graphical clarifications

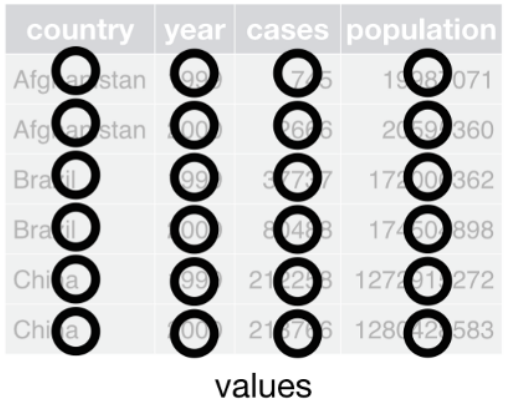


Tidy data

Idea is to structure data to ease analysis and plotting







Tidyverse key features

UNIVERSITY OF OXFORD

- Use `%>%` (piping operator/magrittr) to combine functions
- All packages share an "underlying design philosophy, grammar, and data structures" e.g.
  - Data is the first argument in most functions
  - Use unquoted variable names
  - Use of tibble data structure

Tidyverse packages: tibble

UNIVERSITY OF OXFORD

- tibble (tbl_df) is a type of data.frame in R
- Do less e.g. do not change variable names or types
- Complain more e.g. when variable does not exist
- Enhanced `print()` method:
  - Data type shown with column names
  - Easier to use with large datasets containing complex objects

Tidyverse packages: dplyr

Functions:

- `select()` # select columns
- `filter()` # filter rows based on condition
- `group_by()` # group observations together (original dataset does not change, just the way it is represented)
- `summarise()` # summarise the variables of an existing tbl e.g. mean, sum
- `arrange()` # order observations/rows based on values in given column
- `join()` # join tbls together (left, right, full, inner)
- `mutate()` # create new column

Tidyverse packages: tidyr

Functions:

- `gather()` # "gathers" multiple columns into key-value pairs
- `pivot_longer()` # updated version of `gather()`
- `spread()` # opposite of gather, "spreads" key-value pairs into multiple columns
- `pivot_wider()` # updated version of `spread()`
- `separate()` # split one column into multiple columns
- `unite()` # opposite of `separate()`, join multiple columns into one

Tidyverse packages: readr

Reading data into R and writing to files

Faster than standard R importing and writing methods

Functions:

- `read_delim()` # read in file with specified delimiter
- `read_csv()` / `read_tsv()` # special cases of `read_delim()` for reading in CSV and TSV files
- `write_delim()` # write data to file with specified delimiter
- `write_csv()` / `write_tsv()` # special cases of `write_delim()` for writing data to CSV and TSV files

Tidyverse packages: stringr

UNIVERSITY OF OXFORD

Working with string variables

Functions:

- `str_sub()` # extract substrings from character vector
- `str_trim()` # trim whitespaces
- `str_c()` # combine strings
- `str_length()` # determine length of string
- `str_to_lower()` # convert string to lowercase
- `str_to_upper()` # convert string to uppercase

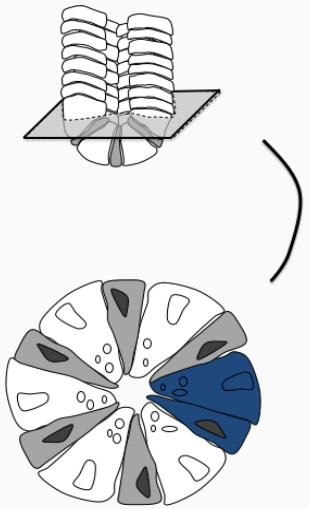
Other Tidyverse packages

UNIVERSITY OF OXFORD

- Importing data - `readxl`, `haven`, `googledrive`
- Data wrangling - `lubridate`, `hms`
- Plotting - `ggplot2`
- Working with categorical variables or factors - `forcats`
- Functional programming - `purrr`

(Untidy) data: labelled cell counts in crypts

- Crypt RFP cell counts
- Induce RFP day 0
  - Just one cell per crypt
- Count number of labelled cells and total cells per crypt
  - Two time points - day 10 and day 21 post labelling
- Over time
  - Expect fewer small clones and more large ones
  - We will explore the data to see if this is true



Dataset files

- File 1
  - Two columns per mouse - RFP+ cells, total cells
- File 2
  - Mouse ID
  - Sex of mouse
  - Day post labelling for measurement
- Rows don't represent an individual observation - untidy!!

What we plan to do

- Tidy data!

Mouse_ID	Sex	Crypt_number	Day_post_label	RFPpos	Total_cells
108421-2	M	1	10	6	11
108422-6	F	1	10	1	13

- General data exploration
  - Total cells per crypt - split by mouse, day, day + sex
  - RFP+ cells per crypt - number and proportion
    - Same for both time points?
  - How many crypts measured per mouse?
  - How many total crypts measured each day?
- Visualise whether clone sizes grow

Demonstrate Tidyverse functions with script

Summary

- We've covered the most used Tidyverse functions:
  - `gather()`, `spread()`
  - `summarise()`
  - `mutate()`
  - `group_by()`
  - `select()`
  - `filter()`
- There is more stuff:
  - `purrr`
    - Functional programming
    - Working with lists
  - Nesting within tibbles

Exercises

1. Run the example code that I have been showing you and make sure that you understand what the different commands are doing

2. Explore bulk RNA-seq data:
    - Read data and metadata in
    - Tidy and annotate
    - Explore
        - Summarise
        - Plot

Data

- RNA-seq (European Nucleotide Archive PRJEB18572)
- Mouse CD4+ and CD8+ T cells extracted from GFP-Egr2 knockin (Egr2 Kin) and $Egr2^{loxP/loxP}$ hCD2-Cre $Egr3^{-/-}$ (Egr2/3 DKO) mice, 7 days after infection with vaccinia virus
- 3 biological replicates per group (12 samples total)
- Two files:
  - obds_countstable.tsv.gz
  - obds_sampletable.tsv

Read and tidy data

- Tidy count file
  - Three columns - Geneid, sample, count
- Join with gene info to get mgi_symbol
  - Use the `biomaRt` package
- Tidy metadata file
  - One variable per column
  - Don't need species and library_layout columns
- Add metadata to table with counts and gene info
- Calculate counts per million (CPM) - use `group_by()` and `mutate()`
- Also calculate log2(CPM + 0.25)

Plot read depth per sample

UNIVERSITY OF OXFORD

- Use `group_by()` and `summarise()`
- Plot with ggplot using `geom_bar()`
- Edit the appearance of the plot to make it easier to read/"prettier"
- Does any sample jump out?

Lowly expressed genes

- How many genes have no counts for any sample?
- Draw a density plot of log2 CPM for all genes
  - Use `geom_density()` and colour by sample
  - Are the samples similar?
- Filter out genes that have low expression in 3 or fewer samples
  - For low expression use CPM < 0.5
  - What proportion of genes are lowly expressed?
- Make a density plot of log2 CPM with the filtered data

Biological exploration of the data

UNIVERSITY OF OXFORD

- Plot CD4 and CD8 expression for all samples - does it make sense?
  - Colour by replicate and facet by genotype against cell type
- Generate the same plot for Egr2 and Egr3 for all samples - does it make sense?
- Choose 8 biologically relevant genes and plot a heatmap using the pheatmap package