Solutions

Research with Computational Biology (ReComBio)

August 26, 2024

Part 1: Introduction to R and the R Syntax

Chapter 3: Data visualization

Load libraries

```
library(ggplot2)
library(maditr)
library(dplyr)
```

Load dataframe

```
r2_gse62564_GSVA_Metadata <- readRDS("../../ReComBio Scientific/ReComBio Book English/recombio bookdown/
```

Make Variables of Numeric Type

```
r2_gse62564_GSVA_Metadata <- r2_gse62564_GSVA_Metadata %>%
   mutate_at("HALLMARK_HYPOXIA", as.numeric) %>%
   mutate_at("HALLMARK_INFLAMMATORY_RESPONSE", as.numeric) %>%
   mutate_at("ADRN_Norm_vs_Hypo_Up_554.txt", as.numeric) %>%
   mutate_at("ADRN_Norm_vs_Hypo_Down_635.txt", as.numeric)
```

Question 1

Use R plot functions to visualize the correlation between Hallmark Hypoxia and Hallmark Inflammatory Response

Solution: HALLMARK_INFLAMMATORY_RESPONSE vs. HALLMARK_HYPOXIA

```
qplot(HALLMARK_INFLAMMATORY_RESPONSE, HALLMARK_HYPOXIA,
      data = r2_gse62564_GSVA_Metadata,
      #colour=quantile,
      ylab = "HALLMARK_HYPOXIA",
      xlab = "HALLMARK_INFLAMMATORY_RESPONSE")
```

Question 2

Use R plot functions to visualize the correlation between Hallmark Hypoxia and Hallmark Inflammatory Response

Solution: HALLMARK_INFLAMMATORY_RESPONSE vs. ADRN_Norm_vs_Hypo_Up_554.txt

```
qplot(HALLMARK_INFLAMMATORY_RESPONSE, ADRN_Norm_vs_Hypo_Up_554.txt,
      data = r2_gse62564_GSVA_Metadata,
      # colour=quantile,
      xlab = "HALLMARK_INFLAMMATORY_RESPONSE",
      ylab = "ADRN_Norm_vs_Hypo_Up_554.txt")
```

Question 3

Use R plot functions to visualize the correlation between Hallmark Hypoxia and Hallmark Inflammatory Response

Solution: HALLMARK_INFLAMMATORY_RESPONSE vs. ADRN_Norm_vs_Hypo_Down_635.txt

```
qplot(HALLMARK_INFLAMMATORY_RESPONSE, ADRN_Norm_vs_Hypo_Down_635.txt,
      data = r2_gse62564_GSVA_Metadata,
      # colour=quantile,
      ylab = "ADRN_Norm_vs_Hypo_Down_635.txt",
      xlab = "HALLMARK_INFLAMMATORY_RESPONSE")
```

Question 4

From the analysis of questions 1-3, choose the correct option:

- ☐ HALLMARK_INFLAMMATORY_RESPONSE and HALLMARK_HYPOXIA have a positive correlation because hypoxia is always beneficial in the tumor microenvironment

- ☐ HALLMARK_INFLAMMATORY_RESPONSE and ADRN_Norm_vs_Hypo_Up_554.txt have a positive correlation because hypoxia upregulation in this case, is beneficial in the tumor microenvironment
- ☐ HALLMARK_INFLAMMATORY_RESPONSE and ADRN_Norm_vs_Hypo_Up_554.txt have a positive correlation because hypoxia upregulation has a negative impact on survival
- ☐ HALLMARK_INFLAMMATORY_RESPONSE and ADRN_Norm_vs_Hypo_Down_635.txt have a negative correlation because hypoxia upregulation in this case, is not beneficial in the tumor microenvironment

Question 5

Which gene expression group has worse survival probability?

- ☐ High HIF1A expression
- ☐ Low HIF1A expression

Question 6

Which phenotype score group has worse survival probability?

- ☐ High Hallmark Hypoxia
- ☐ Low Hallmark Hypoxia

Question 7

Which phenotype score group has worse survival probability?

- ☐ High Hallmark Inflammatory Response
- ☐ Low Hallmark Inflammatory Response

Question 8

Plot the survival curve of the MYCN status variable. Which MYCN status has worse survival outcome?

- ☐ Individuals with MYCN amplification
- ☐ Individuals without MYCN amplification
- ☐ The MYCN group with unknown MYCN status

Question 9

Plot the survival curve of the INSS stage variable. Which INSS stage has worse survival outcome?

- ☐ INSS Stage I
- ☐ INSS Stage II
- ☐ INSS Stage III
- ☐ INSS Stage IV
- ☐ INSS Stage IV A

Question 10

Mark TRUE or FALSE.

- ☐ The higher the age at diagnosis the greater the HIF1A expression difference between HR and non-HR
- ☐ The lower the age at diagnosis the greater the HIF1A expression difference between HR and non-HR

Part 2: Machine Learning

Chapter 1: Classification

P2 Question 1

How can you predict what label a new patient that was sequenced using the UCSC nanopore technology will have using the logistic regression model constructed in R?

☐ I can build a vector to provide as input to R so that the model can use the parameters it calculated to make the prediction

☐ If I have more than 2 patients to predict, I can build a dataframe to input to R

☐ It is not possible to know if a patient has high risk neuroblastoma disease without doing a FISH (Fluorescence *in situ* hybridization) and understanding if the person has MYCN amplification

P2 Question 2

How can you know if a model represents a good indicator for the high risk status of a patient?

☐ If the model has high accuracy

☐ If the model has medium accuracy

☐ If the model has low accuracy

P2 Question 3

Write a command to input information about the predictor variables of a patient into R. To come up with a solution, please discuss the strategy to solve this problem in groups in breakout rooms. Please use 10-15 min to discuss a solution.

- Hint: Use notebook "Classification Using a Logistic Regression Model"

- Please describe a solution in words (**human language**)

P2 Question 4

Write a command to predict the high risk label for the patient based in the patient's gene expression pattern. To come up with a solution, please discuss the strategy to solve this problem in groups in breakout rooms. Please use 10-15 min to discuss a solution.

- Hint 1: Which chunks in the notebook help to explain a possible solution?
- Hint 2: Look at the functions that can possibly present a solution?
- Please describe the solution in pseudocode (**human language + computer language merged**)

P2 Question 5

Write a command to predict the high risk label for the patient based in the patient's gene expression pattern. To come up with a solution, please discuss the strategy to solve this problem in groups. Please use 10-15 min to discuss a solution.

- Hint 1: Which data structures can you use to input the data into the R environment? A dataframe? A vector? A character?
- Please present the solution in a command or algorithm (**computer language**)

Part 3: Terminal

Chapter 1: DNA Sequencing

P3 Question 1

According to Shendure, how many and what are the names of the DNA sequencing technologies?

P3 Question 2

According to Shendure, what is the main clinical application the DNA sequencing technologies?

P3 Question 3

According to Shendure, please identify at least one chemical event in the machanics of DNA sequencing.

P3 Question 4

According to Shendure, please identify at least one significant event in history of DNA sequencing.