Reading, navigating, and plotting on your own

Thoughts and challenges, troubleshooting, and working demo

"Maybe I should have stuck with Excel."

Ed Himelblau

How do you start exploring a new dataset

- How should the data be loaded?

- What type of data are we looking at?
  - What is the size of the dataset?
  - What are the variables?
  - Is the data complete? Why is data missing?

- What are the question(s) I'm trying to address?

How should the data be loaded?

1. Look at the file extension → Do you recognize it?

   → No? Google “what is a **txt** file extension”

   Other extensions: csv, tsv, xls, Rda, Rds, h5, mtx

How should the data be loaded?

1. Look at the file extension → Do you recognize it?

   → No? Google “what is a **txt** file extension”

   Other extensions: csv, tsv, xls, Rda, Rds, h5, mtx

1. Is a package required to look at this data?
   - Large **txt** or **tsv** files: **data.table** package → **fread**()
   - Excel files: **readxl** package → **read_excel**()
   - Google → What package can be used to read an **Rda** file in R? *Is the package installed?*

Reading in your data

```
library(data.table)
variant_file <- "/cloud/project/data/single_cell_rna/cancer_cell_id/mcb6c-exome-somatic.variants.annotated.clean.filtered.tsv"
print(variant_file)
read_variants <- fread(variant_file)
```

Make sure you include the full path of your file.

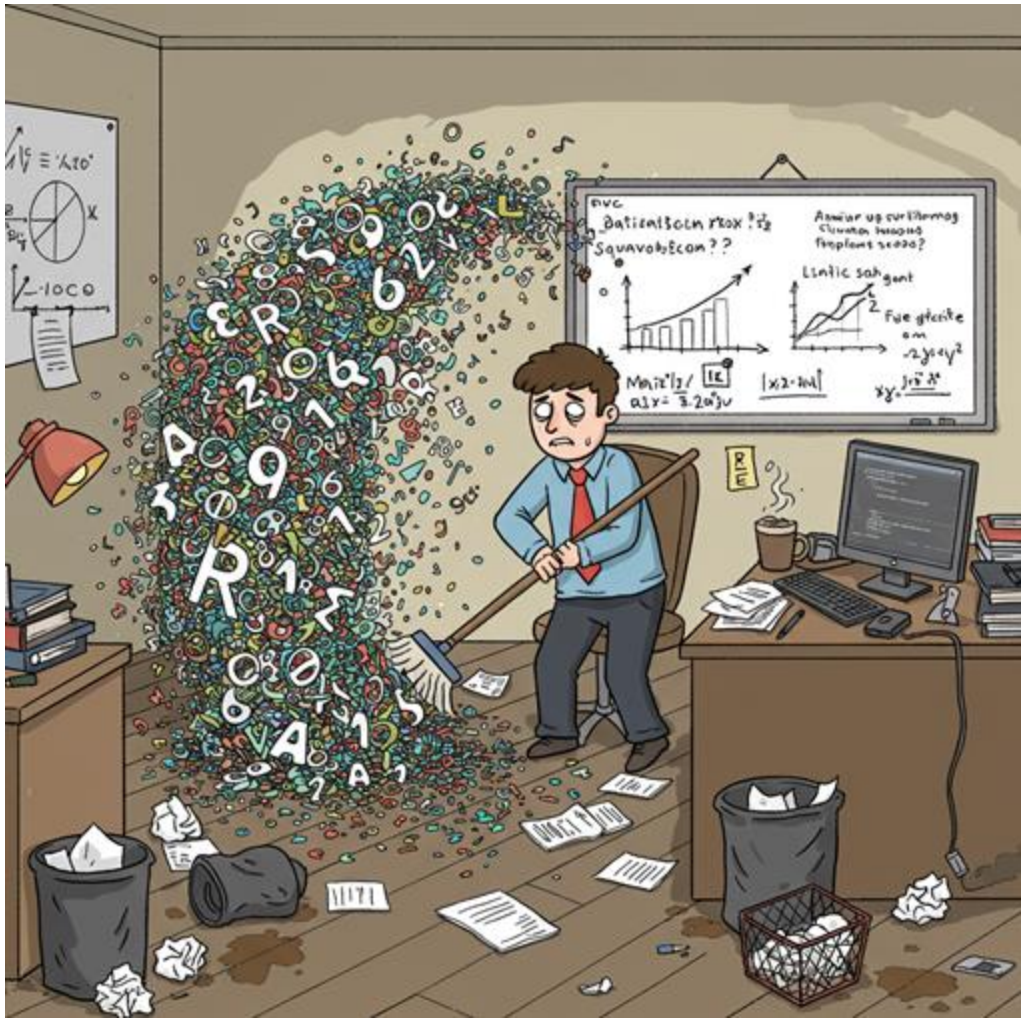
```
head(variant_file)
summary(variant_file)

nrow(variant_file)
ncol(variant_file)
```

Does the data look the way you expect it to?

Is your data ready for analysis?

Is your data ready for analysis?

… ***Probably not***

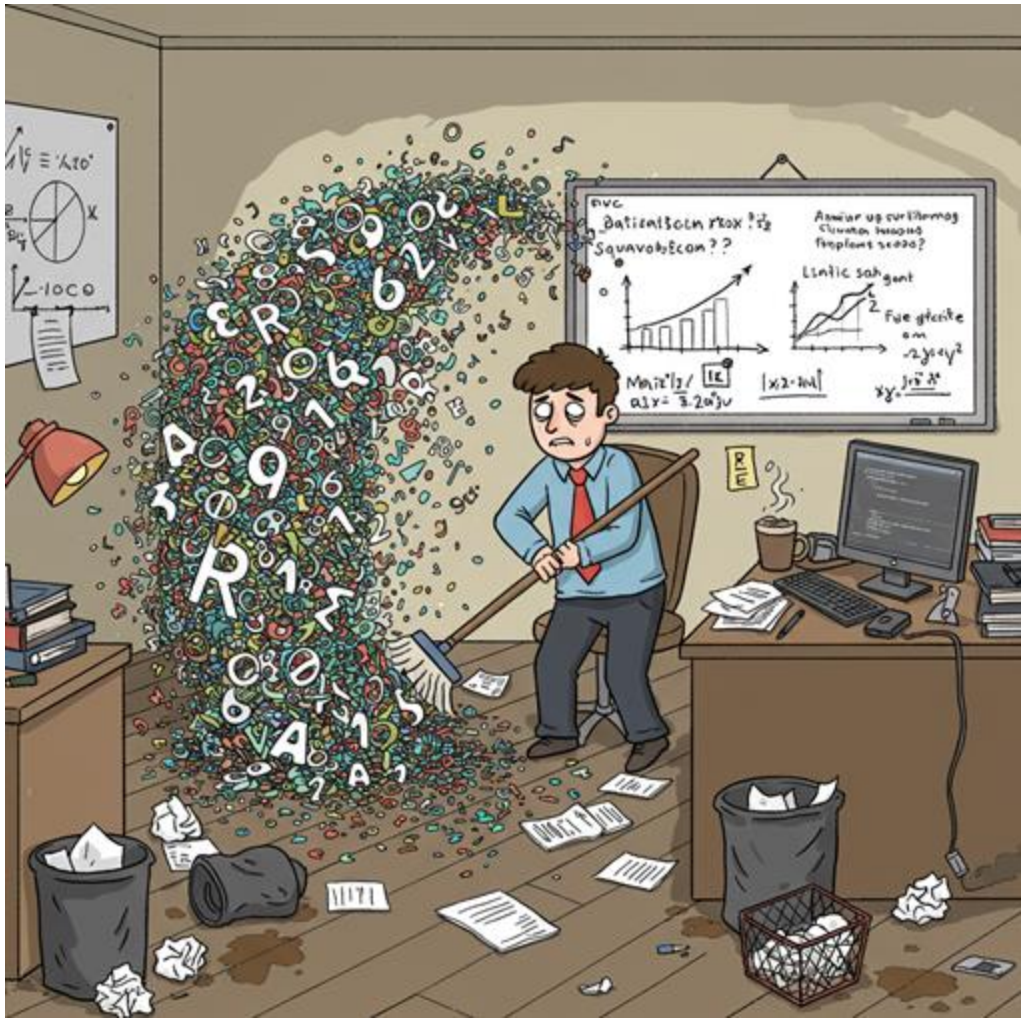


Is your data ready for analysis?

Data cleaning refers to the process of processing datasets to improve their quality for analysis and decision-making.

This may include:

- Removing or correcting missing information
- Identifying inconsistencies
- Performing batch effect correction



Missing data vs. sparse data

Why might data be missing?

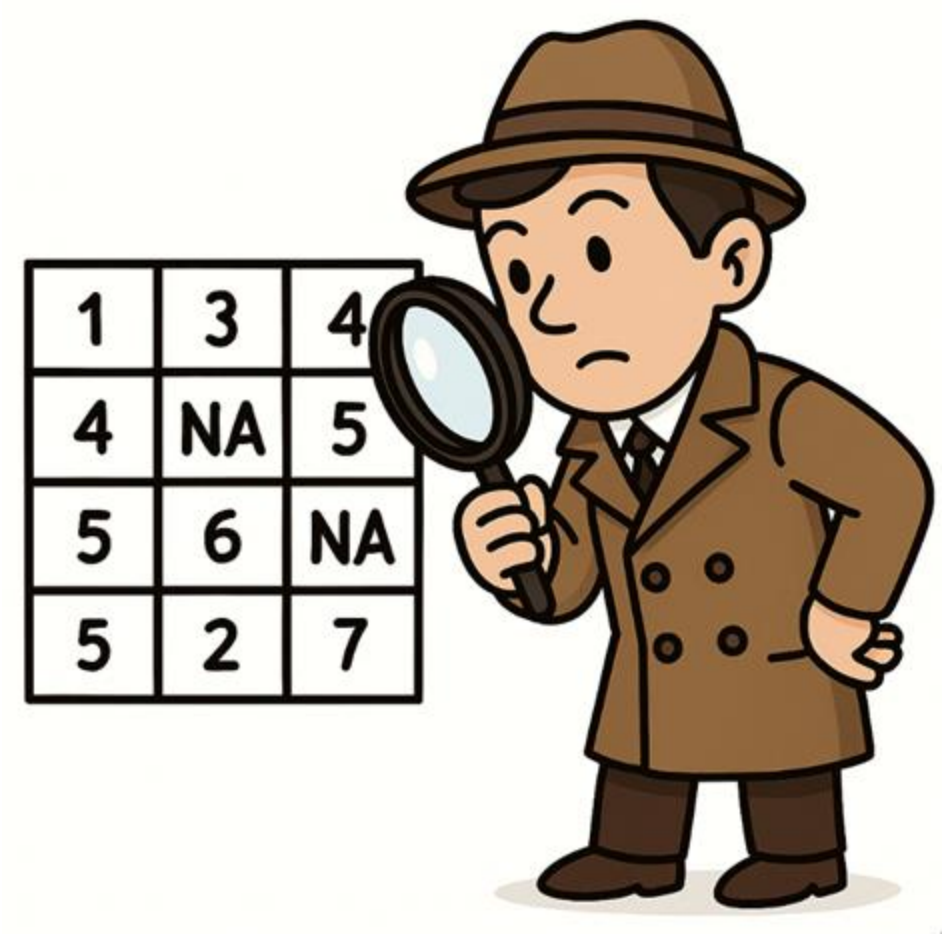
- The file doesn't contain values for a given column in a subset of rows
- Computation/mathematical changes resulted in really large or really small numbers that can't be represented in memory
- You tried to add a column incorrectly
- Incomplete data file
- Technical reasons

Sparse data

- Should there be a lot of zero values?

Missing Data

- **NA** (“Not Available” / “Not Applicable”)
  - Generic missing-value marker in any vector type
- **NaN** (“Not a Number”)
  - Result of undefined numeric operations (e.g. 0/0)
- **NULL**
  - Absence of a value or empty object (length 0)
- **Inf / –Inf**
  - Infinite results (e.g. 1/0, –1/0)
- **Character NA**
  - Written as NA_character_ or as.character(NA) in character vectors

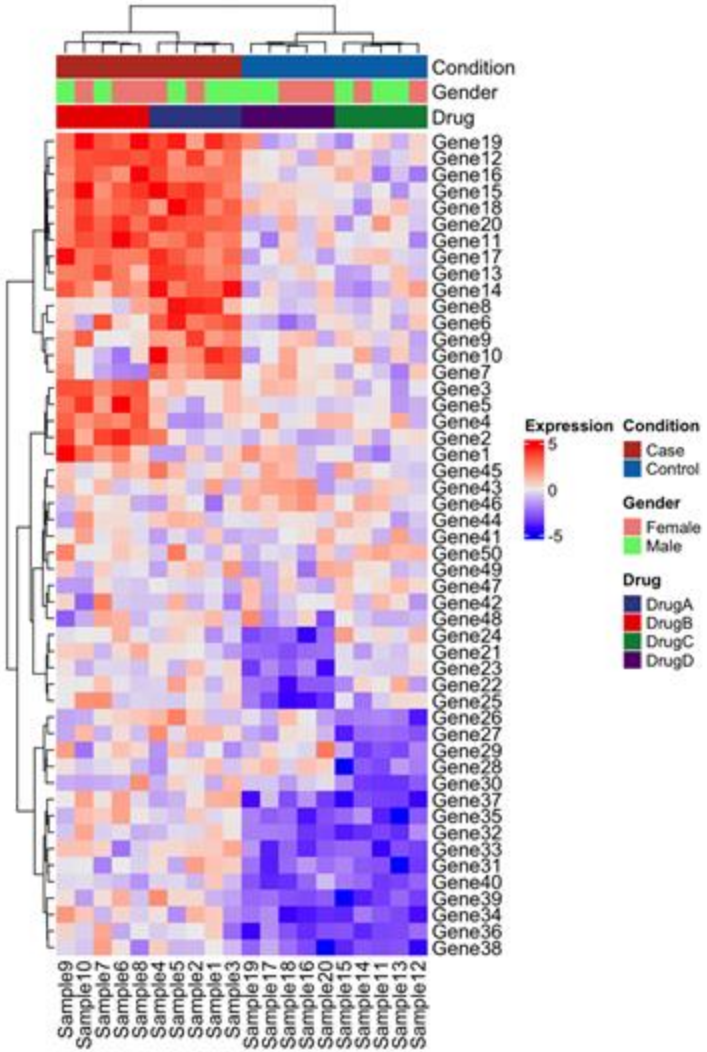


Missing Data

- **Detection functions**
  - is.na(x) → catches both NA and NaN
  - is.nan(x) → isolates NaN only
  - is.null(x) → tests for NULL objects
  - is.infinite(x) → flags Inf/-Inf
- **Common handling strategies**
  - na.omit() / na.exclude() → drop rows with any NA
  - na.rm = TRUE → ignore NA in many base-R functions
  - tidyr::replace_na() or dplyr::coalesce() → fill or impute missing values

Live demo

Replicating analysis by working backwards



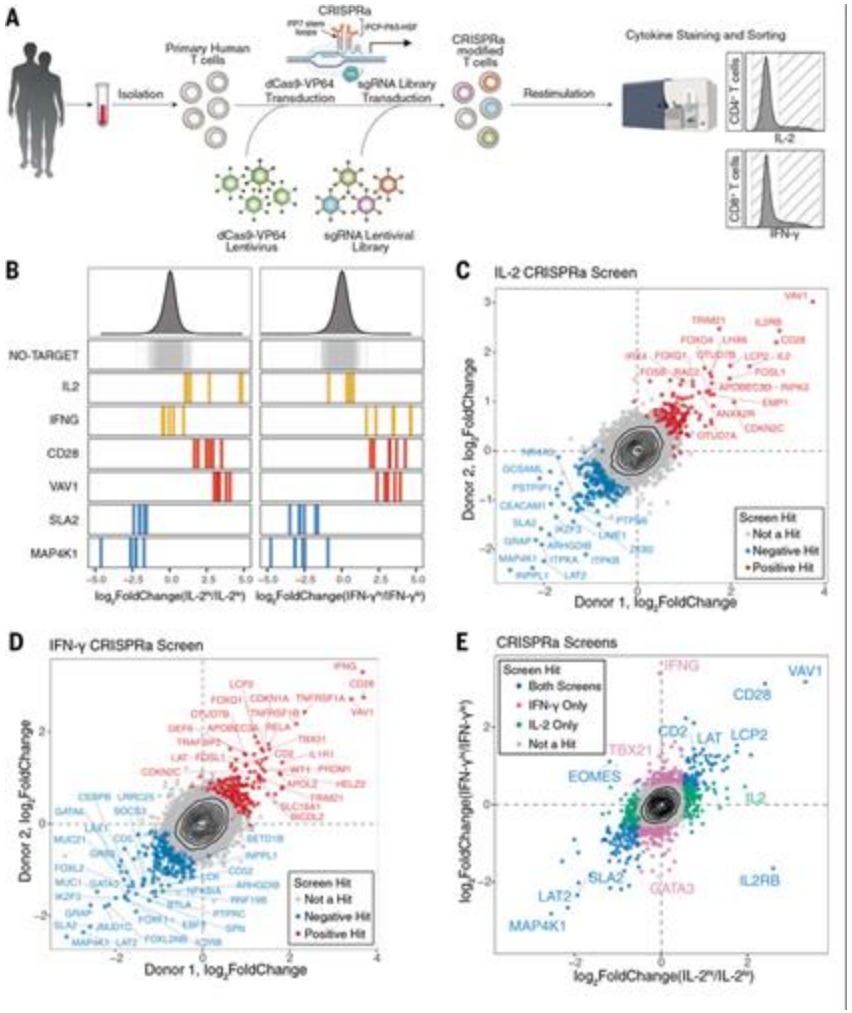
Hands-on: Reading, navigating, and plotting data

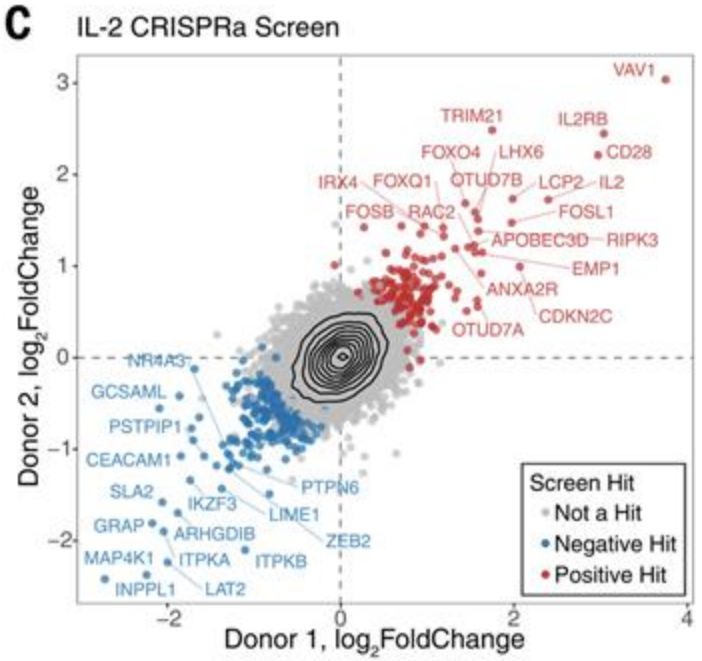
1. Start your first Rscript
2. Work through the blocks of code under the **Course: Reading, navigating, and plotting data** page
   - Review the details on how the code works in the Lecture slides for assistance
   - Put a post-it on your laptop if you get stuck, indicating for a TA to come up to you
   - Work through the blocks of code on this page, practicing in both your Rscript and the console
3. Take the next step
   - There are a list of **Additional exercises** at the bottom of the page for you to try on your own

Goal: Have the foundation to take on your own dataset

Can you load one of your own datasets into R?

Live demo (if time)





Schmidt, et al. Science 2022