VIDO BIOINFO Bioinformatics and Systems Biology Lab

Canadian Bioinformatics Hub

Canadian Bioinformatics Workshops

Introduction to R Programming for Bioinformatics

Day 1- Module 1B: Exploring your data in R

Mohamed Helmy, PhD
Principal Scientist and Adjunct Professor
Bioinformatics and Systems Biology Lab
VIDO, University of Saskatchewan

6-7 October 2025, VIDO, Saskatoon

VIDO

University of Saskatchewan

Why Explore Data?

- Understand structure and types of variables
- Detect missing or incorrect values early
- Check for outliers and unusual patterns
- Guide decisions about cleaning and analysis
- Prevent errors in downstream analysis

Key Tools in R for Data Exploration

- **Preview your data**
  - head() → first few rows
  - tail() → last few rows

- **Inspect structure**
  - str() → data type & structure of each column

- **Quick summary**
  - summary() → basic statistics for each variable

- **Check dimensions**
  - dim(), nrow(), ncol()

- **Frequency counts**
  - table() → frequency counts of categorical variables

```
> summary(df)
      age             name               group
 Min.   :25.00   Length:2           Length:2
 1st Qu.:26.25   Class :character   Class :character
 Median :27.50   Mode  :character   Mode  :character
 Mean   :27.50
 3rd Qu.:28.75
 Max.   :30.00
```

Demo: Basic Data Inspections

- **Inspecting real life data of 128 patients of T- and B-cell Acute Lymphocytic Leukemia (ALL)**

RStudio

Go to file/function | Addins

Basic.R | ALL | Main_alignment_app.R | df

Filter

	X	Code	Diagnosis	Sex	Age	BT	Remission	CR	Cyto.normal	Relapse
1	1	1005	5/21/1997	M	53	B2	CR	CR	FALSE	FALSE
2	2	1010	3/29/2000	M	19	B2	CR	CR	FALSE	TRUE
3	3	3002	6/24/1998	F	52	B4	CR	CR	NA	TRUE
4	4	4006	7/17/1997	M	38	B1	CR	CR	FALSE	TRUE
5	5	4007	7/22/1997	M	57	B2	CR	CR	FALSE	TRUE
6	6	4008	7/30/1997	M	17	B1	CR	CR	FALSE	TRUE
7	7	4010	10/30/1997	F	18	B1	CR	CR	FALSE	TRUE
8	8	4016	2/10/2000	M	16	B1	CR	CR	FALSE	TRUE
9	9	6002	3/19/1997	M	15	B2	CR	CR	TRUE	TRUE
10	10	8001	1/15/1997	M	40	B2	CR	CR	FALSE	TRUE
11	11	8011	8/21/1998	M	33	B3	CR	CR	FALSE	FALSE
12	12	8012	10/22/1998	M	55	B3	CR	CR	FALSE	TRUE
13	13	8018	8/27/1999	M	5	B3	CR	CR	NA	TRUE
14	14	8024	7/20/2000	M	18	B2	CR	DEATH IN CR	FALSE	NA
15	15	9008	12/17/1999	M	41	B3	CR	CR	FALSE	FALSE
16	16	9017	2/3/2000	F	27	B	CR	CR	TRUE	TRUE
17	17	11005	6/1/1998	M	27	B2	CR	DEATH IN CR	FALSE	FALSE
18	18	12006	2/20/1997	M	46	B3	REF	REF	FALSE	NA
19	19	12007	3/11/1997	M	37	B2	CR	CR	FALSE	TRUE
20	20	12012	5/21/1997	F	36	B3	REF	REF	FALSE	NA
21	21	12019	9/4/1997	M	53	B2	CR	CR	TRUE	FALSE

Showing 1 to 22 of 128 entries, 10 total columns

Demo: Basic Data Inspections

- **The ALL dataset:**
  - Microarray gene expression data from acute lymphoblastic leukemia (ALL) patients.

- **Metadata includes patient-level information:**
  - Code: patient ID
  - Diagnosis: date of diagnosis
  - Sex, Age: demographics
  - BT: B-cell tumor subtype (e.g., B1, B2, B3, B4)
  - Remission & CR: clinical outcomes (Complete Remission, Death in CR, etc.)
  - Cyto.normal: cytogenetic normality (TRUE/FALSE)
  - Relapse: indicator of relapse

- **Key Learning Point:**
  - Understanding the structure of your dataset is the first step before any analysis.

Summarizing with Statistics

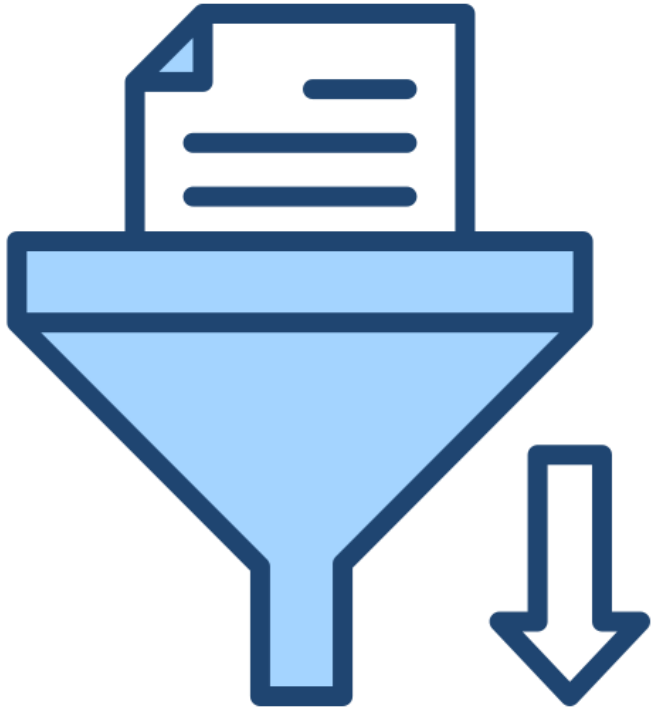
- **Central tendency:**
  - mean()
  - median()
- **Variability:**
  - sd() standard deviation
  - var() variance
- **Extremes:**
  - min()
  - max()
- **Frequency counts:**
  - table() for categorical variables
  - hist() for histogram creation
  - summary () Quick summary

```r
# View patient metadata
data("ALL")
df2 <- pData(ALL)

# Quick summary
summary(pData(ALL)[, c("age", "sex", "BT", "relapse")])

# mean and median age
mn <- mean(df2$age) # this will return NA
md <- median(df2$age) # this will return NA

mn <- mean(df2$age, na.rm = TRUE) # this will work
md <- median(df2$age, na.rm = TRUE) # this will work

# standard deviation and variance
std <- sd(df2$age, na.rm = TRUE)
vr <- var(df2$age, na.rm = TRUE)

# Extremes
mxx <- max(df2$age, na.rm = T)
mnn <- min(df2$age, na.rm = T)

# Table and (Frequency)
age_dit <- table(df2$age)

# Quick summary
summary(df2[, c("age", "sex", "BT", "relapse")])
```

Data Filtering

- **Filtering or Subsetting:**
  - Let us focus on specific patients or conditions of interest
  - Help reduce complexity
  - Prepare data for analysis or visualization

- **Common use cases (in the ALL example):**
  - Select patients above a certain age
  - Select patients by tumor subtype (BT)
  - Identify patients who relapsed



Data Filtering (Combining Conditions)

We can apply multiple filters at once

- **Real data is messy and multi-factorial.**
  - In bioinformatics, we rarely filter by just one variable. For example:
    - You might want patients older than 40 AND with relapse.
    - Or find females OR patients with subtype B2.
- **Supports hypothesis-driven exploration.**
  - Scientists often ask multi-dimensional questions like:
    - Do older male patients relapse more often?
    - Are remission rates different in B2 vs B3 subtypes, but only in females?
- **Use logical operators:**
  - & = AND
  - | = OR
  - ! = NOT

Demo: Data Filtering (Combining Conditions)

- **Subsetting**
  - subset()
- **Indexing with [**
  - df[df$Age > 40,]
  - df[df$Age > 40 & df$Relapse,]
- **Assignment with condition**
  - df <- df[df$Sex == "F",]
- **cleaner syntax: with()**
  - with(df, df[Age > 40 & Relapse == TRUE,])
  - df[which(df$Age > 40),]
- **Matching values**
  - match() / %in%
  - df[df$BT %in% c("B2", "B3"),]
- **Logical indexing directly**
  - df[df$Relapse == TRUE | df$Sex == "F",]

```r
# Subsetting and Filtering
# subset()
subset(df2, Age > 40 & Relapse == TRUE)

# Indexing with [
df2[df2$Age > 40, ]                 # filter rows
df2[df2$Age > 40 & df2$Relapse, ]   # multiple conditions (same as df$Relapse == T)
df2[, c("Age", "BT")]              # select columns
df <- df2[df$Sex == "F", ] # female patients only (#Assignment with condition)

# with()
with(df, df[Age > 40 & Relapse == TRUE, ]) #for cleaner syntax
df[which(df$Age > 40), ] # more cleanr syntax

#match() / %in% (matching values)
  df[df$BT %in% c("B2", "B3"), ]

# Logical indexing directly
  df[df$Relapse == TRUE | df$Sex == "F", ]
```

Hands-on Practice: Subsetting the ALL Metadata

- **Task 1 – Basic Filtering**
  - Extract all patients who are younger than 20 years
- **Task 2 – Single Condition + Column Selection**
  - Get only Age and Sex for patients with BT = "B2"
- **Task 3 – Combining Conditions**
  - Find male patients older than 40
  - Get patients who are female OR had a relapse
- **Mini-Challenge**
  - Find all patients who are male, had relapse, and whose age is greater than 30. How many are there?

THANK YOU

VIDO

VACCINE AND INFECTIOUS DISEASE ORGANIZATION

VIDO.ORG

UNIVERSITY OF SASKATCHEWAN