Data Wrangling with R: Using the Tidyverse

Alexandra Emmons, PhD

Bioinformatics Training and Education Program (BTEP)

8 lessons directed toward data wrangling

- L1: Introduction to R, RStudio, and the Tidyverse
- L2: Getting Started, the basics
- L3: Loading and reshaping data
- L4: Data visualization with ggplot2
- L5: The pipe, filtering, and joining data tables
- L6: Split, apply, combine
- L7: Introduction to Bioconductor -omics classes (containers)
- L8: Data Wrangling Review and Practice

No Coding
Coding

What is R?

- A language and statistical computing environment
- Open source, for and by scientists
- Widespread community
- Extended use through package installation
  - R Packages are collections of R functions, compiled code and sample data

Why should we use R?

- Great for statistical analysis, data visualization, and report generation
- Supports large scale data analysis
- Removes some of the human error associated with excel
- Ever growing community
  - Many ways to get help
  - Field specific packages and workflows
  - Problems are “googlable”

Where do we find R packages?

Comprehensive R Archive Network

Github

Bioconductor

Check out METACRAN

What is R Studio?

An integrated development environment (IDE) for R

Includes a console, code editor, and tools for plotting, history, debugging, and workspace management.

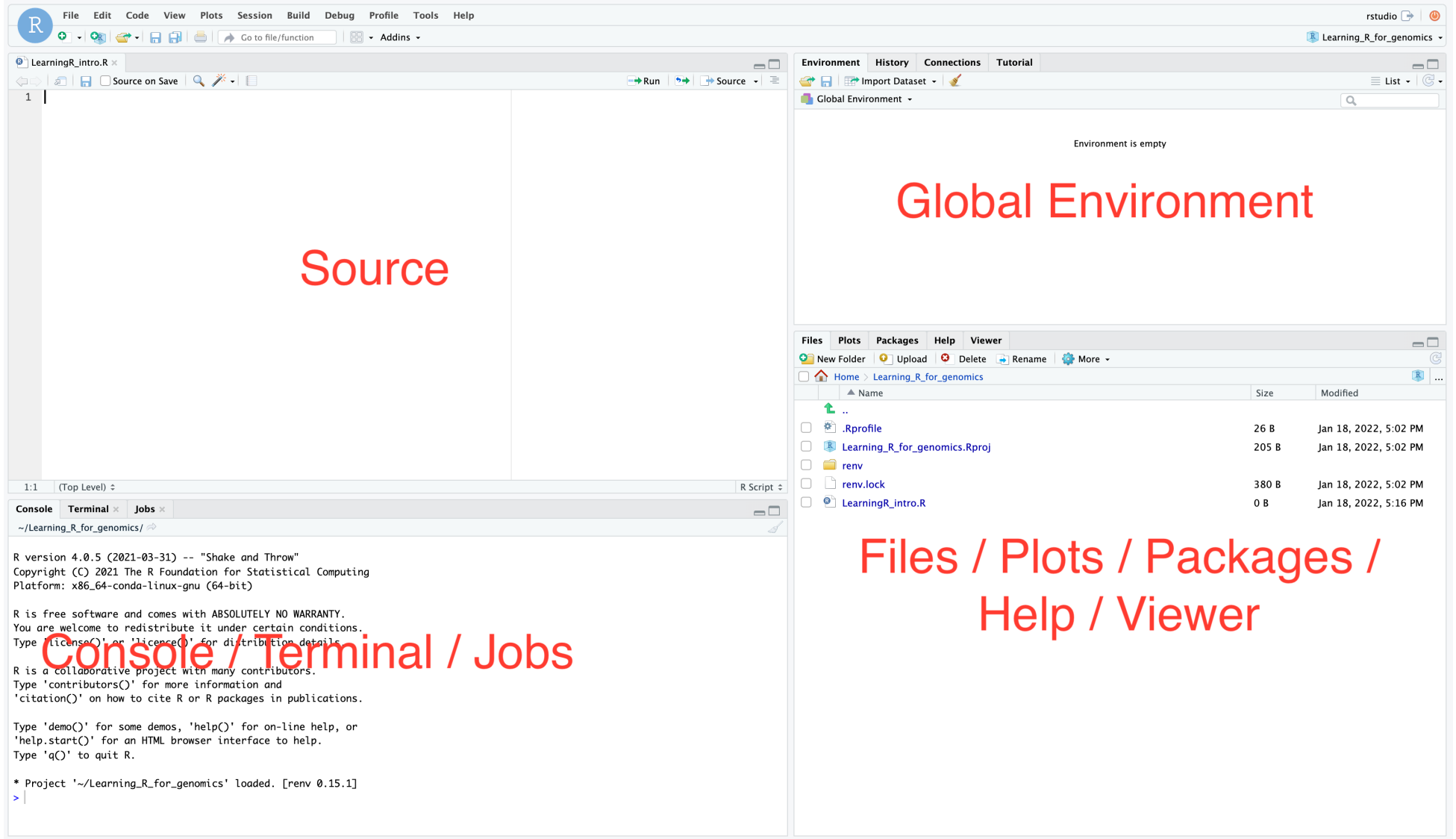
Open-source and can be installed locally or used through a browser (RStudio Server, Posit Cloud)

DNAnexus

- A Cloud-based platform for NextGen Sequence analysis for which CCR has a "*site-license*"
- We will be using this platform to provide a uniform, stable, preinstalled interface for R training.
  - Uses RStudio server
  - Integrates course-notes
  - R packages installed and ready to use
  - The data ready to use and in one place; no need to download

Course registrants, please fill out this form with your DNAnexus information.

Let's take a tour of Rstudio IDE



Data Wrangling

Best Practices for data analysis

1. Keep raw data separate from analyzed data.
2. Keep spreadsheet data Tidy (or as tidy as possible)
3. Trust but Verify

--- From https://datacarpentry.org/genomics-r-intro/03-basics-factors-dataframes/index.html

What is tidy data?

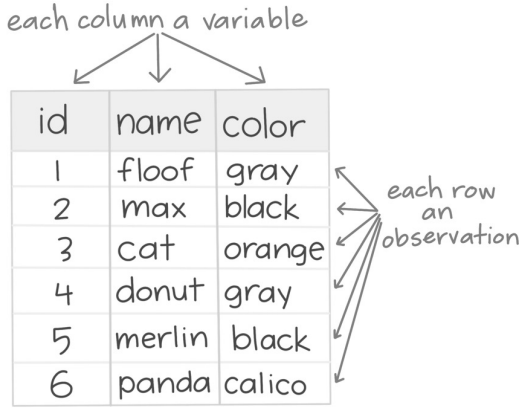
****Having tidy data is useful but not always necessary. Do not worry about strict adherence to the rules. Your data should be in whatever format that makes your life easier for analysis.****

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

Image from Lowndes and Horst 2020: Tidy Data for Efficiency, Reproducibility, and Collaboration

Guidelines to keep spreadsheets tidy

- Be consistent
- Choose meaningful names for things; no spaces
- Write dates as YYYY-MM-DD
- No empty cells
- Put just one thing in a cell
- Don’t use font color or highlighting as data
- Save the data as plain text files

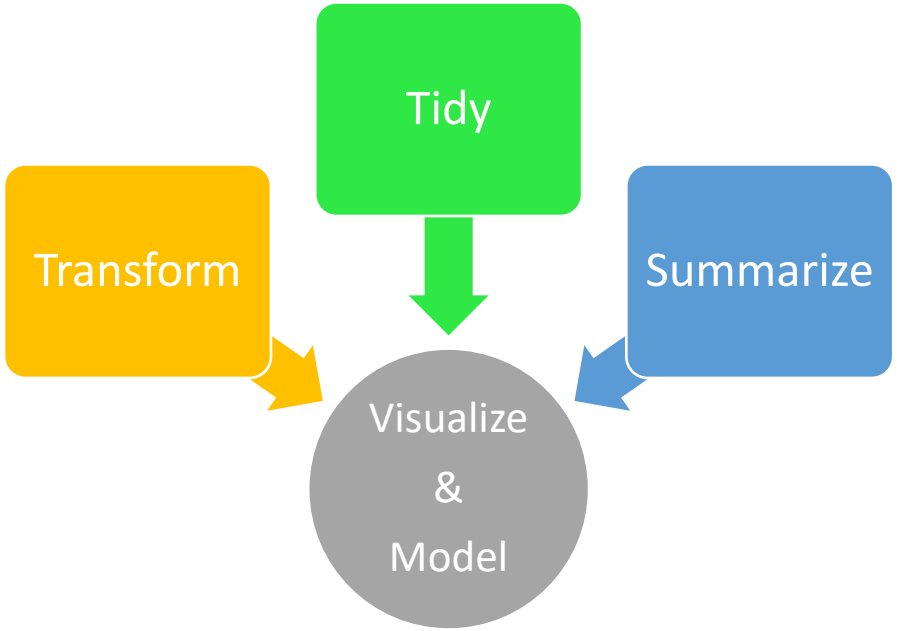
--- https://jhudatascience.org/tidyversecourse/intro.html

Tidyverse

- An opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures. --- tidyverse.org
- Core packages:
  - dplyr, ggplot2, forcats, tibble, readr, stringr, tidyr, and purr

What is data wrangling?

- Data wrangling is a catch all phrase for cleaning, transforming, and summarizing data
- The primary packages we will focus on for this purpose are tidyr and dplyr.



Getting Help

Stack Overflow and other forums

- Public Q&A platform
- Ex: https://stackoverflow.com/questions/65095565/make-some-sample-names-unique-according-to-conditions

Tutorials

Vignettes

- Ex, browseVignettes(package="dplyr")

Coursera

- JHU Tidyverse Skills for Data Science in R Specialization
- Introduction to the Tidyverse
- Importing Data in the Tidyverse
- Wrangling Data in the Tidyverse
- Visualizing Data in the Tidyverse
- Modeling Data in the Tidyverse

Dataquest

- Intro to data analysis in R
- Data Visualization in R
- Data Cleaning in R

Bioconductor tutorials / workflows

Many others…

- glittr

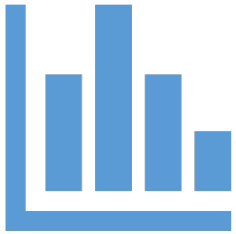
For a Coursera or Dataquest license go to https://bioinformatics.ccr.cancer.gov/btep/self-learning/

Course Materials

Materials for each lesson will be found at https://btep.ccr.cancer.gov/docs/data-wrangle-with-r-2023/.

Course materials will be updated prior to each lesson.

BTEP



Other R courses

R Introductory Series

Data Visualization with R

Email us

ncibtep@nih.gov

BTEP Coding Club

- Once a month
- Tailored bioinformatics training to the NCI community.
- 1-hour demo / tutorial of a bioinformatics tool, software, skill, or platform.
- Ranges in experience level from beginner to advanced.
- **Email us at ncibtep@nih.gov if there is a specific topic you would like to see featured.**

Check out past events here:
https://bioinformatics.ccr.cancer.gov/docs/btep-coding-club/

Helpful things to know before getting started

Terms to Know

- Function - code written to perform a specific task
  - Example: Getwd()
- String – a sequence of one or more characters
  - Enclosed by parentheses
- Data frame – object that stores tabular data; all variables are of the same length
- Directory – location where files are stored
- Working directory – your current directory
- Package – the fundamental unit of shareable code, bundling together code, data, documentation, and tests. This is how we extend the use of R.
- Library – a directory of installed packages
  - Example: library(dplyr)

Directory Structures

- A file path shows us the location of a file. These are nested structures.
- .libPaths()
  - Will show us the location of installed R packages
  - For example:
    - [1] "/Library/Frameworks/R.framework/Versions/4.1/Resources/library"
- Absolute file path
  - The complete file path
- Relative file path
  - A shortcut path from some other directory

In summary

Today we…

Learned about advantages of R and RStudio

Navigated the RStudio environment

Learned about concepts related to data wrangling

Reviewed resources available for getting help

Next time…

Get ready for some coding fun in RStudio (DNAnexus)

Learn R basics