# 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

```r
dim(df2)
```

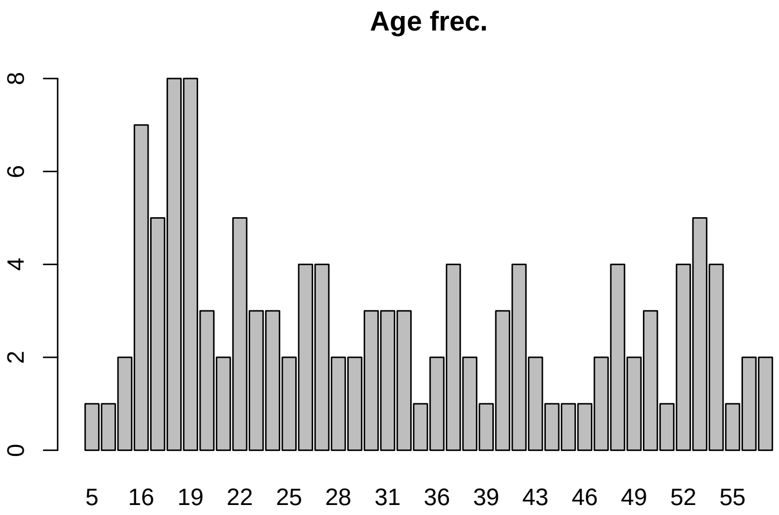
```
## [1] 128  21
```

```r
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 N
##  $ 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

```r
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
```

### 3.3.10 Patients older than 40

```
older_patients <- subset(df2, age > 40)
```

### 3.3.11 Patients who relapsed

```
relapsed_patients <- subset(df2, relapse == TRUE)
```

### 3.3.12 Subsetting and Filtering

#### 3.3.12.1 subset()

```
subset(df2, age > 40 & relapse == TRUE)
```

```
##         cod  diagnosis sex age BT remission CR    date.cr t(4;11) t(9;22)
## 03002  3002  6/24/1998   F  52 B4        CR CR  8/17/1998      NA      NA
## 04007  4007  7/22/1997   M  57 B2        CR CR  9/17/1997   FALSE   FALSE
## 08012  8012 10/22/1998   M  55 B3        CR CR   1/9/1999   FALSE   FALSE
## 15004 15004  2/10/2000   M  44 B1        CR CR   4/3/2000    TRUE   FALSE
## 16004 16004  4/19/1997   F  58 B1        CR CR  7/15/1997    TRUE   FALSE
## 19005 19005 11/15/1997   F  48 B1        CR CR   2/3/1998   FALSE   FALSE
## 20002 20002   5/9/1997   F  58 B2        CR CR  8/19/1997   FALSE    TRUE
## 24005 24005   1/3/1997   F  45 B1        CR CR   4/8/1997    TRUE   FALSE
## 24017 24017  9/15/1998   M  57 B2        CR CR  12/7/1998   FALSE    TRUE
## 26003 26003  2/18/1998   F  49 B4        CR CR  4/21/1998   FALSE   FALSE
## 28028 28028   7/8/1998   M  47 B1        CR CR   9/3/1998    TRUE   FALSE
## 28036 28036 12/23/1998   M  52 B3        CR CR   3/8/1999   FALSE    TRUE
## 43001 43001 11/14/1996   M  41 B1        CR CR  1/30/1997   FALSE    TRUE
## 49006 49006  8/12/1998   F  43 B2        CR CR 11/19/1998      NA      NA
## 62003 62003  12/4/1998   M  53 B4        CR CR  1/28/1999   FALSE    TRUE
## 63001 63001   7/8/1997   M  49 B1        CR CR   9/2/1997      NA      NA
## 84004 84004  9/25/1998   M  50  B        CR CR  12/1/1998      NA      NA
## 16002 16002  4/10/1997   M  50 T3        CR CR  6/10/1997      NA      NA
## 43015 43015  2/29/2000   M  52 T2        CR CR   6/8/2000   FALSE   FALSE
##       cyto.normal         citog mol.biol fusion protein mdr   kinet   ccr
## 03002          NA          <NA>  BCR/ABL           p190 NEG dyploid FALSE
## 04007       FALSE       del(6q)      NEG           <NA> NEG dyploid FALSE
## 08012       FALSE   simple alt.      NEG           <NA> NEG dyploid FALSE
## 15004       FALSE       t(4;11) ALL1/AF4           <NA> NEG dyploid FALSE
## 16004       FALSE       t(4;11) ALL1/AF4           <NA> NEG dyploid FALSE
## 19005        TRUE        normal ALL1/AF4           <NA> NEG dyploid FALSE
## 20002       FALSE t(9;22)+other  BCR/ABL           p190 NEG dyploid FALSE
## 24005       FALSE       t(4;11) ALL1/AF4           <NA> NEG dyploid FALSE
## 24017       FALSE t(9;22)+other  BCR/ABL           p190 NEG hyperd. FALSE
## 26003       FALSE  del(p15/p16)  BCR/ABL           p210 NEG dyploid FALSE
## 28028       FALSE       t(4;11) ALL1/AF4           <NA> NEG dyploid FALSE
## 28036       FALSE       t(9;22)  BCR/ABL           p190 NEG dyploid FALSE
## 43001       FALSE       t(9;22)  BCR/ABL      p190/p210 POS dyploid FALSE
## 49006          NA          <NA>  BCR/ABL           p210 NEG dyploid FALSE
## 62003       FALSE t(9;22)+other  BCR/ABL           p210 NEG hyperd. FALSE
## 63001          NA          <NA> ALL1/AF4           <NA> NEG dyploid FALSE
## 84004          NA          <NA>  BCR/ABL           p190 NEG dyploid FALSE
## 16002          NA          <NA>      NEG           <NA> NEG hyperd. FALSE
## 43015        TRUE        normal      NEG           <NA> NEG dyploid FALSE
##       relapse transplant f.u date last seen
## 03002    TRUE      FALSE REL     10/15/1999
## 04007    TRUE      FALSE REL      11/4/1997
## 08012    TRUE      FALSE REL       4/9/1999
## 15004    TRUE      FALSE REL     12/19/2000
## 16004    TRUE      FALSE REL      12/9/1997
## 19005    TRUE      FALSE REL       2/4/1998
## 20002    TRUE      FALSE REL     12/15/1997
## 24005    TRUE      FALSE REL      8/28/1997
## 24017    TRUE      FALSE REL      2/22/2000
## 26003    TRUE      FALSE REL       7/1/1998
## 28028    TRUE      FALSE REL     10/20/1999
## 28036    TRUE      FALSE REL      3/15/1999
## 43001    TRUE      FALSE REL      6/28/1998
## 49006    TRUE      FALSE REL      4/26/1999
## 62003    TRUE      FALSE REL       8/8/2000
## 63001    TRUE      FALSE REL      6/10/1998
## 84004    TRUE      FALSE REL      1/25/1999
## 16002    TRUE      FALSE REL      12/7/1999
## 43015    TRUE      FALSE REL      3/15/2002
```

### 3.3.13 Indexing with

```
df2[df2$age > 40, ]                # filter rows
```

```
##         cod  diagnosis  sex age   BT remission                 CR     date.cr
## 01005  1005  5/21/1997    M  53   B2        CR                 CR    8/6/1997
## 03002  3002  6/24/1998    F  52   B4        CR                 CR   8/17/1998
## 04007  4007  7/22/1997    M  57   B2        CR                 CR   9/17/1997
## 08012  8012 10/22/1998    M  55   B3        CR                 CR    1/9/1999
## 09008  9008 12/17/1999    M  41   B3        CR                 CR   2/15/2000
## 12006 12006  2/20/1997    M  46   B3       REF                REF        <NA>
## 12019 12019   9/4/1997    M  53   B2        CR                 CR  11/11/1997
## 14016 14016  5/27/1999    M  53   B2      <NA>               <NA>        <NA>
## 15004 15004  2/10/2000    M  44   B1        CR                 CR    4/3/2000
## 16004 16004  4/19/1997    F  58   B1        CR                 CR   7/15/1997
## 16009 16009  7/11/2000    F  43   B2      <NA>               <NA>        <NA>
## 19005 19005 11/15/1997    F  48   B1        CR                 CR    2/3/1998
## 20002 20002   5/9/1997    F  58   B2        CR                 CR   8/19/1997
## 24005 24005   1/3/1997    F  45   B1        CR                 CR    4/8/1997
## 24011 24011   8/5/1997    F  51   B2      <NA> DEATH IN INDUCTION        <NA>
## 24017 24017  9/15/1998    M  57   B2        CR                 CR   12/7/1998
## NA     <NA>       <NA> <NA>   NA <NA>      <NA>               <NA>        <NA>
## 26003 26003  2/18/1998    F  49   B4        CR                 CR   4/21/1998
## 27004 27004 10/20/1998    F  48   B2       REF                REF        <NA>
## 28007 28007  2/21/1997    F  47   B3        CR                 CR    4/7/1997
## 28021 28021  3/18/1998    F  54   B3        CR        DEATH IN CR   5/22/1998
## 28028 28028   7/8/1998    M  47   B1        CR                 CR    9/3/1998
## 28032 28032  9/26/1998    F  52   B1        CR                 CR 10/30/1998
## 28036 28036 12/23/1998    M  52   B3        CR                 CR   3/8/1999
## NA.1   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 30001 30001  1/16/1997    F  54   B3      <NA> DEATH IN INDUCTION       <NA>
## NA.2   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 43001 43001 11/14/1996    M  41   B1        CR                 CR  1/30/1997
## 43007 43007 10/14/1997    M  54   B4        CR                 CR 12/30/1997
## 49006 49006  8/12/1998    F  43   B2        CR                 CR 11/19/1998
## 57001 57001  1/29/1997    F  53   B3      <NA> DEATH IN INDUCTION       <NA>
## 62001 62001 11/11/1997    F  50   B4       REF                REF       <NA>
## 62002 62002  1/15/1998    M  54   B4      <NA> DEATH IN INDUCTION       <NA>
## 62003 62003  12/4/1998    M  53   B4        CR                 CR  1/28/1999
## 63001 63001   7/8/1997    M  49   B1        CR                 CR   9/2/1997
## 84004 84004  9/25/1998    M  50    B        CR                 CR  12/1/1998
## NA.3   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 02020  2020  3/23/2000    F  48   T2      <NA> DEATH IN INDUCTION       <NA>
## 16002 16002  4/10/1997    M  50   T3        CR                 CR  6/10/1997
## 16007 16007  11/1/1998    M  41   T3        CR                 CR  11/5/1998
## 31015 31015  12/3/1998    M  48   T2      <NA> DEATH IN INDUCTION       <NA>
## 43006 43006  6/17/1997    M  41   T2       REF                REF       <NA>
## 43015 43015  2/29/2000    M  52   T2        CR                 CR   6/8/2000
## NA.4   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
##       t(4;11) t(9;22) cyto.normal            citog mol.biol fusion protein  mdr
## 01005   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL           p210  NEG
## 03002      NA      NA          NA             <NA>  BCR/ABL           p190  NEG
## 04007   FALSE   FALSE       FALSE          del(6q)      NEG           <NA>  NEG
## 08012   FALSE   FALSE       FALSE      simple alt.      NEG           <NA>  NEG
## 09008   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL           p190  NEG
## 12006   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL           p210  NEG
## 12019   FALSE   FALSE        TRUE           normal      NEG           <NA>  POS
## 14016   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL           p210  NEG
## 15004    TRUE   FALSE       FALSE          t(4;11) ALL1/AF4           <NA>  NEG
## 16004    TRUE   FALSE       FALSE          t(4;11) ALL1/AF4           <NA>  NEG
## 16009      NA      NA          NA             <NA>      NEG           <NA>  POS
## 19005   FALSE   FALSE        TRUE           normal ALL1/AF4           <NA>  NEG
## 20002   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL           p190  NEG
## 24005    TRUE   FALSE       FALSE          t(4;11) ALL1/AF4           <NA>  NEG
## 24011   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL           p210  POS
## 24017   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL           p190  NEG
## NA         NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 26003   FALSE   FALSE       FALSE     del(p15/p16)  BCR/ABL           p210  NEG
## 27004   FALSE    TRUE       FALSE t(9;22)+del(p15)  BCR/ABL           p190  NEG
## 28007   FALSE   FALSE        TRUE           normal      NEG           <NA>  NEG
## 28021   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL      p190/p210  NEG
## 28028    TRUE   FALSE       FALSE          t(4;11) ALL1/AF4           <NA>  NEG
## 28032    TRUE   FALSE       FALSE          t(4;11) ALL1/AF4           <NA>  NEG
## 28036   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL          p190  NEG
## NA.1       NA      NA          NA             <NA>     <NA>          <NA> <NA>
## 30001   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL          p190  NEG
## NA.2       NA      NA          NA             <NA>     <NA>          <NA> <NA>
## 43001   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL     p190/p210  POS
## 43007   FALSE   FALSE        TRUE           normal      NEG          <NA>  NEG
## 49006      NA      NA          NA             <NA>  BCR/ABL          p210  NEG
## 57001   FALSE   FALSE        TRUE           normal      NEG          <NA>  NEG
## 62001   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL          <NA>  NEG
## 62002   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL          <NA>  NEG
## 62003   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL          p210  NEG
## 63001      NA      NA          NA             <NA> ALL1/AF4          <NA>  NEG
## 84004      NA      NA          NA             <NA>  BCR/ABL          p190  NEG
## NA.3       NA      NA          NA             <NA>     <NA>          <NA> <NA>
## 02020   FALSE   FALSE       FALSE     complex alt.      NEG          <NA>  NEG
## 16002      NA      NA          NA             <NA>      NEG          <NA>  NEG
## 16007      NA      NA          NA             <NA>      NEG          <NA>  NEG
## 31015      NA      NA          NA             <NA>      NEG          <NA>  POS
## 43006   FALSE   FALSE       FALSE      simple alt.      NEG          <NA>  POS
## 43015   FALSE   FALSE        TRUE           normal      NEG          <NA>  NEG
## NA.4       NA      NA          NA             <NA>     <NA>          <NA> <NA>
##         kinet   ccr relapse transplant               f.u date last seen
## 01005 dyploid FALSE   FALSE       TRUE BMT / DEATH IN CR           <NA>
## 03002 dyploid FALSE    TRUE      FALSE               REL     10/15/1999
## 04007 dyploid FALSE    TRUE      FALSE               REL      11/4/1997
## 08012 dyploid FALSE    TRUE      FALSE               REL       4/9/1999
## 09008 hyperd.  TRUE   FALSE       TRUE         BMT / CCR       00/09/01
## 12006 dyploid    NA      NA         NA              <NA>           <NA>
## 12019 dyploid  TRUE   FALSE      FALSE               CCR       6/6/2002
## 14016    <NA>    NA      NA         NA              <NA>           <NA>
## 15004 dyploid FALSE    TRUE      FALSE               REL     12/19/2000
## 16004 dyploid FALSE    TRUE      FALSE               REL      12/9/1997
## 16009 dyploid  TRUE   FALSE      FALSE         CCR / OFF      5/23/2002
## 19005 dyploid FALSE    TRUE      FALSE               REL       2/4/1998
## 20002 dyploid FALSE    TRUE      FALSE               REL     12/15/1997
## 24005 dyploid FALSE    TRUE      FALSE               REL      8/28/1997
## 24011 dyploid    NA      NA         NA              <NA>           <NA>
## 24017 hyperd. FALSE    TRUE      FALSE               REL      2/22/2000
## NA       <NA>    NA      NA         NA              <NA>           <NA>
## 26003 dyploid FALSE    TRUE      FALSE               REL       7/1/1998
## 27004 dyploid    NA      NA         NA              <NA>           <NA>
## 28007 dyploid  TRUE   FALSE      FALSE               CCR      3/22/2002
## 28021 hyperd. FALSE   FALSE      FALSE DEATH IN CR (ICR)           <NA>
## 28028 dyploid FALSE    TRUE      FALSE               REL     10/20/1999
## 28032 dyploid  TRUE   FALSE      FALSE               CCR      5/16/2002
## 28036 dyploid FALSE    TRUE      FALSE               REL      3/15/1999
## NA.1     <NA>    NA      NA         NA              <NA>           <NA>
## 30001 hyperd.    NA      NA         NA              <NA>           <NA>
## NA.2     <NA>    NA      NA         NA              <NA>           <NA>
## 43001 dyploid FALSE    TRUE      FALSE               REL      6/28/1998
## 43007 hyperd.  TRUE   FALSE      FALSE               CCR      5/29/2002
## 49006 dyploid FALSE    TRUE      FALSE               REL      4/26/1999
## 57001 hyperd.    NA      NA         NA              <NA>           <NA>
## 62001 hyperd.    NA      NA         NA              <NA>           <NA>
## 62002 hyperd.    NA      NA         NA              <NA>           <NA>
## 62003 hyperd. FALSE    TRUE      FALSE               REL       8/8/2000
## 63001 dyploid FALSE    TRUE      FALSE               REL      6/10/1998
## 84004 dyploid FALSE    TRUE      FALSE               REL      1/25/1999
## NA.3     <NA>    NA      NA         NA              <NA>           <NA>
## 02020 dyploid    NA      NA         NA              <NA>           <NA>
## 16002 hyperd. FALSE    TRUE      FALSE               REL      12/7/1999
## 16007 dyploid  TRUE   FALSE      FALSE               CCR       1/8/2002
## 31015 dyploid    NA      NA         NA              <NA>           <NA>
## 43006 dyploid    NA      NA         NA              <NA>           <NA>
## 43015 dyploid FALSE    TRUE      FALSE               REL      3/15/2002
## NA.4     <NA>    NA      NA         NA              <NA>           <NA>
```

```
df2[df2$age > 40 & df2$relapse, ]   # multiple conditions (same as df$Relapse == T)
```

```
##         cod  diagnosis  sex age   BT remission   CR    date.cr t(4;11) t(9;22)
## 03002  3002  6/24/1998    F  52   B4        CR   CR  8/17/1998      NA      NA
## 04007  4007  7/22/1997    M  57   B2        CR   CR  9/17/1997   FALSE   FALSE
## 08012  8012 10/22/1998    M  55   B3        CR   CR   1/9/1999   FALSE   FALSE
## NA     <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## NA.1   <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## 15004 15004  2/10/2000    M  44   B1        CR   CR   4/3/2000    TRUE   FALSE
## 16004 16004  4/19/1997    F  58   B1        CR   CR  7/15/1997    TRUE   FALSE
## 19005 19005 11/15/1997    F  48   B1        CR   CR   2/3/1998   FALSE   FALSE
## 20002 20002   5/9/1997    F  58   B2        CR   CR  8/19/1997   FALSE    TRUE
## 24005 24005   1/3/1997    F  45   B1        CR   CR   4/8/1997    TRUE   FALSE
## NA.2   <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## 24017 24017  9/15/1998    M  57   B2        CR   CR  12/7/1998   FALSE    TRUE
## 26003 26003  2/18/1998    F  49   B4        CR   CR  4/21/1998   FALSE   FALSE
## NA.3   <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## 28028 28028   7/8/1998    M  47   B1        CR   CR   9/3/1998    TRUE   FALSE
## 28036 28036 12/23/1998    M  52   B3        CR   CR   3/8/1999   FALSE    TRUE
## NA.4   <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## NA.5   <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## NA.6   <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## 43001 43001 11/14/1996    M  41   B1        CR   CR  1/30/1997   FALSE    TRUE
## 49006 49006  8/12/1998    F  43   B2        CR   CR 11/19/1998      NA      NA
## NA.7   <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## NA.8   <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## NA.9   <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## 62003 62003  12/4/1998    M  53   B4        CR   CR  1/28/1999   FALSE    TRUE
## 63001 63001   7/8/1997    M  49   B1        CR   CR   9/2/1997      NA      NA
## 84004 84004  9/25/1998    M  50    B        CR   CR  12/1/1998      NA      NA
## NA.10  <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## NA.11  <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## 16002 16002  4/10/1997    M  50   T3        CR   CR  6/10/1997      NA      NA
## NA.12  <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## NA.13  <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
## 43015 43015  2/29/2000    M  52   T2        CR   CR   6/8/2000   FALSE   FALSE
## NA.14  <NA>       <NA> <NA>  NA <NA>      <NA> <NA>       <NA>      NA      NA
##       cyto.normal         citog mol.biol fusion protein  mdr   kinet   ccr
## 03002          NA          <NA>  BCR/ABL           p190  NEG dyploid FALSE
## 04007       FALSE       del(6q)      NEG           <NA>  NEG dyploid FALSE
## 08012       FALSE   simple alt.      NEG           <NA>  NEG dyploid FALSE
## NA             NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## NA.1           NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## 15004       FALSE       t(4;11) ALL1/AF4           <NA>  NEG dyploid FALSE
## 16004       FALSE       t(4;11) ALL1/AF4           <NA>  NEG dyploid FALSE
## 19005        TRUE        normal ALL1/AF4           <NA>  NEG dyploid FALSE
## 20002       FALSE t(9;22)+other  BCR/ABL           p190  NEG dyploid FALSE
## 24005       FALSE       t(4;11) ALL1/AF4           <NA>  NEG dyploid FALSE
## NA.2           NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## 24017       FALSE t(9;22)+other  BCR/ABL           p190  NEG hyperd. FALSE
## 26003       FALSE  del(p15/p16)  BCR/ABL           p210  NEG dyploid FALSE
## NA.3           NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## 28028       FALSE       t(4;11) ALL1/AF4           <NA>  NEG dyploid FALSE
## 28036       FALSE       t(9;22)  BCR/ABL           p190  NEG dyploid FALSE
## NA.4           NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## NA.5           NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## NA.6           NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## 43001       FALSE       t(9;22)  BCR/ABL      p190/p210  POS dyploid FALSE
## 49006          NA          <NA>  BCR/ABL           p210  NEG dyploid FALSE
## NA.7           NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## NA.8           NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## NA.9           NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## 62003       FALSE t(9;22)+other  BCR/ABL           p210  NEG hyperd. FALSE
## 63001          NA          <NA> ALL1/AF4           <NA>  NEG dyploid FALSE
## 84004          NA          <NA>  BCR/ABL           p190  NEG dyploid FALSE
## NA.10          NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## NA.11          NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## 16002          NA          <NA>      NEG           <NA>  NEG hyperd. FALSE
## NA.12          NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## NA.13          NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
## 43015        TRUE        normal      NEG           <NA>  NEG dyploid FALSE
## NA.14          NA          <NA>     <NA>           <NA> <NA>    <NA>    NA
##       relapse transplant  f.u date last seen
## 03002    TRUE      FALSE  REL      10/15/1999
## 04007    TRUE      FALSE  REL       11/4/1997
## 08012    TRUE      FALSE  REL        4/9/1999
## NA         NA         NA <NA>           <NA>
## NA.1       NA         NA <NA>           <NA>
## 15004    TRUE      FALSE  REL      12/19/2000
## 16004    TRUE      FALSE  REL       12/9/1997
## 19005    TRUE      FALSE  REL        2/4/1998
## 20002    TRUE      FALSE  REL      12/15/1997
## 24005    TRUE      FALSE  REL       8/28/1997
## NA.2       NA         NA <NA>           <NA>
## 24017    TRUE      FALSE  REL       2/22/2000
## 26003    TRUE      FALSE  REL        7/1/1998
## NA.3       NA         NA <NA>           <NA>
## 28028    TRUE      FALSE  REL      10/20/1999
## 28036    TRUE      FALSE  REL       3/15/1999
## NA.4       NA         NA <NA>           <NA>
## NA.5       NA         NA <NA>           <NA>
## NA.6       NA         NA <NA>           <NA>
## 43001    TRUE      FALSE  REL       6/28/1998
## 49006    TRUE      FALSE  REL       4/26/1999
## NA.7       NA         NA <NA>           <NA>
## NA.8       NA         NA <NA>           <NA>
## NA.9       NA         NA <NA>           <NA>
## 62003    TRUE      FALSE  REL        8/8/2000
## 63001    TRUE      FALSE  REL       6/10/1998
## 84004    TRUE      FALSE  REL       1/25/1999
## NA.10      NA         NA <NA>           <NA>
## NA.11      NA         NA <NA>           <NA>
## 16002    TRUE      FALSE  REL       12/7/1999
## NA.12      NA         NA <NA>           <NA>
## NA.13      NA         NA <NA>           <NA>
## 43015    TRUE      FALSE  REL       3/15/2002
## NA.14      NA         NA <NA>           <NA>
```

```
df2[, c("age", "BT")]             # select columns
```

```
##       age BT
## 01005  53 B2
## 01010  19 B2
## 03002  52 B4
## 04006  38 B1
## 04007  57 B2
## 04008  17 B1
## 04010  18 B1
## 04016  16 B1
## 06002  15 B2
## 08001  40 B2
## 08011  33 B3
## 08012  55 B3
## 08018   5 B3
## 08024  18 B2
## 09008  41 B3
## 09017  27  B
## 11005  27 B2
## 12006  46 B3
## 12007  37 B2
## 12012  36 B3
## 12019  53 B2
## 12026  39 B2
## 14016  53 B2
## 15001  20 B1
## 15004  44 B1
## 15005  28 B2
## 16004  58 B1
## 16009  43 B2
## 19005  48 B1
## 20002  58 B2
## 22009  19  B
## 22010  26  B
## 22011  19 B2
## 22013  32 B2
## 24001  17 B2
## 24005  45 B1
## 24008  20 B2
## 24010  16 B2
## 24011  51 B2
## 24017  57 B2
## 24018  29 B2
## 24019  16 B4
## 24022  32 B4
## 25003  15 B2
## 25006  NA B2
## 26001  21 B2
## 26003  49 B4
## 26005  38 B2
## 26008  17 B1
## 27003  26 B2
## 27004  48 B2
## 28001  16 B3
## 28003  18 B4
## 28005  17 B3
## 28006  22 B3
## 28007  47 B3
## 28019  21 B4
## 28021  54 B3
## 28023  26 B3
## 28024  19 B1
## 28028  47 B1
## 28031  18 B1
## 28032  52 B1
## 28035  27 B3
## 28036  52 B3
## 28037  18 B3
## 28042  18 B3
## 28043  23 B3
## 28044  16 B3
## 28047  NA B3
## 30001  54 B3
## 31007  25 B1
## 31011  31 B3
## 33005  19 B1
## 36001  24 B4
## 36002  23 B2
## 37013  NA B2
## 43001  41 B1
## 43004  37 B3
## 43007  54 B4
## 43012  18 B4
## 48001  19 B2
## 49006  43 B2
## 57001  53 B3
## 62001  50 B4
## 62002  54 B4
## 62003  53 B4
## 63001  49 B1
## 64001  20 B2
## 64002  26 B2
## 65005  22 B2
## 68001  36 B1
## 68003  27 B2
## 84004  50  B
## LAL5   NA  B
## 01003  31  T
## 01007  16 T3
## 02020  48 T2
## 04018  17 T2
## 09002  40 T3
## 10005  22 T2
## 11002  30  T
## 12008  18 T4
## 15006  22 T2
## 16002  50 T3
## 16007  41 T3
## 17003  40  T
## 18001  28 T2
## 19002  25 T3
## 19008  16 T2
## 19014  31 T2
## 19017  14 T2
## 20005  24 T1
## 24006  19 T4
## 26009  37  T
## 28008  23 T2
## 28009  30 T3
## 31015  48 T2
## 37001  22 T2
## 43006  41 T2
## 43015  52 T2
## 44001  32 T3
## 49004  24 T3
## 56007  37 T3
## 64005  19 T2
## 65003  30 T3
## 83001  29 T2
## LAL4   NA  T
```

```
df <- df2[df2$sex == "F", ] # female patients only (#Assignment with condition)
```

### 3.3.14 with()

```
dfrt2 <- with(df2, df2[age > 40 & relapse == TRUE, ]) #for cleaner syntax
df2[which(df2$age > 40 & df2$relapse == TRUE), ] # more cleanr syntax
```

```
##         cod  diagnosis sex age BT remission CR    date.cr t(4;11) t(9;22)
## 03002  3002  6/24/1998   F  52 B4        CR CR  8/17/1998      NA      NA
## 04007  4007  7/22/1997   M  57 B2        CR CR  9/17/1997   FALSE   FALSE
## 08012  8012 10/22/1998   M  55 B3        CR CR   1/9/1999   FALSE   FALSE
## 15004 15004  2/10/2000   M  44 B1        CR CR   4/3/2000    TRUE   FALSE
## 16004 16004  4/19/1997   F  58 B1        CR CR  7/15/1997    TRUE   FALSE
## 19005 19005 11/15/1997   F  48 B1        CR CR   2/3/1998   FALSE   FALSE
## 20002 20002   5/9/1997   F  58 B2        CR CR  8/19/1997   FALSE    TRUE
## 24005 24005   1/3/1997   F  45 B1        CR CR   4/8/1997    TRUE   FALSE
## 24017 24017  9/15/1998   M  57 B2        CR CR  12/7/1998   FALSE    TRUE
## 26003 26003  2/18/1998   F  49 B4        CR CR  4/21/1998   FALSE   FALSE
## 28028 28028   7/8/1998   M  47 B1        CR CR   9/3/1998    TRUE   FALSE
## 28036 28036 12/23/1998   M  52 B3        CR CR   3/8/1999   FALSE    TRUE
## 43001 43001 11/14/1996   M  41 B1        CR CR  1/30/1997   FALSE    TRUE
## 49006 49006  8/12/1998   F  43 B2        CR CR 11/19/1998      NA      NA
## 62003 62003  12/4/1998   M  53 B4        CR CR  1/28/1999   FALSE    TRUE
## 63001 63001   7/8/1997   M  49 B1        CR CR   9/2/1997      NA      NA
## 84004 84004  9/25/1998   M  50  B        CR CR  12/1/1998      NA      NA
## 16002 16002  4/10/1997   M  50 T3        CR CR  6/10/1997      NA      NA
## 43015 43015  2/29/2000   M  52 T2        CR CR   6/8/2000   FALSE   FALSE
##       cyto.normal         citog mol.biol fusion protein mdr   kinet   ccr
## 03002          NA          <NA>  BCR/ABL           p190 NEG dyploid FALSE
## 04007       FALSE       del(6q)      NEG           <NA> NEG dyploid FALSE
## 08012       FALSE   simple alt.      NEG           <NA> NEG dyploid FALSE
## 15004       FALSE       t(4;11) ALL1/AF4           <NA> NEG dyploid FALSE
## 16004       FALSE       t(4;11) ALL1/AF4           <NA> NEG dyploid FALSE
## 19005        TRUE        normal ALL1/AF4           <NA> NEG dyploid FALSE
## 20002       FALSE t(9;22)+other  BCR/ABL           p190 NEG dyploid FALSE
## 24005       FALSE       t(4;11) ALL1/AF4           <NA> NEG dyploid FALSE
## 24017       FALSE t(9;22)+other  BCR/ABL           p190 NEG hyperd. FALSE
## 26003       FALSE  del(p15/p16)  BCR/ABL           p210 NEG dyploid FALSE
## 28028       FALSE       t(4;11) ALL1/AF4           <NA> NEG dyploid FALSE
## 28036       FALSE       t(9;22)  BCR/ABL           p190 NEG dyploid FALSE
## 43001       FALSE       t(9;22)  BCR/ABL      p190/p210 POS dyploid FALSE
## 49006          NA          <NA>  BCR/ABL           p210 NEG dyploid FALSE
## 62003       FALSE t(9;22)+other  BCR/ABL           p210 NEG hyperd. FALSE
## 63001          NA          <NA> ALL1/AF4           <NA> NEG dyploid FALSE
## 84004          NA          <NA>  BCR/ABL           p190 NEG dyploid FALSE
## 16002          NA          <NA>      NEG           <NA> NEG hyperd. FALSE
## 43015        TRUE        normal      NEG           <NA> NEG dyploid FALSE
##       relapse transplant f.u date last seen
## 03002    TRUE      FALSE REL     10/15/1999
## 04007    TRUE      FALSE REL      11/4/1997
## 08012    TRUE      FALSE REL       4/9/1999
## 15004    TRUE      FALSE REL     12/19/2000
## 16004    TRUE      FALSE REL      12/9/1997
## 19005    TRUE      FALSE REL       2/4/1998
## 20002    TRUE      FALSE REL     12/15/1997
## 24005    TRUE      FALSE REL      8/28/1997
## 24017    TRUE      FALSE REL      2/22/2000
## 26003    TRUE      FALSE REL       7/1/1998
## 28028    TRUE      FALSE REL     10/20/1999
## 28036    TRUE      FALSE REL      3/15/1999
## 43001    TRUE      FALSE REL      6/28/1998
## 49006    TRUE      FALSE REL      4/26/1999
## 62003    TRUE      FALSE REL       8/8/2000
## 63001    TRUE      FALSE REL      6/10/1998
## 84004    TRUE      FALSE REL      1/25/1999
## 16002    TRUE      FALSE REL      12/7/1999
## 43015    TRUE      FALSE REL      3/15/2002
```

### 3.3.15 match() / %in% (matching values)

```
dfrt <- df2[df2$BT %in% c("B2", "B3"), ]
```

### 3.3.16 Logical indexing directly

```
df2[df2$relapse == F | df2$sex == "F", ]
```

```
##         cod  diagnosis  sex age   BT remission                 CR    date.cr
## 01005  1005  5/21/1997    M  53   B2        CR                 CR   8/6/1997
## 03002  3002  6/24/1998    F  52   B4        CR                 CR  8/17/1998
## 04010  4010 10/30/1997    F  18   B1        CR                 CR   1/7/1998
## 08011  8011  8/21/1998    M  33   B3        CR                 CR  10/8/1998
## NA     <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 09008  9008 12/17/1999    M  41   B3        CR                 CR  2/15/2000
## 09017  9017   2/3/2000    F  27    B        CR                 CR  3/23/2000
## 11005 11005   6/1/1998    M  27   B2        CR        DEATH IN CR   8/3/1998
## NA.1   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 12012 12012  5/21/1997    F  36   B3       REF                REF       <NA>
## 12019 12019   9/4/1997    M  53   B2        CR                 CR 11/11/1997
## 12026 12026  5/29/1998    M  39   B2       REF                REF       <NA>
## NA.2   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 15001 15001   9/3/1997    M  20   B1        CR                 CR 11/11/1997
## 16004 16004  4/19/1997    F  58   B1        CR                 CR  7/15/1997
## 16009 16009  7/11/2000    F  43   B2      <NA>               <NA>       <NA>
## 19005 19005 11/15/1997    F  48   B1        CR                 CR   2/3/1998
## 20002 20002   5/9/1997    F  58   B2        CR                 CR  8/19/1997
## 22009 22009  8/10/1999    F  19    B      <NA>               <NA>       <NA>
## 22010 22010 12/31/1999    F  26    B      <NA>               <NA>       <NA>
## 22011 22011   4/7/2000    M  19   B2        CR                 CR  5/19/2000
## NA.3   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 24001 24001  10/4/1996    F  17   B2        CR                 CR 12/20/1996
## 24005 24005   1/3/1997    F  45   B1        CR                 CR   4/8/1997
## 24008 24008  5/14/1997    F  20   B2        CR                 CR  7/31/1997
## 24010 24010   6/3/1997    F  16   B2        CR                 CR  8/11/1997
## 24011 24011   8/5/1997    F  51   B2      <NA> DEATH IN INDUCTION       <NA>
## 24018 24018  2/18/1999    F  29   B2        CR                 CR   5/4/1999
## 24022 24022 12/21/1999    F  32   B4       REF                REF       <NA>
## 25003 25003  5/22/1998    M  15   B2        CR                 CR   8/4/1998
## 25006 25006  3/18/2000 <NA>  NA   B2        CR                 CR   5/8/2000
## 26001 26001  9/27/1997    M  21   B2        CR                 CR 12/11/1997
## 26003 26003  2/18/1998    F  49   B4        CR                 CR  4/21/1998
## 26008 26008  8/25/1999    F  17   B1        CR                 CR 10/14/1999
## 27003 27003  1/17/1998    F  26   B2        CR                 CR  3/16/1998
## 27004 27004 10/20/1998    F  48   B2       REF                REF       <NA>
## NA.4   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 28003 28003 11/28/1996    M  18   B4        CR                 CR  1/17/1997
## 28007 28007  2/21/1997    F  47   B3        CR                 CR   4/7/1997
## 28019 28019  2/10/1998    M  21   B4        CR                 CR   4/2/1998
## 28021 28021  3/18/1998    F  54   B3        CR        DEATH IN CR  5/22/1998
## 28024 28024  4/19/1998    F  19   B1        CR                 CR  6/17/1998
## 28032 28032  9/26/1998    F  52   B1        CR                 CR 10/30/1998
## 28035 28035 12/21/1998    M  27   B3        CR                 CR  2/12/1999
## 30001 30001  1/16/1997    F  54   B3      <NA> DEATH IN INDUCTION       <NA>
## 33005 33005  2/10/1998    F  19   B1        CR                 CR  4/29/1998
## 36001 36001  9/29/1997    F  24   B4        CR                 CR  12/5/1997
## NA.5   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 43004 43004   2/4/1997    F  37   B3        CR                 CR   4/1/1997
## 43007 43007 10/14/1997    M  54   B4        CR                 CR 12/30/1997
## 48001 48001  3/22/1997    M  19   B2        CR                 CR  5/20/1997
## 49006 49006  8/12/1998    F  43   B2        CR                 CR 11/19/1998
## 57001 57001  1/29/1997    F  53   B3      <NA> DEATH IN INDUCTION       <NA>
## 62001 62001 11/11/1997    F  50   B4       REF                REF       <NA>
## NA.6   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 64002 64002 10/21/1997    F  26   B2        CR                 CR  1/21/1998
## NA.7   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 68001 68001  5/15/1997    M  36   B1        CR                 CR  7/22/1997
## 68003 68003  4/11/2000    F  27   B2      <NA> DEATH IN INDUCTION       <NA>
## NA.8   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 01007  1007  9/30/1998    F  16   T3        CR                 CR 11/30/1998
## 02020  2020  3/23/2000    F  48   T2      <NA> DEATH IN INDUCTION       <NA>
## 04018  4018  3/24/2000    M  17   T2        CR                 CR  5/23/2000
## 09002  9002  5/14/1998    F  40   T3        CR                 CR  7/21/1998
## NA.9   <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 16007 16007  11/1/1998    M  41   T3        CR                 CR  11/5/1998
## 17003 17003   4/8/1997    F  40    T       REF                REF       <NA>
## 18001 18001  4/23/1997    F  28   T2       REF                REF       <NA>
## 19008 19008  4/29/1998    F  16   T2       REF                REF       <NA>
## 20005 20005  3/15/2000    M  24   T1        CR                 CR   5/5/2000
## 24006 24006  1/14/1997    F  19   T4        CR                 CR  not known
## 28008 28008  3/27/1997    M  23   T2        CR                 CR  5/27/1997
## 28009 28009  4/19/1997    F  30   T3        CR                 CR  6/13/1997
## NA.10  <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## NA.11  <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 49004 49004  9/18/1997    M  24   T3        CR                 CR 11/11/1997
## 56007 56007   8/6/1999    M  37   T3        CR                 CR  9/24/1999
## NA.12  <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
## 83001 83001 10/23/1998    M  29   T2        CR                 CR 12/21/1998
## NA.13  <NA>       <NA> <NA>  NA <NA>      <NA>               <NA>       <NA>
##       t(4;11) t(9;22) cyto.normal            citog mol.biol fusion protein  mdr
## 01005   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL           p210  NEG
## 03002      NA      NA          NA             <NA>  BCR/ABL           p190  NEG
## 04010   FALSE   FALSE       FALSE     complex alt.      NEG           <NA>  POS
## 08011   FALSE   FALSE       FALSE     del(p15/p16)  BCR/ABL      p190/p210  NEG
## NA         NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 09008   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL           p190  NEG
## 09017   FALSE   FALSE        TRUE           normal      NEG           <NA>  NEG
## 11005   FALSE   FALSE       FALSE  del(7q) + altro  BCR/ABL           p190  NEG
## NA.1       NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 12012   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL           p190  NEG
## 12019   FALSE   FALSE        TRUE           normal      NEG           <NA>  POS
## 12026   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL      p190/p210 <NA>
## NA.2       NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 15001   FALSE   FALSE        TRUE           normal      NEG           <NA>  NEG
## 16004    TRUE   FALSE       FALSE          t(4;11) ALL1/AF4           <NA>  NEG
## 16009      NA      NA          NA             <NA>      NEG           <NA>  POS
## 19005   FALSE   FALSE        TRUE           normal ALL1/AF4           <NA>  NEG
## 20002   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL           p190  NEG
## 22009   FALSE   FALSE       FALSE      simple alt.      NEG           <NA>  NEG
## 22010   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL      p190/p210  NEG
## 22011   FALSE   FALSE        TRUE           normal      NEG           <NA>  NEG
## NA.3       NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 24001      NA      NA          NA             <NA>  BCR/ABL           p190  NEG
## 24005    TRUE   FALSE       FALSE          t(4;11) ALL1/AF4           <NA>  NEG
## 24008      NA      NA          NA             <NA>      NEG           <NA>  NEG
## 24010   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL      p190/p210  NEG
## 24011   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL           p210  POS
## 24018      NA      NA          NA             <NA>      NEG           <NA>  POS
## 24022   FALSE    TRUE       FALSE          t(9;22)  BCR/ABL           p190  POS
## 25003   FALSE   FALSE       FALSE      simple alt.      NEG           <NA>  POS
## 25006      NA      NA          NA             <NA>      NEG           <NA>  NEG
## 26001      NA      NA          NA             <NA>      NEG           <NA>  POS
## 26003   FALSE   FALSE       FALSE     del(p15/p16)  BCR/ABL           p210  NEG
## 26008   FALSE   FALSE        TRUE           normal ALL1/AF4           <NA>  NEG
## 27003      NA      NA          NA             <NA>  BCR/ABL      p190/p210  POS
## 27004   FALSE    TRUE       FALSE t(9;22)+del(p15)  BCR/ABL           p190  NEG
## NA.4       NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 28003      NA      NA          NA             <NA> E2A/PBX1           <NA>  NEG
## 28007   FALSE   FALSE        TRUE           normal      NEG           <NA>  NEG
## 28019      NA      NA          NA             <NA>  BCR/ABL           p190  NEG
## 28021   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL      p190/p210  NEG
## 28024   FALSE   FALSE       FALSE     complex alt.      NEG           <NA>  NEG
## 28032    TRUE   FALSE       FALSE          t(4;11) ALL1/AF4           <NA>  NEG
## 28035      NA      NA          NA             <NA>      NEG           <NA>  POS
## 30001   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL           p190  NEG
## 33005   FALSE   FALSE       FALSE     complex alt.      NEG           <NA>  NEG
## 36001   FALSE   FALSE        TRUE           normal E2A/PBX1           <NA>  NEG
## NA.5       NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 43004      NA      NA          NA             <NA>      NEG           <NA>  NEG
## 43007   FALSE   FALSE        TRUE           normal      NEG           <NA>  NEG
## 48001   FALSE   FALSE       FALSE     complex alt.      NEG           <NA>  NEG
## 49006      NA      NA          NA             <NA>  BCR/ABL           p210  NEG
## 57001   FALSE   FALSE        TRUE           normal      NEG           <NA>  NEG
## 62001   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL           <NA>  NEG
## NA.6       NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 64002      NA      NA          NA             <NA>      NEG           <NA>  NEG
## NA.7       NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 68001   FALSE   FALSE        TRUE           normal      NEG           <NA>  NEG
## 68003   FALSE    TRUE       FALSE    t(9;22)+other  BCR/ABL           p190  NEG
## NA.8       NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 01007   FALSE   FALSE       FALSE      simple alt.   NUP-98           <NA>  NEG
## 02020   FALSE   FALSE       FALSE     complex alt.      NEG           <NA>  NEG
## 04018   FALSE   FALSE       FALSE      simple alt.      NEG           <NA>  NEG
## 09002      NA      NA          NA             <NA>      NEG           <NA>  NEG
## NA.9       NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 16007      NA      NA          NA             <NA>      NEG           <NA>  NEG
## 17003      NA      NA          NA             <NA>      NEG           <NA>  POS
## 18001      NA      NA          NA             <NA>      NEG           <NA>  NEG
## 19008   FALSE   FALSE        TRUE           normal      NEG           <NA>  POS
## 20005      NA      NA          NA             <NA>      NEG           <NA>  NEG
## 24006   FALSE   FALSE       FALSE      simple alt.      NEG           <NA>  NEG
## 28008   FALSE   FALSE        TRUE           normal      NEG           <NA>  NEG
## 28009      NA      NA          NA             <NA>      NEG           <NA>  NEG
## NA.10      NA      NA          NA             <NA>     <NA>           <NA> <NA>
## NA.11      NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 49004   FALSE   FALSE       FALSE          del(7q)      NEG           <NA>  POS
## 56007      NA      NA          NA             <NA>      NEG           <NA>  NEG
## NA.12      NA      NA          NA             <NA>     <NA>           <NA> <NA>
## 83001   FALSE   FALSE       FALSE     complex alt.      NEG           <NA>  NEG
## NA.13      NA      NA          NA             <NA>     <NA>           <NA> <NA>
##          kinet   ccr relapse transplant               f.u  date last seen
## 01005 dyploid FALSE   FALSE       TRUE BMT / DEATH IN CR            <NA>
## 03002 dyploid FALSE    TRUE      FALSE               REL      10/15/1999
## 04010 hyperd. FALSE    TRUE      FALSE               REL        3/5/1998
## 08011 dyploid FALSE   FALSE       TRUE BMT / DEATH IN CR            <NA>
## NA       <NA>    NA      NA         NA              <NA>            <NA>
## 09008 hyperd.  TRUE   FALSE       TRUE         BMT / CCR        00/09/01
## 09017 dyploid FALSE    TRUE      FALSE               REL       9/11/2001
## 11005 dyploid FALSE   FALSE      FALSE       DEATH IN CR            <NA>
## NA.1     <NA>    NA      NA         NA              <NA>            <NA>
## 12012 dyploid    NA      NA         NA              <NA>            <NA>
## 12019 dyploid  TRUE   FALSE      FALSE               CCR        6/6/2002
## 12026 dyploid FALSE   FALSE      FALSE       DEATH IN CR            <NA>
## NA.2     <NA>    NA      NA         NA              <NA>            <NA>
## 15001 dyploid  TRUE   FALSE      FALSE               CCR       6/21/2002
## 16004 dyploid FALSE    TRUE      FALSE               REL       12/9/1997
## 16009 dyploid  TRUE   FALSE      FALSE         CCR / OFF       5/23/2002
## 19005 dyploid FALSE    TRUE      FALSE               REL        2/4/1998
## 20002 dyploid FALSE    TRUE      FALSE               REL      12/15/1997
## 22009 dyploid    NA      NA         NA              <NA>            <NA>
## 22010 dyploid    NA      NA         NA              <NA>            <NA>
## 22011 dyploid  TRUE   FALSE      FALSE               CCR       7/31/2002
## NA.3     <NA>    NA      NA         NA              <NA>            <NA>
## 24001 dyploid FALSE    TRUE      FALSE               REL       2/10/1997
## 24005 dyploid FALSE    TRUE      FALSE               REL       8/28/1997
## 24008 dyploid FALSE   FALSE       TRUE         BMT / CCR 00/09/20+T12501
## 24010 dyploid FALSE    TRUE       TRUE         BMT / REL       8/24/1998
## 24011 dyploid    NA      NA         NA              <NA>            <NA>
## 24018 dyploid FALSE    TRUE      FALSE               REL       7/22/2000
## 24022 dyploid    NA      NA         NA              <NA>            <NA>
## 25003 dyploid  TRUE   FALSE      FALSE               CCR       6/10/2002
## 25006    <NA>  TRUE   FALSE      FALSE               CCR        3/3/2002
## 26001 dyploid  TRUE   FALSE      FALSE               CCR       7/31/2002
## 26003 dyploid FALSE    TRUE      FALSE               REL        7/1/1998
## 26008 dyploid FALSE    TRUE      FALSE               REL       6/26/2000
## 27003 dyploid FALSE    TRUE      FALSE               REL        5/6/1998
## 27004 dyploid    NA      NA         NA              <NA>            <NA>
## NA.4     <NA>    NA      NA         NA              <NA>            <NA>
## 28003 hyperd.  TRUE   FALSE      FALSE               CCR      12/31/2002
## 28007 dyploid  TRUE   FALSE      FALSE               CCR       3/22/2002
## 28019 hyperd.  TRUE   FALSE       TRUE         BMT / CCR       3/21/2001
## 28021 hyperd. FALSE   FALSE      FALSE DEATH IN CR (ICR)            <NA>
## 28024 hyperd.  TRUE   FALSE      FALSE               CCR      12/31/2002
## 28032 dyploid  TRUE   FALSE      FALSE               CCR       5/16/2002
## 28035 hyperd.  TRUE   FALSE      FALSE               CCR       5/20/2002
## 30001 hyperd.    NA      NA         NA              <NA>            <NA>
## 33005 dyploid  TRUE   FALSE      FALSE               CCR       6/28/2002
## 36001 dyploid FALSE    TRUE      FALSE               REL        1/7/1998
## NA.5     <NA>    NA      NA         NA              <NA>            <NA>
## 43004 dyploid  TRUE   FALSE      FALSE               CCR       3/20/2001
## 43007 hyperd.  TRUE   FALSE      FALSE               CCR       5/29/2002
## 48001 hyperd. FALSE   FALSE      FALSE MUD / DEATH IN CR      12/18/1998
## 49006 dyploid FALSE    TRUE      FALSE               REL       4/26/1999
## 57001 hyperd.    NA      NA         NA              <NA>            <NA>
## 62001 hyperd.    NA      NA         NA              <NA>            <NA>
## NA.6     <NA>    NA      NA         NA              <NA>            <NA>
## 64002 hyperd. FALSE   FALSE       TRUE BMT / DEATH IN CR            <NA>
## NA.7     <NA>    NA      NA         NA              <NA>            <NA>
## 68001 dyploid  TRUE   FALSE      FALSE               CCR       5/10/2002
## 68003    <NA>    NA      NA         NA              <NA>            <NA>
## NA.8     <NA>    NA      NA         NA              <NA>            <NA>
## 01007 hyperd. FALSE   FALSE       TRUE BMT / DEATH IN CR            <NA>
## 02020 dyploid    NA      NA         NA              <NA>            <NA>
## 04018 dyploid  TRUE   FALSE      FALSE               CCR       5/14/2001
## 09002 dyploid FALSE    TRUE      FALSE         REL / SNC       9/14/1999
## NA.9     <NA>    NA      NA         NA              <NA>            <NA>
## 16007 dyploid  TRUE   FALSE      FALSE               CCR        1/8/2002
## 17003 dyploid    NA      NA         NA              <NA>            <NA>
## 18001 hyperd.    NA      NA         NA              <NA>            <NA>
## 19008 dyploid    NA      NA         NA              <NA>            <NA>
## 20005 dyploid  TRUE   FALSE      FALSE               CCR       3/20/2002
## 24006 dyploid  TRUE   FALSE      FALSE               CCR        6/5/2002
## 28008 hyperd.  TRUE   FALSE      FALSE               CCR        4/9/2002
## 28009 dyploid FALSE    TRUE      FALSE               REL       6/30/1998
## NA.10    <NA>    NA      NA         NA              <NA>            <NA>
## NA.11    <NA>    NA      NA         NA              <NA>            <NA>
## 49004 dyploid  TRUE   FALSE      FALSE               CCR       6/14/2001
## 56007 dyploid  TRUE   FALSE      FALSE               CCR       1/26/2001
## NA.12    <NA>    NA      NA         NA              <NA>            <NA>
## 83001 hyperd.  TRUE   FALSE      FALSE               CCR       5/24/2002
## NA.13    <NA>    NA      NA         NA              <NA>            <NA>
```

## 3.4 Lab 1B Tasks

```
# Task 1
# Patients younger than 20
subset(ALL_df, age < 20)

# Task 2
# Age and Sex for patients with BT = "B2"
subset(ALL_df, BT == "B2", select = c(age, sex))

# Task 3
# Male patients older than 40
subset(ALL_df, sex == "M" & age > 40)

# Female OR Relapse = Yes
subset(ALL_df, sex == "F" | relapse == T)

# Mini-Challenge
# Male + Relapse + Age > 30
subset(ALL_df, sex == "M" & relapse == T & age > 30)

#OR
ALL_df[ALL_df$sex == "M" & ALL_df$relapse == T & ALL_df$age > 30, ]
```

# 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

read CSV - base functions

```
bp <- read.csv2("Desktop/R/data/BloodPressure_Data.csv") # no sepqration
# take a quick look at the data
head(bp)
```

```
bp <- read.csv2("Desktop/R/data/BloodPressure_Data.csv", sep = ",") # no sepqration
# take another look at the data
head(bp)
str(bp)
```

readr functions

```
library(readr)
```

Read the CSV file

```
bp_data <- read_csv("Desktop/R/data/BloodPressure_Data.csv")
```

Take a quick look at the data

```
head(bp_data)
str(bp_data)
```

Work with date

```
library(readr)
library("lubridate")
```

read ALL data

```
bp <-  read.csv2("Desktop/R/data/BloodPressure_wDates.csv", sep = ",")
```

Convert date column and extract year

```
bp$Date <- ymd(bp$Date)
bp$Year <- year(bp$Date)
```

Filtering blood pressure patients by year and gender

```
subset(bp, Year == 2003 & Gender == "f")
```

### 4.2.2 Conditions and loops

If {} else {} statement

```r
if (condition) {
  # code if TRUE
} else {
  # code if FALSE
}
```

```r
# If else example
age <- 55
if (age > 50) {
  print("Older patient")
} else {
  print("Younger patient")
}
```

```r
# If {} else if {} else statement

if (condition1) {
  # code if condition1 is TRUE
} else if (condition2) {
  # code if condition2 is TRUE
} else {
  # code if none are TRUE
}
```

```r
# If else if example
age <- 35

if (age < 18) {
  print("Child")
} else if (age >= 18 & age < 60) {
  print("Adult")
} else {
  print("Senior")
}
```

```r
# for loops
for (i in 1:5) {
  print(i)
}
```

```
# for loop example
patients <- c("P1", "P2", "P3")
for (p in patients) {
  print(paste("Processing:", p))
}
```

```
# the apply () family vs. loops
# Using a for loop
m <- matrix(1:9, nrow=3)
row_sums <- c()
for (i in 1:nrow(m)) {
  row_sums[i] <- sum(m[i, ])
}
```

```
# Using apply()
row_sums2 <- apply(m, 1, sum)
```

```
# the apply() family
# apply()
apply(m, 1, sum)    # row sums
apply(m, 2, mean)   # column means
```

```
# lapply()
lapply(list(1:3, 4:6), mean)
```

```
# sapply()
sapply(list(1:3, 4:6), mean)
```

```
# lapply()
lapply(list(1:3, 4:6), mean)
```

```
# tapply()
ages <- c(21, 25, 30, 40, 35)
gender <- c("M", "M", "F", "F", "M")
tapply(ages, gender, mean)    # mean age by gender
```

```
# mapply()
nums1 <- 1:5
nums2 <- 6:10
mapply(sum, nums1, nums2)   # adds 1+6, 2+7, … 5+10
```

## 4.3 Lab 2A Tasks

Task 1 – Basic Filtering

```
#Use a for loop with if/else conditions
Go through each row of the dataset and:
Print a message if the patient has High BP (> 140) # Tip: use the paste function,
Print a message if the patient has Low BP (< 90)
Otherwise, mark them as Normal
# read data
bp_data <- read.csv2("Desktop/R/data/BloodPressure_wDates.csv", sep = ",")

# Use for loop
for (i in 1:nrow(bp_data)) {
  if (bp_data$BloodPressure[i] > 140) {
    print(paste("Patient", bp_data$ID[i], "has HIGH blood pressure"))
  } else if (bp_data$BloodPressure[i] < 90) {
    print(paste("Patient", bp_data$ID[i], "has LOW blood pressure"))
  } else {
    print(paste("Patient", bp_data$ID[i], "is NORMAL"))
  }
}


# Use apply() instead of loops
bp_data$BP_Status <- apply(bp_data, 1, function(row) {
  if (as.numeric(row["BloodPressure"]) > 140) {
    "HIGH"
  } else if (as.numeric(row["BloodPressure"]) < 90) {
    "LOW"
  } else {
    "NORMAL"
  }
})
```

## 4.4 Lab 2B

### 4.4.1 Basic plotting in R

Plot multiple panels in one plot

```
par(mfrow = c(3, 1))
layout(matrix(c(1, 1, 2, 3), nrow = 2, byrow = TRUE))
```

Histogram of Blood Pressure

```
hist(bp_data$BloodPressure, main="Blood Pressure Distribution",
     xlab="Blood Pressure", col="lightblue")
```

Boxplot of BP by Gender

```
boxplot(BloodPressure ~ Gender, data=bp_data,
        main="BP by Gender", xlab="Gender", ylab="Blood Pressure")
```

Scatterplot Age vs BP

```
plot(bp_data$Age, bp_data$BloodPressure,
     main="Age vs Blood Pressure", xlab="Age", ylab="BP")
```

### 4.4.2 ggplot2

```
# example syntax
#ggplot(data, aes(x, y)) + geom_*()
```

```
library(ggplot2)
# Scatterplot: Age vs Blood Pressure
ggplot(bp_data, aes(x=Age, y=BloodPressure)) +
  geom_point()
```

boxplot

```
ggplot(bp_data, aes(x=Group, y=BloodPressure)) +
  geom_boxplot()
```

Customizing ggplot2

```
ggplot(bp_data, aes(x = Age, y = BloodPressure)) +
  geom_point(color = "blue") +
  labs(
    title = "Age vs Blood Pressure",
    x = "Patient Age",
    y = "BP (mmHg)"
  ) +
  theme_minimal()
```

```
# Save last plot as PNG
ggsave("Age_BP_Scatter.png", width=6, height=4)
```

```
# Save specific plot object
p <- ggplot(bp_data, aes(x=Age, y=BloodPressure)) +
  geom_point()
ggsave("data/scatter_plot.png", plot=p)
```

## 4.5 Lab 2B Tasks

ggplot hands on tasks

```
library(lubridate)
library(ggplot2)
library(patchwork)
```

Load data

```
bp_data <- read.csv2("data/BloodPressure_wDates.csv", sep = ",")
```

Get the year in a new column

```
bp_data$Year <- year(bp_data$Date)
```

1. Bar plot: patient counts per group

```
# ggplot hands on tasks
ggplot(bp_data, aes(x=Group)) +
  geom_bar(fill="steelblue") +
  labs(title="Number of Patients per Group", x="Group", y="Count")
```

2. Histogram: Age distribution

```
ggplot(bp_data, aes(x=as.numeric(Age)))  +
  geom_histogram(binwidth=5, fill="lightgreen", color="black") +
  labs(title="Age Distribution of Patients", x="Age", y="Frequency")
```

3. Scatterplot: Age vs BloodPressure, colored by Group

```
ggplot(bp_data, aes(x=Age, y=BloodPressure, color=Group)) +
  geom_point() +
  labs(title="Age vs Blood Pressure by Group",
       x="Age", y="Blood Pressure")
```

Bonus hands on

```
library(ggplot2)
library(patchwork)   # install.packages("patchwork") if needed
```

1. Bar plot

```
p1 <- ggplot(bp_data, aes(x = Gender, y = BloodPressure, fill = Gender)) +
  stat_summary(fun = "mean", geom = "bar") +
  labs(title = "Average BP by Gender") +
  theme_minimal()
```

2. Histogram

```
p2 <- ggplot(bp_data, aes(x=as.numeric(Age)))  +
  geom_histogram(binwidth=5, fill="lightgreen", color="black") +
  labs(title="Age Distribution of Patients", x="Age", y="Frequency")
```

3. Boxplot

```
p3 <- ggplot(bp_data, aes(x = Group, y = BloodPressure, fill = Group)) +
  geom_boxplot() +
  labs(title = "BP by Group") +
  theme_minimal()
```

Combine plots into one figure

```
# horizontal layout
(p1 | p2 | p3)

# vertical layout
(p1 / p2 / p3)

# 2x2 grid
(p1 | p2) / p3
```

# Chapter 5

# Module 3

## 5.1 Lecture

### 5.1.1 3A

### 5.1.2 3B

## 5.2 Lab 3A

Install a package, e.g. GenomicRanges

```
BiocManager::install("GenomicRanges")
```

Load a package

```
library(GenomicRanges)
library(SummarizedExperiment)
```

Create simple SummarizedExperiment

```
counts <- matrix(rpois(20, 10), ncol=4)
colData <- DataFrame(condition=c("A","A","B","B"))
rowData <- DataFrame(gene=letters[1:5])

se <- SummarizedExperiment(assays=list(counts=counts),
                           colData=colData,
                           rowData=rowData)

se
```

Demo 2: ALL dataset

```
BiocManager::install("ALL")
library(ALL)
data(ALL)
ALL
```

## 5.3 Lab 3A Tasks

Extract and preview sample (patient) metadata

```
meta <- pData(ALL)
head(meta)   # first 6 rows

# Gender distribution
table(meta$sex)

# Mean age (ignoring missing values)
mean(meta$age, na.rm = TRUE)
```

Visualization in Bioconductor

```
boxplot(exprs(ALL)[,1:10], las=2, main="Expression values (first 10 samples)")
```

Load All package and data

```
# BiocManager::install("ALL")
library(SummarizedExperiment)
library(ALL)
data(ALL)
```

```
# Subset patients < 20
young_patients <- ALL[, pData(ALL)$age < 20]
dim(young_patients)

# Count patients by Immunophenotype (BT)
barplot(table(pData(ALL)$BT), main="Patients by Immunophenotype (BT)", ylab="Patients", xlab="Imm

# PCA on first 50 genes
expr <- exprs(ALL)[1:50, ]
pca <- prcomp(t(expr), scale. = TRUE)
plot(pca$x[,1:2], col = as.factor(pData(ALL)$BT),
     pch=19, main="PCA of 50 genes")


#Boxplot of Age by Sex
boxplot(age ~ sex, data = pData(ALL),
        main="Age Distribution by Sex", xlab="Sex", ylab="Age")

# Challenge (Filter missing age & re-run PCA)
ALL_clean <- ALL[, !is.na(pData(ALL)$age)]
expr_clean <- exprs(ALL_clean)[1:50, ]
pca_clean <- prcomp(t(expr_clean), scale. = TRUE)
plot(pca_clean$x[,1:2], col = as.factor(pData(ALL_clean)$BT),
     pch=19, main="PCA after removing NA ages")
```

## 5.4 Lab 3B

### Bioconductor Packages and Data sets

Install airway package

```
BiocManager::install("airway")

# load package and data
library("airway")
data("airway")   # loads the dataset into your environment
airway
```

Explore airway package

```r
ex <- assay(airway)[1:5, 1:5]   # expression counts
cols <- colData(airway)[1:5, ]    # sample metadata
rows <- rowData(airway)[1:5, ]    # gene metadata
```

### Hands on tasks

Subsetting treated vs untreated

```r
treated <- airway[, airway$dex == "trt"]
untreated <- airway[, airway$dex == "untrt"]

dim(treated)
dim(untreated)
```

Count treated vs untreated

```r
table(airway$dex)
```

Extract samples from a specific cell line

```r
subset_cell <- airway[, airway$cell == "N061011"]
```

Get number of genes

```r
nrow(airway)
```

ExperimentHub Demo

```r
# Load ExperimentHub
library(ExperimentHub)

# Create a hub object
eh <- ExperimentHub()

# Search for RNA-seq datasets
query(eh, "RNA-seq")

# Access a specific dataset by ID (example)
eh[["EH1234"]]   # Loads dataset into R
```

AnnotationHub Demo

```r
# Load AnnotationHub
library(AnnotationHub)
library("rtracklayer")
# Create a hub object
ah <- AnnotationHub()

# Search for human genome resources
query(ah, "Homo sapiens")

# Access an annotation dataset by ID (example)
ah[["AH83281"]]   # Loads GRCh38 GTF annotation into R
```

org.Hs.eg.db Demo

```r
# Install packages
BiocManager::install("AnnotationDbi")
BiocManager::install("org.Hs.eg.db")

# load packages
library(org.Hs.eg.db)
library(AnnotationDbi)

ids <- rownames(airway)[1:5]
mapIds(org.Hs.eg.db,
       keys = ids,
       keytype = "ENSEMBL",
       column = "SYMBOL")
```

## 5.5 Lab 3B Tasks

**Task 1: Take the first 20 genes from airway. Map ENSEMBL IDs → gene symbols.**

Retrieve gene descriptions

```r
library(airway)
data("airway")

library(org.Hs.eg.db)
library(AnnotationDbi)
```

Get first 20 ENSEMBL IDs from airway

```
ids20 <- rownames(airway)[1:20]
```

Map ENSEMBL → Gene Symbol

```
symbols <- mapIds(org.Hs.eg.db,
                  keys = ids20,
                  keytype = "ENSEMBL",
                  column = "SYMBOL")
```

Map ENSEMBL → Full Gene Name

```
descriptions <- mapIds(org.Hs.eg.db,
                       keys = ids20,
                       keytype = "ENSEMBL",
                       column = "GENENAME")
```

Combine into a data frame

```
annotated20 <- data.frame(ENSEMBL_ID = ids20,
                          Symbol = symbols,
                          Description = descriptions)

head(annotated20)
```

**Task 2: Subset airway to treated samples only. Select the first 5 genes.**

Annotate them with symbols + full names

```
# Subset treated samples
treated <- airway[, airway$dex == "trt"]

# Get first 5 ENSEMBL IDs from treated dataset
ids5 <- rownames(treated)[1:5]

# Map ENSEMBL → Symbol
symbols5 <- mapIds(org.Hs.eg.db,
                   keys = ids5,
```

```r
                   keytype = "ENSEMBL",
                   column = "SYMBOL")

# Map ENSEMBL → Gene Name
names5 <- mapIds(org.Hs.eg.db,
                 keys = ids5,
                 keytype = "ENSEMBL",
                 column = "GENENAME")

# Combine results
annotated5 <- data.frame(ENSEMBL_ID = ids5,
                         Symbol = symbols5,
                         Full_Name = names5)

annotated5
```

# Chapter 6

# Module 4

## 6.1 Lecture

### 6.1.1 4A

### 6.1.2 4B

## 6.2 Lab 4A

**Normalization and Preprocessing**

Install and load DESeq2

```
BiocManager::install("DESeq2")
library(DESeq2) # Load DESeq2
```

Create a DESeq2 dataset object

```
dds <- DESeqDataSet(airway, design = ~ dex)
```

Run the DESeq pipeline

```
dds <- DESeq(dds)
```

Raw and normalized counts

```r
# Normalized
norm_counts <- counts(dds, normalized=TRUE)
head(norm_counts)

# Raw
norm_counts <- counts(dds, normalized=FALSE)
head(norm_counts)
```

### How to use VST in DESeq2

Create a new transformed dataset

```r
vsd <- vst(dds, blind=FALSE) # takes into account the experimental design
# Us the transformed expression matrix
assay(vsd)[1:5, 1:5]
```

### Sample Clustering (Dendrogram/Heatmap)

```r
# install pheatmap
BiocManager::install("pheatmap", force = T)

# Load pheatmap
library(pheatmap)

# Calculate sample-to-sample distances
sampleDists <- dist(t(assay(vsd)))

# Convert distances into a matrix
sampleDistMatrix <- as.matrix(sampleDists)

# Heatmap of distances between samples
pheatmap(sampleDistMatrix,
         annotation_col = as.data.frame(colData(vsd)[, "dex", drop=FALSE]),
         main = "Sample-to-sample distances")

# PCA with DESeq2 VST Data colored by treatment (dex)
plotPCA(vsd, intgroup="dex")
```

## 6.3 Lab 4A Tasks

Pretasks

```r
# Load DESeq2
library(DESeq2)
library(airway)
data("airway")

# Create a DESeq2 dataset object
dds <- DESeqDataSet(airway, design = ~ dex)

# Run the DESeq pipeline
dds <- DESeq(dds)

# creates a new transformed dataset
vsd <- vst(dds, blind=FALSE) # takes into account the experimental design
```

### Task 1: Variance check (top variable genes)

```r
# Calculate variance for each gene
geneVars <- rowVars(assay(vsd))

# Top 10 most variable genes
top10 <- head(order(geneVars, decreasing=TRUE), 10)
rownames(vsd)[top10]
```

### Task 2: Custom PCA on top 500 genes

```r
library(ggplot2)

# Select top 500 variable genes
top500 <- head(order(geneVars, decreasing=TRUE), 500)
mat <- assay(vsd)[top500, ]

# Run PCA
pca <- prcomp(t(mat), scale. = TRUE)

# Create data frame for plotting
pca_df <- as.data.frame(pca$x)
pca_df$dex <- colData(vsd)$dex

# PCA plot
ggplot(pca_df, aes(x=PC1, y=PC2, color=dex)) +
  geom_point(size=3) +
  labs(title="PCA on Top 500 Variable Genes")
```

### Task 3: Challenge: Add cell line

```
pca_df$cell <- colData(vsd)$cell

ggplot(pca_df, aes(x=PC1, y=PC2, color=dex, shape=cell)) +
  geom_point(size=3) +
  labs(title="PCA: Treatment (color) vs Cell Line (shape)")
```

## 6.4 Lab 4B

### Bioconductor Packages and Data sets

Install airway package

```
BiocManager::install("airway")
library(airway)
library(DESeq2)
```

DGE Analysis

```
dds <- DESeqDataSet(airway, design = ~ dex)
dds <- DESeq(dds)
```

Extract DGE results

```
res <- results(dds)
head(res)
```

Filter significant genes

```
sig_res <- res[which(res$padj < 0.05), ]
head(sig_res)
summary(sig_res)
```

### Visualization of DGE Results

```r
# Load required packages
library(DESeq2)
library(ggplot2)

# Run DESeq2 analysis
dds <- DESeqDataSet(airway, design = ~ dex)
dds <- DESeq(dds)
res <- results(dds)

# Convert to data frame for plotting
res_df <- as.data.frame(res)

# Remove rows with missing p-values or fold change (optional but helps avoid warnings)
res_df <- na.omit(res_df)

# Create a new column indicating regulation direction
res_df$Regulation <- "Not significant"
res_df$Regulation[res_df$log2FoldChange > 1 & res_df$padj < 0.05] <- "Upregulated"
res_df$Regulation[res_df$log2FoldChange < -1 & res_df$padj < 0.05] <- "Downregulated"
```

*Volcano plot*

```r
ggplot(res_df, aes(x = log2FoldChange, y = -log10(padj), color = Regulation)) +
  geom_point(alpha = 0.6, size = 1.5) +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed", color = "gray40") +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "gray40") +
  scale_color_manual(values = c("Upregulated" = "red",
                                "Downregulated" = "blue",
                                "Not significant" = "gray70")) +
  labs(title = "Volcano Plot: Treated vs Untreated",
       x = "log2 Fold Change",
       y = "-log10(Adjusted p-value)",
       color = "Regulation") +
  theme_minimal()
```

## 6.5 Mini Project

### Task 1: Load Required Libraries and Dataset

Load packages

```r
library(DESeq2)
library(airway)
library(ggplot2)
library(org.Hs.eg.db)
library(AnnotationDbi)
library(pheatmap)


data("airway"). # Load data

# Inspect dataset
airway
```

### Task 2: Preprocessing and Normalization

```r
# Create DESeq2 object
dds <- DESeqDataSet(airway, design = ~ dex)

# Run the DESeq2 pipeline (includes normalization)
dds <- DESeq(dds)

# Variance Stabilizing Transformation
vsd <- vst(dds, blind = FALSE)
```

### Task 3: Differential Expression Analysis

```r
# Extract DE results
res <- results(dds)
summary(res)

# Filter significant genes (adjusted p < 0.05)
sig_res <- res[which(res$padj < 0.05), ]

# Order by log2 fold change
sig_res <- sig_res[order(sig_res$log2FoldChange, decreasing = TRUE), ]

# Show top results
head(sig_res)
```

### Task 4: Annotate Significant Genes

```
# Take top 20 genes
top_genes <- rownames(sig_res)[1:20]

# Map to gene symbols and names
symbols <- mapIds(org.Hs.eg.db,
                  keys = top_genes,
                  keytype = "ENSEMBL",
                  column = "SYMBOL")
names <- mapIds(org.Hs.eg.db,
                keys = top_genes,
                keytype = "ENSEMBL",
                column = "GENENAME")

annotated <- data.frame(ENSEMBL = top_genes,
                        Symbol = symbols,
                        Name = names)
head(annotated)
```

### Task 5: Visualization

*A: PCA*

```
plotPCA(vsd, intgroup = "dex")
```

*B: Volcano plot*

```
res_df <- as.data.frame(res)
res_df <- na.omit(res_df)
res_df$Regulation <- "Not significant"
res_df$Regulation[res_df$log2FoldChange > 1 & res_df$padj < 0.05] <- "Upregulated"
res_df$Regulation[res_df$log2FoldChange < -1 & res_df$padj < 0.05] <- "Downregulated"

ggplot(res_df, aes(x = log2FoldChange, y = -log10(padj), color = Regulation)) +
  geom_point(alpha = 0.6, size = 1.5) +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed", color = "gray40") +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "gray40") +
  scale_color_manual(values = c("Upregulated" = "red",
                                "Downregulated" = "blue",
                                "Not significant" = "gray70")) +
  labs(title = "Volcano Plot: Treated vs Untreated",
       x = "log2 Fold Change", y = "-log10(Adjusted p-value)") +
  theme_minimal()
```

*C: Heatmap of Top Variable Genes*

```
topVarGenes <- head(order(rowVars(assay(vsd)), decreasing = TRUE), 30)
pheatmap(assay(vsd)[topVarGenes, ],
         scale = "row",
         annotation_col = as.data.frame(colData(vsd)[, "dex", drop=FALSE]),
         main = "Top 30 Variable Genes")
```

*D: Optional Challenge (Barplot of top 10 upregulated genes)*

```
top_up <- head(sig_res[order(sig_res$log2FoldChange, decreasing = TRUE), ], 10)
barplot(top_up$log2FoldChange, names.arg = rownames(top_up),
        las = 2, col = "tomato", main = "Top 10 Upregulated Genes",
        ylab = "log2 Fold Change")
```