Introduction to R 2025

Faculty: Mohamed Helmy

October 6-7, 2025

Contents

Part I

Introduction

Chapter 1

Workshop Info

Welcome to the 2025 Introduction to R Canadian Bioinformatics Workshop webpage!

1.1 Pre-work

You can find your pre-work here.

1.2 Class Photo

1.3 Schedule

Chapter 2

Meet Your Faculty

2.0.0.1 Mohamed Helmy

Principal Scientist and Adjunct Professor Vaccine and Infectious Disease Organization (VIDO), University of Saskatchewan Saskatoon, Saskatchewan, Canada

mohamed.helmy@usask.ca

Mohamed is a Computational Systems Biologist and Principal Scientist leading the Bioinformatics and Systems Biology Lab (BSBL) at the Vaccine and Infectious Disease Organization (VIDO), University of Saskatchewan. He received his MSc and PhD in Computational Systems Biology from Keio University (Tokyo, Japan) and completed his postdoctoral training in bioinformatics at Kyoto University and the University of Toronto. Mohamed's interdisciplinary research profile bridges biology, computer science, and public health.

2.0.0.2 Sylvia Li

Graduate student Vaccine and Infectious Disease Organization (VIDO), University of Saskatchewan Saskatoon, Saskatchewan, Canada

Sylvia is a Computer science MSc student at the University of Saskatchewan, supervised by Dr. Helmy. She holds dual BSc degrees in Bioinformatics and Computer science. Currently her work focuses on bacterial genomic data.

Data and Compute Setup

2.0.0.3 Course data downloads

Coming soon!

2.0.0.4 Compute setup

Coming soon!

Part II

Modules

Chapter 3

Module 1

3.1 Lecture

3.1.1 1A

3.1.2 1B

3.2 Lab 1A

3.2.1 Variables

Create 2 numeric variables and assign values for each

```
x = 10
y = 6
```

Calculate the sum of them

```
total = x + y
total
```

```
## [1] 16
```

Calculate the square root of the total

```
sr = sqrt(total)
sr
```

```
## [1] 4
```

3.2.2 Data Structures

Vector

```
v <- c(1,2,3,4)
v
```

```
## [1] 1 2 3 4
```

Matrix

```
m <- matrix(1:6, nrow = 2)
m
```

```
##      [,1] [,2] [,3]
## [1,]    1    3    5
## [2,]    2    4    6
```

Dataframe

```
df <- data.frame(age=c(25,30), name=c("Mo","Tom"), group=c("A", "B"))
df
```

```
##   age name group
## 1  25   Mo     A
## 2  30  Tom     B
```

List

```
lst <- list(numbers=v, info=df)
lst
```

```
## $numbers
## [1] 1 2 3 4
##
## $info
##   age name group
## 1  25   Mo     A
## 2  30  Tom     B
```

3.3 Lab 1B

3.3.1 Install BioconductoR packages

```
install.packages("BiocManager")
BiocManager::install("ALL")
```

3.3.2 View patient metadata

```
library(BiocManager)
library(ALL)
data(ALL)
df2 <- pData(ALL)
```

3.3.3 Quick summary

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

```
##       age          sex           BT      relapse
##  Min.   : 5.00   F   :42   B2     :36   Mode :logical
##  1st Qu.:19.00   M   :83   B3     :23   FALSE:35
##  Median :29.00   NA's: 3   B1     :19   TRUE :65
##  Mean   :32.37             T2     :15   NA's :28
##  3rd Qu.:45.50             B4     :12
##  Max.   :58.00             T3     :10
##  NA's   :5                 (Other):13
```

3.3.4 str() and dim() functions

```
dim(df2)
```

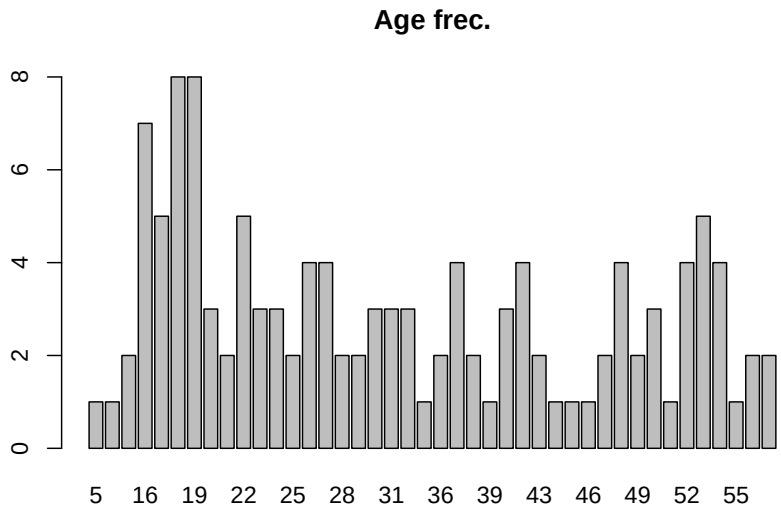
```
## [1] 128  21
```

```
str(df2)
```

```
## 'data.frame':    128 obs. of  21 variables:
##  $ cod           : chr  "1005" "1010" "3002" "4006" ...
##  $ diagnosis     : chr  "5/21/1997" "3/29/2000" "6/24/1998" "7/17/1997" ...
##  $ sex           : Factor w/ 2 levels "F","M": 2 2 1 2 2 2 1 2 2 2 ...
##  $ age           : int  53 19 52 38 57 17 18 16 15 40 ...
##  $ BT            : Factor w/ 10 levels "B","B1","B2",..: 3 3 5 2 3 2 2 2 3 3 ...
##  $ remission     : Factor w/ 2 levels "CR","REF": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CR            : chr  "CR" "CR" "CR" "CR" ...
##  $ date.cr       : chr  "8/6/1997" "6/27/2000" "8/17/1998" "9/8/1997" ...
##  $ t(4;11)       : logi  FALSE FALSE NA TRUE FALSE FALSE ...
##  $ t(9;22)       : logi  TRUE FALSE NA FALSE FALSE FALSE ...
##  $ cyto.normal   : logi  FALSE FALSE NA FALSE FALSE FALSE ...
##  $ citog         : chr  "t(9;22)" "simple alt." NA "t(4;11)" ...
##  $ mol.biol      : Factor w/ 6 levels "ALL1/AF4","BCR/ABL",..: 2 4 2 1 4 4 4 4 4 2
##  $ fusion protein: Factor w/ 3 levels "p190","p190/p210",..: 3 NA 1 NA NA NA NA NA I
##  $ mdr           : Factor w/ 2 levels "NEG","POS": 1 2 1 1 1 1 2 1 1 1 ...
##  $ kinet         : Factor w/ 2 levels "dyploid","hyperd.": 1 1 1 1 1 2 2 1 1 NA ...
##  $ ccr           : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ relapse       : logi  FALSE TRUE TRUE TRUE TRUE TRUE ...
##  $ transplant    : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ f.u           : chr  "BMT / DEATH IN CR" "REL" "REL" "REL" ...
##  $ date last seen: chr  NA "8/28/2000" "10/15/1999" "1/23/1998" ...
```

3.3.5 the table() function

```
af <- table(df2$age)
barplot(af, main = "Age frec.")
```



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
```

3.3.6 standard deviation and variance

```
std <- sd(df2$age, na.rm = TRUE)
vr <- var(df2$age, na.rm = TRUE)
```

3.3.7 Extremes

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

3.3.8 Table and (Frequency)

```
age_dit <- table(df2$age)
```

3.3.9 Quick summary

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

```
##       age          sex          BT      relapse
##  Min.   : 5.00   F   :42   B2     :36   Mode :logical
##  1st Qu.:19.00   M   :83   B3     :23   FALSE:35
##  Median :29.00   NA's: 3   B1     :19   TRUE :65
##  Mean   :32.37             T2     :15   NA's :28
##  3rd Qu.:45.50             B4     :12
##  Max.   :58.00             T3     :10
##  NA's   :5                 (Other):13
```

Chapter 4

Module 2

4.1 Lecture

4.1.1 2A

4.1.2 2B

4.2 Lab 2A

4.2.1 Read data in to R

4.2.2 read CSV - base functions

Chapter 5

Module 3

5.1 Lecture

5.1.1 3A

5.1.2 3B

5.2 Lab

Chapter 6

Module 4

6.1 Lecture

6.1.1 4A

6.1.2 4B

6.2 Lab