

Master Real-world Bioinformatics analysis in R

Ming Tommy Tang

2024-07-28

Contents

1 Preface	7
2 Introduction	9
2.1 Meet your Instructor	9
2.2 Leveraging our online community	11
2.3 Install R and R studio	11
2.4 The role of programming in Biology	11
3 Introduction to programming	15
3.1 What is an algorithm?	15
3.2 Why are Algorithms Important in Programming?	15
3.3 A Real World Example:	16
3.4 Variables, Data Types, and Expressions	16
4 Getting started with R and RStudio	21
4.1 What is R and why is it used in Biology?	21
4.2 What is R?	21
4.3 Why is R Used in Biology?	21
4.4 Introduction to R-Studio	22
5 Introduction to R for biologists	23
5.1 Basic Data Types in R	23
5.2 The key concept of R: Vectors	27
5.3 Subsetting and Indexing	32

5.4	Understanding and Manipulating Matrices	33
5.5	Essential Functions in R	39
5.6	Functions: Organizing Your Code for Reusability	41
5.7	Common Mistakes to avoid	50
6	Controlling the flow of our programs	53
6.1	Boolean Operators	53
6.2	Conditional statements (<code>if</code> , <code>else</code>)	56
6.3	Loops	59
6.4	Gene Expression Annotation using Loops and Control Structures	64
6.5	Let's solve a Challenge	67
6.6	Solution	68
6.7	Section complete	70
7	Going more in-depth with R	73
7.1	Handling Missing Values	73
7.2	Introduction to Statistical Tests and P-Values	76
7.3	Understanding R Packages	79
7.4	Exploring Functions in Different Packages: Avoiding Collisions and Accessing Them	83
7.5	Writing Custom Scripts	85
7.6	Data I/O	87
7.7	Best Practices in Modular Programming and Project Management	91
7.8	Section complete	94
8	Fundamental Data Structures in R	97
8.1	Named Vectors	97
8.2	Lists	100
8.3	Dataframes	103
8.4	How to Represent Categorical Data in R? Understanding Factors	105
8.5	Section Complete	109

CONTENTS	5
9 Introduction to the tidyverse ecosystem	111
9.1 What is the Tidyverse?	111
9.2 Key Tidyverse Packages	112
9.3 Tibble and Readr - Modern Data Structures in R	117
9.4 The tidy data format	123
9.5 Introducing dplyr: Your Data Wrangling Toolkit	125
9.6 stringr: your essential toolkit to manipulate strings	131
9.7 purrr: ditch your for loops	132
9.8 Tidying metadata from GEO.	143
9.9 Section Complete	153
10 Data visualization with ggplot2	155
10.1 Creating Scatter Plots	156
10.2 Understanding Distributions with Histograms	162
10.3 Visualizing Data Distribution with Boxplots and Violin Plots . .	165
10.4 Creating Bar Plots to Visualize Median EPCAM Levels by Cancer Type	173
11 Introduction to BioConductor	179
11.1 Introduction to BiocManager	179
11.2 Working with Genomic Coordinates and Regions in R	181
11.3 Exploring CpG Islands and Shores in Genomic Data	198
11.4 Real-World Applications: ChIP-seq	202
11.5 Analyzing and Visualizing Genomic Data	202
11.6 Real-World Example - TCGA Analysis	210
11.7 Section completed	229
12 Final Project: Analyzing RNAseq Data from GEO	231
12.1 Final Project Overview	231
12.2 How to pre-process RNAseq data	232
12.3 Download and subset Count Matrix	234
12.4 Calculate the total exon length per gene	238
12.5 Normalizing Raw Counts to Transcripts per Million (TPM) . .	245

12.6 Analyzing Gene Expression Data Using t-Tests	247
12.7 Analyzing Gene Expression Data with ggplot2	250
12.8 Correcting for Multiple Comparisons in Statistical Analysis . . .	256
12.9 Analyzing Differential Gene Expression with DESeq2	262
12.10 Principal Component Analysis (PCA) using DESeq2	272
12.11 Creating a Perfect Heatmap	278
12.12 Pathway Analysis Using Over-Representation and Gene Set Enrichment Analysis	282
12.13 Congratulations for successfully completing this final project! . .	301

Chapter 1

Preface



Chapter 2

Introduction

2.1 Meet your Instructor

Hello, Welcome! all the students!

this is Tommy, your instructor for this course. Congratulations on signing up for this course. I am sure you will learn a lot. I created the first draft of this course during the holidays and with the help of ChatGPT, I was able to polish it quite a bit. This is not a perfect course, your feedback is greatly appreciated!



A little bit more about me. With over a decade of experience in computational biology, I specialize in genomics, epigenomics, and (single-cell) transcriptomics data analysis. I have taken on pivotal roles in various cancer research projects, notably contributing to the NCI's Cancer Moonshot initiative at the Dana-Farber Cancer Institute.

I am now the Director of Computational Biology at Immunitas Therapeutics, we employ machine-learning techniques to investigate immune cells in human tumors by analyzing single-cell RNAseq, single-cell TCRseq, and spatial transcriptome data. Our goal is to develop novel therapeutics for cancer patients.

I am a self-trained computational biologist. I fully understand how challenging it is to learn computational biology from scratch. That's why beyond my professional work, I am passionate about promoting open science and improving bioinformatics education to equip biologists with computational skills. More about me can be found at my website <https://divingintogeneticsandgenomics.com/>.

Enjoy this course! Let's go!

2.2 Leveraging our online community

2.3 Install R and R studio

2.4 The role of programming in Biology

In this lesson, we will discover how programming languages empower biologists to unravel the mysteries of life, make critical discoveries, and automate complex tasks. While we won't be diving into intricate technicalities, we'll explore the fundamental concepts that will serve as the foundation for your exploration into the world of computational biology and bioinformatics.

2.4.1 What are programming languages?

Programming languages are a set of rules and instructions used by humans to communicate with computers. They serve as a bridge between human thought and machine execution, allowing us to convey complex tasks and algorithms to computers in a way they can understand and execute.

- **Communication Tool:** Programming languages are a means of communication between humans and computers. They provide a structured and understandable way for programmers to convey their intentions to the computer.
- **Instructions:** In a programming language, you write instructions or commands that specify what the computer should do. These instructions can range from simple tasks like adding numbers to complex processes like data analysis or simulation.
- **Syntax:** Programming languages have their syntax or grammar rules that programmers must follow. Syntax defines how instructions should be structured, including the order of words, punctuation, and formatting. Following the correct syntax is crucial for the computer to interpret the code correctly.

- Abstraction: Programming languages provide a level of abstraction. They allow us to work at a higher level of understanding, dealing with concepts like variables, functions, and data structures, without needing to worry about the low-level details of how the computer processes these instructions.
- Interpreter or Compiler: To execute code written in a programming language, you need either an interpreter or a compiler. An interpreter reads and executes code line by line, while a compiler translates the entire code into machine code before execution.

2.4.2 Why programming is important in biology?

Ever wondered why programming is such a big deal in biology? It's because it gives biologists the superpower to handle mountains of data, uncover hidden patterns, and automate those repetitive lab tasks. Read this paper: All Biology is computational Biology. So, what makes programming tick in the world of biology? Let's delve deeper:

- Efficient Data Handling: In biology, we deal with enormous volumes of data, from DNA sequences to ecological observations. Programming allows us to efficiently manage and process this data. By automating data collection and analysis, we save time and minimize errors, ensuring that our research is based on accurate and comprehensive information.
- Complex Analysis: Biological research often involves intricate analyses, such as genetic sequence comparisons, statistical modeling, and simulations. Programming languages provide the tools to perform these complex tasks with precision and speed. These analyses can unveil hidden patterns, relationships, and insights that would be challenging or impossible to discover manually.
- Reproducibility: Reproducibility is a cornerstone of scientific research. Programming ensures that experiments and analyses can be replicated precisely. By sharing code, scientists can validate each other's findings and build upon existing research, fostering collaboration and advancing the field collectively.
- Automation: Many biological experiments and processes are repetitive. Programming enables the automation of these tasks, freeing researchers from mundane and time-consuming work. This automation not only improves efficiency but also reduces the risk of human error.
- Visualization: Visualization is a crucial aspect of biology, allowing researchers to represent complex data in understandable ways. Programming languages provide libraries and tools to create stunning visualizations, aiding in the interpretation and communication of research findings.

2.4.3 What are the most used programming languages within biology?

2.4.3.1 Python: The Swiss Army Knife of Biology

Python is the go-to programming language in the field of biology, and for good reasons. It's known for its simplicity and readability, making it an ideal choice for biologists who may not have extensive programming backgrounds. Python offers a vast ecosystem of libraries and tools tailored for bioinformatics, data analysis, and scientific computing.

- Bioinformatics: Python excels in bioinformatics, with libraries like Biopython, which provides tools for sequence analysis, structural biology, and more. Biologists use Python to parse and manipulate DNA, RNA, and protein sequences effortlessly.
- Data Analysis: Python's libraries, such as NumPy, pandas, and matplotlib, make data analysis and visualization a breeze. Researchers can explore and visualize complex biological data, from gene expression profiles to ecological datasets.
- Machine Learning: Python's machine learning libraries like scikit-learn enable biologists to build predictive models for disease classification, drug discovery, and more. It's a valuable tool for harnessing the power of data.

2.4.3.2 R: The Statistical Powerhouse

R is another programming language highly favored by biologists, particularly for statistical analysis and data visualization. It's renowned for its statistical packages and robust graphing capabilities, making it an indispensable tool for researchers dealing with biological data.

- Statistical Analysis: R boasts an extensive collection of statistical packages and libraries, including Bioconductor, designed specifically for biological data analysis. Biologists rely on R for hypothesis testing, regression analysis, and experimental design.
- Data Visualization: With libraries like ggplot2, R allows biologists to create intricate and publication-quality visualizations. It's instrumental in presenting research findings effectively.

2.4.3.3 Julia: The Rising Star

Julia is an emerging programming language that has garnered attention in the scientific community, including biology. It's prized for its exceptional perfor-

mance and versatility, making it suitable for computationally intensive tasks in genomics, proteomics, and more.

- Performance: Julia's speed rivals low-level languages like C and Fortran, making it a compelling choice for high-performance computing in biology. It's used for tasks like simulating biological systems and analyzing large datasets.
- Ease of Use: Julia's syntax is intuitive and easy to learn, appealing to both programmers and scientists. Its interactive environment fosters quick experimentation.

2.4.3.4 Other Programming Languages:

While Python, R, and Julia are the prominent choices, other programming languages find their niche in specific areas of biology:

- Perl: Historically used in bioinformatics for tasks like text processing and sequence analysis.
- Java: Commonly employed in developing bioinformatics software and applications.
- C/C++: Reserved for computationally intensive tasks where speed is critical, such as molecular dynamics simulations.

Chapter 3

Introduction to programming

3.1 What is an algorithm?

Welcome to the first lesson of our course! Today, we're going to explore two fundamental concepts in programming: algorithms and flowcharts. Don't worry if these terms sound a bit technical; we'll break them down into simple ideas.

Imagine you're following a recipe to bake a cake. The recipe gives you step-by-step instructions on what to do, right? An algorithm is similar. It's a set of instructions or steps designed to perform a specific task. In programming, we use algorithms to tell the computer exactly what we want it to do.

3.2 Why are Algorithms Important in Programming?

- Clarity: Algorithms serve as an essential tool for strategizing and planning our code. They provide us with a clear roadmap of the steps we need to follow before we start writing the actual code. This pre-coding stage can help us avoid potential issues and ensure that our solutions are well thought out.
- Problem-Solving: Algorithms play a crucial role in problem-solving. They allow us to break down complex tasks into a series of simpler steps, making them easier to manage and understand. By using algorithms, we can tackle large problems by solving each small part one at a time, thus making the overall problem-solving process more efficient and manageable.

- Efficiency: A well-designed algorithm can save significant time and resources. It can help us optimize our code to perform tasks in the fastest and most efficient way possible. By improving the efficiency of our code, we can ensure that it runs smoothly and quickly, thus enhancing the performance of our software or application.

3.3 A Real World Example:

In this algorithm, we'll learn how to find the length of a DNA sequence. Knowing the length of a DNA sequence is important for various biological analyses.

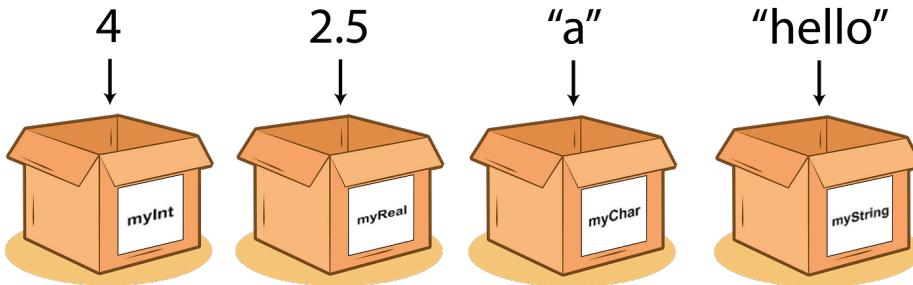
1. Start: Begin the algorithm.
2. Input DNA Sequence: Ask the user to input a DNA sequence. For example, the sequence could be a string of letters like "ATCGATGCTA."
3. Initialize Length: Set a variable called "Length" to 0. This variable will be used to keep track of the length of the sequence.
4. For Each Base in DNA Sequence:
 - Start a loop that goes through each base in the DNA sequence, one by one.
 - For the input "ATCGATGCTA," the loop will start with "A."
5. Increase Length by 1: For each base you encounter in the sequence, add 1 to the "Length" variable. This counts the number of bases in the sequence.
6. Repeat: Continue the loop until you have processed all the bases in the DNA sequence.
7. Output Length: Once the loop is finished, the "Length" variable will contain the length of the DNA sequence. Display this value as the output.
8. End: End the algorithm.

3.4 Variables, Data Types, and Expressions

In this lesson, we're going to explore some fundamental concepts of programming that are crucial in understanding how to write code for computational biology. We'll focus on three key ideas: variables, data types, and expressions. Think of these as the building blocks for creating a language that your computer can understand and use to solve biological problems.

3.4.1 Variables

A variable is quite similar to a labeled box. It's a container where you can store information or data. Once you have this box (variable), you can do various things with it: you can put things into it, take things out of it, or even change what's inside it to something else. This flexibility is immensely useful when dealing with large volumes of data or when you're accessing often to a specific piece of that data, a common occurrence in the field of computational biology.



```
# Step 1: Define a variable and store the DNA sequence
```

```
gene_sequence = "ATCGAGCTAGCTGCTAGCTAGCTAGCT"
```

```
# Step 2: Print the stored DNA sequence
```

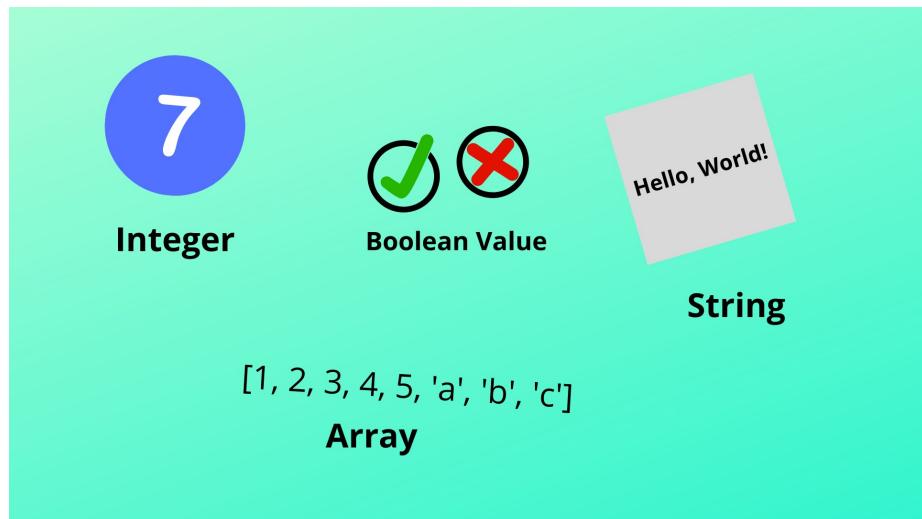
```
print(gene_sequence)
```

```
## [1] "ATCGAGCTAGCTGCTAGCTAGCTAGCT"
```

For instance, consider a situation where you're working with gene sequences. You can use a variable like 'gene_sequence' to store a particular gene's sequence. Later, you can access this stored sequence, manipulate it, or compare it with other sequences as needed in your computational biology tasks.

3.4.2 Data Types

In programming, similar to the real world, we come across various data types that help us to structure and understand information. These data types include numbers, which could be whole numbers or numbers with decimal points; text, often referred to as strings in programming terms; booleans, representing True or False values; and lists, which are utilized to hold collections of items.



Each of these data types serve a unique purpose and are used in different contexts, playing an integral role in how we write and interpret code.

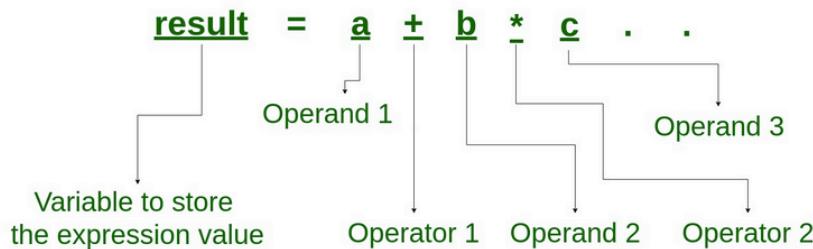
- Numbers: In programming, numbers can be whole or decimal. They represent measurements or quantities in computational biology, like DNA length or chemical concentration.
- Text (Strings): Strings are sequences of characters, used for gene names, DNA sequences, or any textual information in computational biology.
- Lists/Arrays: Lists group items together, useful in handling multiple genes, proteins, or biological elements. They can store different gene names or DNA sequences.
- Booleans: Booleans represent true or false values and are used to express conditions or decisions based on a yes/no scenario, like determining if a specific gene is present.

Understanding data types is crucial because it helps you work with biological data accurately and efficiently. When you're coding in computational biology, knowing whether you're dealing with numbers, text, or lists allows you to use the right tools and operations for each type of data.

For instance, if you need to perform calculations on gene lengths (numbers), you'll use mathematical operations. If you want to search for a specific gene name (text), you'll use string manipulation techniques. And when you're handling multiple genes (lists), you'll employ list-related functions to process them effectively.

3.4.3 Expressions

What is an Expression?



In the world of programming, expressions are like simple puzzles or equations that you can use to do things with data. These expressions are created by putting together a few essential elements:

- Values: Think of these as numbers, like 2 or 3, that you want to work with. In computational biology, these could be things like the length of a gene or the number of amino acids in a protein.
- Variables: These are like labeled containers where you can store information. For example, you might have a variable called ‘gene_length’ that holds the length of a specific gene sequence.
- Operators: Operators are special symbols like `+`, `-`, `*`, or `/` that you use to perform operations on values and variables. They tell the computer what kind of action to take.

Now, let's dive into some beginner-level examples in computational biology:

3.4.4 Example: Finding the average number of bases appearances in a set of DNA strings

Let's suppose we want to calculate the average number of bases in a DNA string. Let's assume we already processed the DNA string and we know the counts for each one.

Given counts of each base:

```
count_A = 120
count_T = 90
count_C = 80
count_G = 110
```

Calculate the total number of bases

```
total_bases = count_A + count_T + count_C + count_G
```

Calculate the average number of bases

```
average_bases = total_bases / 4
```

Print the result

```
print(average_bases)
```

```
## [1] 100
```

In this code:

1. We start by declaring the counts of each base using variables (number_a, number_t, number_c, number_g).
2. We calculate the total number of bases by summing up the counts.
3. Then, we compute the average number of bases by dividing the total by the number of different bases (4 in this case, representing A, T, C, and G).
4. Finally, we print the result, which gives us the average number of bases in the DNA string.

Chapter 4

Getting started with R and RStudio

4.1 What is R and why is it used in Biology?

In this lesson, we'll dive into the world of R, a powerful programming language and environment used extensively in the field of biology for data analysis and visualization. We'll explore what R is, why it's so useful, and how it can be a valuable tool for biologists.

4.2 What is R?

R is a free and open-source statistical programming language specialized for statistical analysis and data visualization. R can supercharge your data analysis and enable you to go beyond the Excel spreadsheet.

4.3 Why is R Used in Biology?

Biologists use R for a variety of reasons:

- Data Analysis: R provides a wide range of tools and packages for data analysis, making it easier to handle and analyze complex biological datasets. Whether you're studying gene expression, population genetics, or ecological data, R can help you make sense of your data. You can take advantage of a lot of ready-to-use Bioconductor packages for almost any data type in biology. We will learn several essential bioconductor packages in the future sections.

- Statistics: In biology, statistical analysis is crucial for drawing meaningful conclusions from experiments and observations. R offers an extensive collection of statistical functions and libraries, allowing biologists to perform advanced statistical tests and modeling.
- Data Visualization: R excels at creating stunning visualizations of biological data. You can generate graphs, charts, and plots to visualize trends, relationships, and patterns in your data. Visualization is essential for communicating your findings effectively. You can make publication-ready figures with packages such as ggplot2 which we will cover in depth too.
- Reproducibility: R promotes reproducibility in scientific research. You can write scripts or programs to automate your analyses, ensuring that others can replicate your work and verify your results. Tools such as Rmarkdown make the analysis reproducible in Rstudio or Jupyter Notebook.
- Community Support: R has a vibrant and active user community, which means you'll find plenty of resources, tutorials, and forums where you can seek help and share your knowledge. You can visit helpful communities such as the Posit community and Bioconductor support site.
- Integration: R can be integrated with other tools and languages, such as Python and SQL, making it flexible for various research needs.

In conclusion, R is a versatile and essential tool for biologists. It empowers researchers to handle, analyze, and visualize data effectively, leading to a deeper understanding of biological phenomena.

“Like learning anything, it takes effort to master R. However, if you take the effort to learn the basics and relevant bioinformatics packages, you can conduct your analysis 100 times faster than the point-and-click tools. The added benefit is that you can make your analysis more reproducible.” – Tommy

4.4 Introduction to R-Studio

In this chapter, we will introduce you to RStudio, a powerful integrated development environment (IDE) for the R programming language. RStudio provides a user-friendly interface for writing, running, and managing R code. We will explore the various panes, functions, and features of RStudio to help you get started on your journey with R programming.

Chapter 5

Introduction to R for biologists

5.1 Basic Data Types in R

5.1.1 Before Starting

When you're working with R, it's crucial to name your variables properly. Here are some simple rules to follow:

1. Allowed Characters: Variable names can include letters, numbers, underscores (_), and periods (.).
2. Starting Characters: A variable name must start with a letter or a period. It can't start with a number.
3. Case Sensitivity: R treats variable names as case-sensitive. For example, myvar is different from MyVar.
4. Dots in Names: While dots are allowed in variable names, it's best to avoid them except at the beginning.
5. Avoiding Conflicts: Don't use names that match existing functions in R, like mean or c.

5.1.1.1 Examples:

Valid:

- myvar

- my.var
- var1
- var_1

Invalid:

- 1var (can't start with a number)
- *temp* (*can't start with a*)
- c (matches existing function)
- my-var (hyphens aren't allowed)
- my var (spaces aren't allowed)

5.1.1.2 Tips for Naming:

- Meaningful Names: Choose descriptive names like patient_data instead of generic ones like x, y, or z.
- Short and Clear: Keep your names short but make sure they clearly represent what the variable contains.
- Naming Style: You can use camelCase or underscores_between_words for multi-word names. Make sure to stick to one style and be consistent.

Case Consistency: Decide whether you want all lowercase names or CapWords (also known as PascalCase), and stick to it throughout your code.

By following these naming conventions, your code will be easier to understand, and you'll avoid unexpected errors. Consistency is key when naming variables across your R scripts.

In R, data can be classified into several fundamental types, each serving specific purposes in data analysis and manipulation. Understanding these data types is crucial for effective data handling. Let's explore the primary data types in R:

5.1.2 1. Numeric

Numeric data represents continuous numerical values. These can be integers or real numbers (floating-point). Numeric data is used for mathematical calculations and statistical analysis.

Example:

```
# Numeric data
age <- 28
height <- 1.75
```

The `<-` operator and the `=` operator in R are both used for assignment but have some key differences.

The `<-` operator is the standard assignment operator in R. It assigns values to objects.

- The arrow can be read as “gets”. So age gets the value of 28.
- All R objects should be created using the `<-` operator.

The `=` Operator can also be used for assignments.

```
age = 28
```

- This also assigns 28 to age.
- The `=` should be read as “is equal to”. This is mainly used to specify function arguments.

So in a function call like:

```
plot(x = mydata)
```

We are specifying that the `x` argument is equal to `mydata`.

In summary, `<-` is the operator you should use when creating R objects, while `=` is reserved for function arguments. For assignments, both `<-` and `=` can be used. If you want to read more differences, take a look at <https://stat.ethz.ch/R-manual/R-patched/library/base/html/assignOps.html>.

5.1.3 2. Character (String)

Character data represents text or strings of characters. You use character data for storing and manipulating text-based information, such as names, descriptions, and labels.

Example:

```
# Character data
name <- "John Doe"
city <- "New York"
```

5.1.4 3. Integer

Integer data represents whole numbers. Unlike numeric data, which includes decimal points, integer data includes only whole numbers. It's commonly used when dealing with counts or discrete quantities.

Example:

```
# Integer data
count <- 10
students <- 42
```

5.1.5 4. Logical (Boolean)

Logical data consists of two possible values: TRUE or FALSE. These values are used for binary decisions, conditions, and logical operations.

Example:

```
# Logical data
is_student <- TRUE
has_permission <- FALSE
```

5.1.6 5. Factor

Factor data represents categorical variables with predefined levels or categories. Factors are used when you have data that can be divided into distinct categories, such as “High,” “Medium,” and “Low.”

Example:

```
# Factor data
grade <- factor(c("A", "B", "C", "B", "A"))
```

5.1.7 6. Date and Time

Date and time data types are used for representing dates, times, or both. These data types are crucial when dealing with time series data or conducting temporal analysis.

Example:

```
# Date and time data
birth_date <- as.Date("1990-05-15")
timestamp <- as.POSIXct("2023-01-09 14:30:00")
```

5.1.8 7. Complex

Complex data types represent complex numbers, which have both real and imaginary parts. Complex numbers are used in advanced mathematical and engineering applications.

Example:

```
# Complex data
z <- 3 + 2i
```

5.1.9 8. Missing Values (NA)

In R, missing values are represented as NA. These values indicate the absence of data or an undefined value. Handling missing data is essential in data analysis.

Example:

```
# Missing value
missing_data <- NA
```

Understanding these data types and their characteristics is fundamental to effective data manipulation and analysis in R. Different operations and functions may behave differently depending on the data type, so being familiar with these types will help you work with data effectively.

5.2 The key concept of R: Vectors

In the world of R, vectors are the building blocks of data. In fact, everything in R is a vector. Whether you're dealing with numbers or characters, R treats them all as vectors, which are simply collections of these elements. There is no concept of a scalar in R; even a single value is considered a one-element vector. To create a vector, you'll use the `c` function, which stands for “combine” or “concatenate.”

5.2.1 Creating Numeric and Character Vectors

Let's dive right in with examples:

```
# A numeric vector
c(1, 2, 3)

## [1] 1 2 3

# Output: [1] 1 2 3

# A character vector
c("A", "T", "C", "G")

## [1] "A" "T" "C" "G"

# Output: [1] "A" "T" "C" "G"
```

Notice how `c()` is used to combine values into vectors. Even a single element, such as “A”, is a vector in R. Similarly, numeric values like 5 are considered one-element vectors.

5.2.2 Saving Vectors to Variables

Now, let's save these vectors into variables with meaningful names:

```
number_vec <- c(1, 2, 3)
dna_base <- c("A", "T", "C", "G")
gene_ids <- c("PAX6", "TP53", "MYC")
```

Remember, it's good practice to use informative variable names. Variable names cannot start with a number, so opt for names like `number_vec` and `dna_base`.

5.2.3 Vectorized Calculations

One of the powerful features of R is vectorization, where operations are automatically applied element-wise to vectors. For example:

```
number_vec + 1

## [1] 2 3 4
```

Here, we've added 1 to each element of `number_vec`. This vectorized behavior simplifies many calculations.

5.2.3.1 Understanding Indexing

You may have noticed the [1] that appears in the output. In R, it's called an index, and it indicates the position of each element in the result. For instance:

```
x <- 1:100
x

## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
## [19] 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
## [37] 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54
## [55] 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
## [73] 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90
## [91] 91 92 93 94 95 96 97 98 99 100
```

In this case, [1] denotes the first position of each element. Understanding indexing helps when working with large datasets.

5.2.4 Performing Vectorized Calculations

You can perform various calculations on vectors in R:

```
number_vec * 2

## [1] 2 4 6

# Output: [1] 2 4 6

number_vec / 2

## [1] 0.5 1.0 1.5

# Output: [1] 0.5 1.0 1.5
```

Remember, all these calculations are vectorized, making R a powerful tool for data manipulation.

5.2.5 Operations with character vectors

Character vectors in R offer a wide range of operations for text manipulation and analysis. Let's explore some of the essential operations:

5.2.5.1 Concatenation

You can combine character vectors using the `c` function:

```
new_bases <- c(dna_base, "N")
new_bases
```

```
## [1] "A" "T" "C" "G" "N"
```

This operation is useful for extending or combining character vectors.

5.2.5.2 Changing Case

Transforming the case of characters is straightforward in R. You can convert characters to uppercase or lowercase:

```
toupper(dna_base)
```

```
## [1] "A" "T" "C" "G"
```

```
tolower(dna_base)
```

```
## [1] "a" "t" "c" "g"
```

This is handy when you need consistent formatting.

5.2.6 Logical Vectors

Character vectors also allow for logical operations, which can be incredibly powerful

5.2.6.1 Finding Matches

To check which elements in a character vector meet certain criteria, use the `%in%` operator:

```
dna_base %in% c("A", "C", "T")
```

```
## [1] TRUE TRUE TRUE FALSE
```

This produces a logical vector where `TRUE` indicates a match. This is because the vector `dna_base` contains A, T, C, G and G does not match any element in the vector created by `c("A", "C", "T")`.

5.2.6.2 Saving a Logical Vector

Save the resulting logical vector to a new variable for future use:

```
logical_vec <- dna_base %in% c("A", "C", "T")
logical_vec
```

```
## [1] TRUE TRUE TRUE FALSE
```

The length of the logical vector matches the original character vector.

5.2.6.3 Negating a Logical Vector

You can negate a logical vector using the `!` operator:

```
!logical_vec
```

```
## [1] FALSE FALSE FALSE TRUE
```

Now, `TRUE` represents elements that do not match the criteria.

5.2.6.4 Subsetting with Logical Vectors

Using a logical vector for subsetting another vector is a common operation. It allows you to filter and extract specific elements:

```
# Subsetting elements that meet the criteria
dna_base[logical_vec]
```

```
## [1] "A" "T" "C"
```

```
# Subsetting elements that do not meet the criteria
dna_base[!logical_vec]
```

```
## [1] "G"
```

This powerful technique helps you extract and manipulate data efficiently based on specified conditions.

5.2.7 Conclusion

You've learned the fundamental operations that can be performed on both numeric and character vectors. These essential skills will serve as a strong foundation as you delve deeper into the world of data analysis and manipulation using R.

Remember that vectors are the building blocks of R, and they are used extensively in various data analysis tasks. Whether you're combining elements, changing case, or using logical operations to filter and extract data, you have now gained valuable insights into how vectors can be harnessed to accomplish your data analysis goals.

As you continue your journey in R programming, you'll encounter more complex data structures and operations, but the understanding of vectors will remain a cornerstone of your proficiency.

5.3 Subsetting and Indexing

In this guide, we will explain one of the fundamental concepts for dealing with vectors: indexing and slicing. Understanding how R handles these operations is crucial as it differs from some other programming languages, like Python.

5.3.1 Indexing in R

In R, unlike Python where indexing starts at 0, it begins at 1. This means that the first element in a sequence is located at position 1, the second at position 2, and so on. Let's dive into some practical examples to illustrate this concept.

```
# Let's create a vector of DNA bases
dna_base <- c("A", "T", "C", "G")

# Accessing the second element (T) using indexing
dna_base[2]

## [1] "T"

# second and fourth element
dna_base[c(2,4)]
```

```
## [1] "T" "G"
```

Here, we use square brackets [] to access elements by their position in the dna_base vector. So, `dna_base[2]` returns the second element, which is “T”.

5.3.2 Slicing in R

Slicing in R allows you to extract a chunk of elements from a vector or sequence. You can specify the start and end positions within the square brackets to define the slice.

```
# Slicing the first two elements (A and T) from the dna_base vector
dna_base[1:2]
```

```
## [1] "A" "T"
```

In this example, `dna_base[1:2]` retrieves elements from position 1 to 2, giving us “A” and “T”.

5.3.3 Negative Indexing

R also allows negative indexing to remove that element at that position:

```
# remove first element by negative indexing
remove_first_element <- dna_base[-1]
```

```
# remove second and fourth
dna_base[-c(2,4)]
```

```
## [1] "A" "C"
```

```
# Output: [1] "A" "C"
```

5.3.4 Conclusion

Remember that R starts indexing at 1, not 0, and you can use square brackets `[]` to access elements and slices within vectors and sequences. This is essential for working with data and performing various operations in R.

5.4 Understanding and Manipulating Matrices

Matrices are essential for organizing and processing data, especially when dealing with gene expression data from technologies like RNA sequencing (RNAseq). In this tutorial, we will explore how to create, manipulate, and extract information from matrices using the R programming language. We will cover topics ranging from basic matrix operations to more advanced tasks like normalization for RNAseq data analysis.

5.4.1 Creating a Matrix

To create a matrix in R, you can use the `matrix()` function. A matrix is essentially a 2-dimensional table for storing numerical data. Let's start by creating a simple matrix:

```
expression_mat <- matrix(1:12, nrow = 3, ncol = 4)
expression_mat
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

Here, `expression_mat` is a dummy gene expression matrix with 3 rows (genes) and 4 columns (samples), where the entries represent counts for each gene in each sample.

5.4.2 Adding Row and Column Names

You can enhance the clarity of your matrix by adding row and column names. This is particularly useful when dealing with real biological data. For example:

```
rownames(expression_mat) <- c("gene1", "gene2", "gene3")
colnames(expression_mat) <- c("sample1", "sample2", "sample3", "sample4")
expression_mat
```

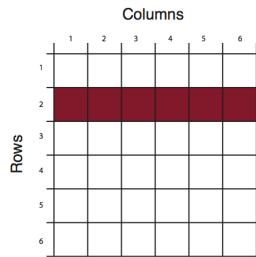
```
##           sample1 sample2 sample3 sample4
## gene1        1        4        7       10
## gene2        2        5        8       11
## gene3        3        6        9       12
```

Now, instead of numerical indices, your matrix displays gene and sample names, making it easier to interpret.

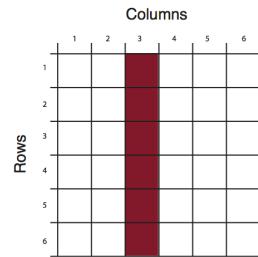
5.4.3 Subsetting a Matrix

Subsetting allows you to extract specific rows and columns from a matrix. You can use numerical indices, row/column names, or logical vectors. Remember, R is 1-based. Indices start at 1 while Python starts at 0.

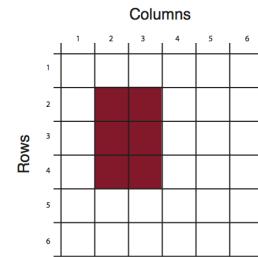
```
mat[2, ]
```



```
mat[, 3]
```



```
mat[2:4, 2:3]
```



5.4.3.1 Subsetting using numerical indices

```
# Accessing a single element
expression_mat[1, 2]
```

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

slice a chunk

```
expression_mat[1:2, 1:2]
```

```
##      sample1 sample2
## gene1      1      4
## gene2      2      5
```

If you leave either the row index blank or the column index, it will subset all the rows or columns. Subset the first two rows and all columns:

```
expression_mat[1:2, ]
```

```
##      sample1 sample2 sample3 sample4
## gene1      1      4      7     10
## gene2      2      5      8     11
```

Subset the columns 2 and 4 and all rows

```
expression_mat[, c(2,4)]
```

```
##      sample2 sample4
## gene1      4     10
## gene2      5     11
## gene3      6     12
```

5.4.3.2 Subsetting using row names

```
# Accessing a specific gene's data
expression_mat["gene3", ]
```

```
## sample1 sample2 sample3 sample4
##      3       6       9      12
```

When only one row or one column is left after subsetting, R returns a vector instead of a matrix. To return a single row or column matrix, add `drop=FALSE`.

```
expression_mat["gene3", , drop=FALSE]
```

```
##      sample1 sample2 sample3 sample4
## gene3      3       6       9      12
```

5.4.3.3 Subsetting using column names

```
# Using predefined gene and sample names
differential_genes<- c("gene3", "gene1")
```

```
expression_mat[differential_genes, c("sample1", "sample2")]
```

```
##      sample1 sample2
## gene3      3       6
## gene1      1       4
```

You see how the matrix is subsetted and the row names are ordered as in `differential_genes`.

5.4.3.4 Subsetting using logical vectors

We have a matrix called `expression_mat` that contains gene expression data, and you want to subset it to include only the rows corresponding to certain “differential genes.” Here’s how you can do it:

```
logical_vec_genes <- rownames(expression_mat) %in% differential_genes
expression_mat[logical_vec_genes, ]
```

```
##      sample1 sample2 sample3 sample4
## gene1      1       4       7      10
## gene3      3       6       9      12
```

5.4.4 Calculations with Matrices

You can perform various calculations on matrices, such as calculating the sum of counts for each sample (column level) or gene (row level):

```
colSums(expression_mat)

## sample1 sample2 sample3 sample4
##       6      15      24      33

rowSums(expression_mat)

## gene1 gene2 gene3
##    22    26    30
```

5.4.5 Normalization

Normalization is crucial in RNAseq data analysis to account for differences in sequencing depth and gene length. Two common methods for normalization are RPKM (Reads Per Kilobase per Million) and TPM (Transcripts Per Million). Watch this video to understand better of their differences

RPKM (Reads Per Kilobase per Million) and TPM (Transcripts Per Million) are two widely-used methods for normalizing gene expression data in RNAseq, with RPKM considering gene length and total reads per sample and TPM further accounting for differences in sequencing depth between samples.

Here's how you can normalize a matrix to RPKM and TPM:

```
# Gene lengths (in kilobases)
gene_lengths <- c(1000, 2000, 3000)

# Normalizing to RPKM
rpkm_normalized <- t(t(expression_mat)/colSums(expression_mat))/gene_lengths * 1e6
rpkm_normalized

##           sample1 sample2 sample3 sample4
## gene1 166.6667 266.6667 291.6667 303.0303
## gene2 166.6667 166.6667 166.6667 166.6667
## gene3 166.6667 133.3333 125.0000 121.2121
```

```
# Normalizing to TPM
tpm_normalized <- t(t(expression_mat/gene_lengths) / colSums((expression_mat/gene_lengths))

##           sample1   sample2   sample3   sample4
## gene1 333333.3 470588.2 500000.0 512820.5
## gene2 333333.3 294117.6 285714.3 282051.3
## gene3 333333.3 235294.1 214285.7 205128.2
```

Note, when you divide a matrix by a vector, the operation is row-wise.

```
expression_mat

##           sample1   sample2   sample3   sample4
## gene1      1        4        7       10
## gene2      2        5        8       11
## gene3      3        6        9       12

gene_lengths

## [1] 1000 2000 3000

expression_mat/gene_lengths

##           sample1   sample2   sample3   sample4
## gene1 0.001 0.0040 0.007 0.0100
## gene2 0.001 0.0025 0.004 0.0055
## gene3 0.001 0.0020 0.003 0.0040
```

That's why if we want to divide the matrix by the column sum, we use the `t()` to transpose the matrix first.

5.4.6 Conclusion

Understanding matrices and their manipulation is fundamental when working with biological data in R. Whether you're analyzing gene expression or any other numerical data, these matrix operations are essential tools for data exploration and analysis in the field of bioinformatics.

5.5 Essential Functions in R

In this foundational lesson, we will explore several fundamental R functions that are indispensable for your daily bioinformatics tasks. As a beginner in programming, mastering these basic building blocks will provide you with a sturdy groundwork upon which to build your bioinformatics skills.

5.5.1 `length` Function

The `length` function is your go-to tool for determining the size of a vector or list in R. It's immensely useful when working with gene expression data, as it allows you to gauge the number of elements in your dataset. Let's dive into an example:

```
expression_vec <- c(10, 25, 30, 12, 20)
names(expression_vec) <- c("gene1", "gene2", "gene3", "gene4", "gene5")

length(expression_vec)

## [1] 5
```

In this case, our `expression_vec` contains gene expression values, and the `length` function tells us that it comprises five elements. This straightforward function provides a crucial dimension for managing your data.

5.5.2 `unique` Function

When working with genomics data, you'll often encounter lists of genes or sequences. The `unique` function is a valuable asset for identifying and extracting unique elements from such lists. Let's illustrate with a simple example:

```
genes <- c("GeneC", "GeneA", "GeneB", "GeneA")
unique_genes <- unique(genes)

unique_genes

## [1] "GeneC" "GeneA" "GeneB"
```

In this snippet, we have a list of genes, and `unique` helps us extract the unique gene names, which can be essential for various genomics analyses.

5.5.3 `sort` Function

Sorting is a fundamental operation in data manipulation. The `sort` function in R allows you to arrange your data in ascending or descending order. It's particularly handy when dealing with gene lists, as it helps you organize genes alphabetically or numerically. Let's explore some examples:

```
unique_genes <- c("GeneC", "GeneA", "GeneB")

# Sort alphabetically
sort(unique_genes)

## [1] "GeneA" "GeneB" "GeneC"

# Sort alphabetically in descending order
sort(unique_genes, decreasing = TRUE)

## [1] "GeneC" "GeneB" "GeneA"

# Sort numeric values
sort(expression_vec)

## gene1 gene4 gene5 gene2 gene3
##    10     12     20     25     30

# Sort numeric values in descending order
sort(expression_vec, decreasing = TRUE)

## gene3 gene2 gene5 gene4 gene1
##    30     25     20     12     10
```

These examples demonstrate how the `sort` function can be applied to both character and numeric data. Sorting can be particularly useful when you need to organize and prioritize genes or data for downstream analyses.

5.5.4 `cor` Function

The `cor` function is indispensable for bioinformatics tasks that involve assessing the relationships between variables, such as gene expression levels. It calculates the correlation coefficient, which measures the degree of association between two variables. Let's explore how it works:

```

gene1 <- c(10, 15, 20, 25)
gene2 <- c(8, 12, 18, 22)

correlation_coefficient <- cor(gene1, gene2)

correlation_coefficient

## [1] 0.9965458

```

In this example, we calculate the correlation between the expression levels of two genes, `gene1` and `gene2`. The resulting correlation coefficient value provides insights into the similarity of their expression patterns.

It's worth noting that the `cor` function supports various correlation methods, such as Pearson and Spearman. Understanding these correlations is crucial for deciphering gene interactions and conducting network analyses.

5.6 Functions: Organizing Your Code for Reusability

In this section, we will delve into the concept of functions in R—a fundamental building block for creating organized and reusable code. Functions serve as a means to encapsulate specific tasks within your code and can significantly enhance your programming capabilities. In essence, a function is like a black box that takes input, processes it, and produces output, shielding you from the inner workings of the logic. We will explore how to create functions, define input arguments, and ensure they provide meaningful output.

5.6.1 What is a Function?

A function in R is a way to bundle code that accomplishes a specific task or computation. It comprises defined input arguments and a code block, and you can invoke the function whenever you need to execute that particular logic.

Let's begin by creating a simple function to calculate the mean of a numeric vector.

5.6.2 Creating a Basic Function: Calculating the Mean

To create a function, you should follow these steps:

1. Name Your Function: Give your function a meaningful name. In our example, we'll call it `mean_customer`.
2. Use the function Keyword: Begin your function with the function keyword, followed by parentheses. Inside the parentheses, define your input arguments. You can have multiple arguments for a function.
3. Body of the Function: The actual code of the function is enclosed within curly braces {}.

Let's create a `mean_customer` function to compute the mean of a numeric vector:

```
mean_customer <- function(x) {
  total <- sum(x)
  mean_value <- total / length(x)
}
```

In this function, we first calculate the total sum of the input vector `x` using the built-in `sum` function. Then, we divide this sum by the length of the vector to obtain the mean.

5.6.3 Using the Custom Function

Now that we have defined our `mean_customer` function, let's use it with an example vector:

```
input_vec <- c(1, 2, 3, 4)
result <- mean_customer(input_vec)

result

## [1] 2.5
```

You might have noticed that our initial function did not print anything to the console. To make a function display an output, you need a `return` statement. Let's add it to our function.

5.6.4 Returning a Value from the Function

Returning a value in a function in a programming language like R is a fundamental concept that determines what the function does with its computed result. When you include a `return` statement within a function, you are specifying the value that the function should provide as output when it's called elsewhere in your code.

1. Calculation or Operation: Inside the function, there is a block of code that performs some calculation, operation, or task based on the input arguments provided to the function.
2. return Statement: When you include a return statement, it signifies that the function should terminate its execution at that point and immediately provide the value specified after return as its output. This value can be a single variable, an expression, or even a complex data structure.
3. Function Execution: When you call the function in your code, it starts executing from the beginning and proceeds until it encounters the return statement. At that point, the function stops executing further code within its body and exits, returning the value specified in the return statement.
4. Assigning to a Variable: Typically, you capture the returned value by assigning it to a variable when you call the function. This allows you to store and use the result elsewhere in your code.

For example, in R, following the previous example

```
mean_customer <- function(x) {
  total <- sum(x)
  mean_value <- total / length(x)
  return(mean_value)
}
```

Now, when you use `mean_customer(input_vec)`, it will correctly display the mean value.

```
# Calculate the mean of a numeric vector
input_vec <- c(1, 2, 3, 4)
result <- mean_customer(input_vec)
result

## [1] 2.5
```

5.6.5 Optional: Omitting the return Statement

You can also omit the return statement. By default, the last expression in the function will be returned as the output. Here's the updated function:

```
mean_customer <- function(x) {
  total <- sum(x)
  mean_value <- total / length(x)
  mean_value
}
```

The behavior remains the same as before when you use `mean_customer(input_vec)`.

5.6.6 Using more than one argument

Missing data is a common occurrence in real-world biological datasets, and learning how to handle it is crucial for robust data analysis in R. Let's create a custom function to calculate the mean of a numeric vector while accommodating missing values (NAs). We'll introduce the concept of the `na.rm` argument and illustrate how it allows us to decide whether or not to remove NAs before performing calculations.

5.6.7 Understanding Missing Values (NAs) in R

In R, “NA” stands for “Not Available,” and it is used to represent missing or undefined values in your data. Let's start by creating a couple of example vectors:

```
genes <- c("TP53", NA, "MYC")
NA_vec <- c(1, 2, 3, NA)
```

As you can see, our `genes` vector contains a missing value (`NA`). To identify and handle NAs, we can use the `is.na()` function, which returns a logical vector indicating which elements are NAs:

```
is.na(genes)

## [1] FALSE  TRUE FALSE

is.na(NA_vec)

## [1] FALSE FALSE FALSE  TRUE
```

5.6.8 Initial Attempt: A Function Without Handling NAs

Let's start by creating a custom function to calculate the mean of a numeric vector. However, if the vector contains NAs, our initial function doesn't handle them correctly:

```
mean_customer <- function(x){
  total <- sum(x)
  mean_value <- total / length(x)
  return(mean_value)
}
```

When we try to calculate the mean of `NA_vec`, it returns `NA`:

```
mean_customer(NA_vec)
```

```
## [1] NA
```

This outcome is not ideal, especially when we want to calculate the average of the non-missing values.

5.6.9 Adding the `remove_na` Argument

To address this issue, we can enhance our function by introducing a new argument called `remove_na`, which allows us to control whether `NAs` should be removed before performing calculations. By default, we set `remove_na` to `TRUE`, indicating that `NAs` should be removed:

You'll see that we declare the `remove_na` argument with an `=`. That means that if no value is provided, by default it will take the value `TRUE`.

```
mean_customer_NA <- function(x, remove_na = TRUE){
  if (remove_na){
    x <- x[!is.na(x)]
  }
  total <- sum(x)
  mean_value <- total / length(x)
  return(mean_value)
}
```

Now, our function behaves differently based on the value of `remove_na`. If set to `TRUE`, it removes `NAs` from the vector before calculating the mean; if set to `FALSE`, it includes `NAs` in the calculation.

5.6.10 Practical Application

Let's see how this enhanced function works with our example vector:

```
mean_customer_NA(NA_vec) # Default behavior (remove_na = TRUE)
```

```
## [1] 2
```

```
mean_customer_NA(NA_vec, remove_na = TRUE) # Explicitly removing NAs

## [1] 2

mean_customer_NA(NA_vec, remove_na = FALSE) # Including NAs

## [1] NA
```

In the first two calls, we get a result of 2 by removing the `NAs`, while in the last call, we receive `NA` since we chose not to remove them.

5.6.11 Getting help with functions

5.6.11.1 Using the `?` Operator for Documentation.

The `?` operator is a quick and convenient way to access documentation directly within the R environment. Simply type in the console `?` followed by the name of the function you want to learn more about. For example:

```
?mean
```

This command will open the documentation for the `mean()` function, providing details on its usage, arguments, examples, and related functions.

5.6.11.2 Accessing Documentation via the `help()` Function

Alternatively, you can use the `help()` function to retrieve documentation for a specific function. Syntax:

```
help(mean)
```

Executing this command will display the documentation for the `mean()` function in the Help pane of R Studio.

5.6.11.3 Utilizing the `help.search()` Function

If you are unsure about the exact function name or wish to search for functions related to a certain topic, you can use the `help.search()` function. This function allows you to search for keywords across all available documentation. Syntax:

```
help.search("keyword")
```

For example:

```
help.search("linear regression")
```

This command will return a list of relevant documentation entries containing the specified keyword, assisting you in finding relevant functions and packages.

5.6.11.4 Exploring Online Resources and Community Forums

In addition to built-in documentation, online resources such as the official R documentation website (<https://www.rdocumentation.org/>) and community forums like Stack Overflow (<https://stackoverflow.com/>) are valuable sources of information and support. You can search for specific functions, read user discussions, and even ask questions if needed.

5.6.12 A Real World Example

In the field of bioinformatics, analyzing gene expression data is a fundamental task. One common analysis involves identifying the most highly variable genes within a gene expression matrix. Let's create a custom R function for this purpose.

Before diving into the creation our function, it's essential to understand some key pre-concepts. Firstly, gene expression data typically consists of a matrix where rows represent genes and columns represent samples or conditions. Second, the variability of gene expression across samples is a crucial metric, often measured by variance or standard deviation. Highly variable genes can provide valuable insights into biological processes.

Let's create a custom function, `findVariableGenes`, which identifies the top N most highly variable genes in a gene expression matrix. Below is the code for the function, with explanations:

```
# Define the custom function
findVariableGenes <- function(expr_matrix, n = 10) {
  # Calculate variances for each gene
  gene_variances <- apply(expr_matrix, MARGIN = 1, var)

  # Sort genes by variance in descending order
  sorted_genes <- sort(gene_variances, decreasing = TRUE)
  top_n_genes <- sorted_genes[1:n]
}
```

```

sorted_genes <- sort(gene_variances, decreasing=TRUE)

# Select the top N variable genes
top_n <- names(sorted_genes[1:n])

# Return the names of top variable genes
return(top_n)
}

```

- expr_matrix: This is the gene expression matrix you want to analyze.
- n: The number of top variable genes you want to identify (default is 10).
- apply function applies the var function to the matrix for rows (MARGIN =1). if you want to apply a function for columns, use MARGIN =2.

Now, let's put our custom function to the test using some randomly generated gene expression data.

5.6.12.1 Generating Random Data

We'll create a gene expression matrix with 25 genes (rows) and 4 samples (columns) using normally distributed random data. To ensure reproducibility, we'll set a seed value (123). Here's the code and the resulting data matrix:

```

# Set a seed for reproducibility (ensure all get the same random data)
set.seed(123) # This sets a random seed to ensure that the random data generated below is reproducible

data <- matrix(rnorm(100), ncol = 4) # This line generates a matrix with 100 random numbers
rownames(data) <- paste0("gene", 1:25) # This line assigns row names to the matrix, labeling them as genes

# Display the generated data
print(data)

```

	[,1]	[,2]	[,3]	[,4]
## gene1	-0.56047565	-1.68669331	0.25331851	1.025571370
## gene2	-0.23017749	0.83778704	-0.02854676	-0.284773007
## gene3	1.55870831	0.15337312	-0.04287046	-1.220717712
## gene4	0.07050839	-1.13813694	1.36860228	0.181303480
## gene5	0.12928774	1.25381492	-0.22577099	-0.138891362
## gene6	1.71506499	0.42646422	1.51647060	0.005764186
## gene7	0.46091621	-0.29507148	-1.54875280	0.385280401
## gene8	-1.26506123	0.89512566	0.58461375	-0.370660032
## gene9	-0.68685285	0.87813349	0.12385424	0.644376549

```
## gene10 -0.44566197  0.82158108  0.21594157 -0.220486562
## gene11  1.22408180  0.68864025  0.37963948  0.331781964
## gene12  0.35981383  0.55391765 -0.50232345  1.096839013
## gene13  0.40077145 -0.06191171 -0.33320738  0.435181491
## gene14  0.11068272 -0.30596266 -1.01857538 -0.325931586
## gene15 -0.55584113 -0.38047100 -1.07179123  1.148807618
## gene16  1.78691314 -0.69470698  0.30352864  0.993503856
## gene17  0.49785048 -0.20791728  0.44820978  0.548396960
## gene18 -1.96661716 -1.26539635  0.05300423  0.238731735
## gene19  0.70135590  2.16895597  0.92226747 -0.627906076
## gene20 -0.47279141  1.20796200  2.05008469  1.360652449
## gene21 -1.06782371 -1.12310858 -0.49103117 -0.600259587
## gene22 -0.21797491 -0.40288484 -2.30916888  2.187332993
## gene23 -1.02600445 -0.46665535  1.00573852  1.532610626
## gene24 -0.72889123  0.77996512 -0.70920076 -0.235700359
## gene25 -0.62503927 -0.08336907 -0.68800862 -1.026420900
```

5.6.12.2 Using the Custom Function

Now that we have our gene expression data, let's apply our `findVariableGenes` function to identify the top 10 most highly variable genes:

```
# Use the custom function to find highly variable genes
highly_variable <- findVariableGenes(data, n = 10)

# Display the list of highly variable genes
print(highly_variable)

## [1] "gene22" "gene23" "gene1"  "gene19" "gene3"  "gene20" "gene18" "gene16"
## [9] "gene4"  "gene8"
```

Encapsulating logic into functions makes our code more organized, reusable, and scalable. Functions make code more organized, reusable, and scalable. As you code more in R, you'll want to encapsulate logic into functions just as shown here.

5.6.13 Conclusion

Understanding functions and their ability to encapsulate code and return specific values is crucial in R programming. Functions enhance code organization, maintainability, and reusability, making them a valuable tool for any data analyst or scientist. The flexibility and efficiency they offer become increasingly evident as you tackle more complex data analysis tasks in R.

5.7 Common Mistakes to avoid

In this lesson we will discuss some common mistakes that absolute beginners often make when learning R. Learning a new programming language can be challenging, and it's natural to encounter stumbling blocks along the way. By understanding these common mistakes, you can avoid them and become a more proficient R programmer.

5.7.1 Mixing Data Types

One common mistake is mixing different data types in operations. For instance, trying to add a number to a string or perform mathematical operations on non-numeric data types.

```
# Example:
x <- "5"
y <- 3
z <- x + y # This will result in an error because you cannot add a string and a number
```

5.7.2 Forgetting Function Parentheses

Another mistake is forgetting to include parentheses when calling functions. In R, functions typically require parentheses, even if they don't have any arguments.

```
# Incorrect:
print "Hello, World!"

# Correct:
print("Hello, World!")
```

5.7.3 Overwriting Built-in Functions

Sometimes you might unintentionally overwrite built-in functions or variable names, causing unexpected behavior in your code.

```
# Example:
# Incorrect:
mean <- function(x) {
  sum(x) / length(x)
}
# Now, mean function is overwritten and will not work as expected.
# e.g., it can not handle na.rm argument
```

5.7.4 Misunderstanding Variable Scoping

Variable scoping in R defines where a variable can be used in a program. If a variable is defined inside a function, it's only accessible within that function (local scope). Variables defined outside functions, usually at the start of a script or in the main program, can be used anywhere (global scope).

Understanding variable scoping is crucial for avoiding errors and writing maintainable code. One common mistake is assuming that variables defined in one part of the program will be accessible from another part.

Let's look at an example:

```
# Example 1: Incorrect variable scoping
calculate_sum <- function(a, b) {
  result <- a + b
}

# Trying to access 'result' outside the function will result in an error.
print(result)

## [1] 2.5
```

5.7.5 Best practices for variable scoping include

1. Explicitly Pass Variables: When variables are needed in different parts of the program, it's better to explicitly pass them as arguments to functions rather than relying on global variables.

```
# Example 2: Explicitly passing variables
calculate_sum <- function(a, b) {
  result <- a + b
  return(result)
}

# Call the function with arguments
result <- calculate_sum(3, 5)
print(result)

## [1] 8
```

2. Use Meaningful Variable Names: Clear and meaningful variable names can help avoid confusion about variable scope and improve code readability.

```
# Example 3: Clear variable names
calculate_area <- function(length, width) {
  area <- length * width
  return(area)
}

# Call the function with arguments
area <- calculate_area(4, 5)
print(area)

## [1] 20
```

3. Avoid Modifying Global Variables Within Functions: Modifying global variables within functions can lead to unexpected behavior and make code harder to understand and debug. Instead, prefer returning values from functions.

```
# Example 4: Avoid modifying global variables
x <- 10

modify_variable <- function() {
  x <- 20 # This creates a new local variable 'x', it does not modify the global 'x'
}

modify_variable()
print(x) # Output: 10 (global 'x' remains unchanged)
```

```
## [1] 10
```

5.7.6 Conclusion

Learning R can be a rewarding experience, but it's common to encounter challenges along the way. By being aware of these common mistakes and understanding the allowed operations in R, you can avoid many pitfalls and become a more proficient R programmer. Remember to practice regularly and don't hesitate to seek help from resources like documentation, online tutorials, and community forums. Happy coding!

Chapter 6

Controlling the flow of our programs

6.1 Boolean Operators

In R, understanding boolean operators is crucial for making logical comparisons and decisions in your code. Boolean operators are used to compare values and evaluate conditions, providing a foundation for decision-making processes in programming.

In this tutorial, we will explore the basics of boolean operators, including equality, greater-than, and less-than comparisons. We will also delve into logical operations, introducing the essential concepts of “and” and “or” with both vectorized and non-vectorized forms.

6.1.1 Comparison Operators

Comparison operators allow us to compare values and return a boolean result, either `TRUE` or `FALSE`. Let’s start with some common comparison operators:

6.1.1.1 Equality (==)

The equality operator (`==`) checks if two values are equal. For example:

```
"A" == "A"
```

```
## [1] TRUE
```

```
3 == 3
```

```
## [1] TRUE
```

6.1.1.2 Greater Than (>)

The greater-than operator (>) checks if one value is greater than another:

```
5 > 3
```

```
## [1] TRUE
```

6.1.1.3 Less Than (<)

The less-than operator (<) checks if one value is less than another:

```
5 < 3
```

```
## [1] FALSE
```

6.1.2 Logical Operators

Now, let's explore logical operators, which allow us to combine multiple conditions and make more complex decisions in our code.

6.1.2.1 Vectorized “AND” (&)

The vectorized “and” operator (&) allows us to perform element-wise comparisons on vectors. It returns a vector of boolean values, which is extremely useful when dealing with data sets. For example:

```
# Check if elements in -2:2 are greater than or equal to 0
-2:2 >= 0
```

```
## [1] FALSE FALSE  TRUE  TRUE  TRUE
```

```
# Check if elements in -2:2 are less than or equal to 0
-2:2 <= 0
```

```
## [1]  TRUE  TRUE  TRUE FALSE FALSE
```

```
# Combine the two conditions using vectorized "and"
(-2:2 >= 0) & (-2:2 <= 0)
```

```
## [1] FALSE FALSE TRUE FALSE FALSE
```

6.1.2.2 Vectorized “OR” (|)

Similar to the “and” operator, the vectorized “or” operator (|) performs element-wise comparisons and returns a vector of boolean values. Here’s an example:

```
# Check if elements in -2:2 are greater than or equal to 0
-2:2 >= 0
```

```
## [1] FALSE FALSE TRUE TRUE TRUE
```

```
# Check if elements in 2:6 are less than or equal to 0
2:6 <= 0
```

```
## [1] FALSE FALSE FALSE FALSE FALSE
```

```
# Combine the two conditions using vectorized "or"
(-2:2 >= 0) | (2:6 <= 0)
```

```
## [1] FALSE FALSE TRUE TRUE TRUE
```

6.1.2.3 Non-Vectorized “AND” (&&) and “OR” (||)

Non-vectorized “and” (&&) and “or” (||) operators in R are used for performing logical operations that return a **single** boolean value based on the evaluation of multiple conditions

The non-vectorized forms of “and” (&&) and “or” (||) return a single value and are typically used for non-vectorized logical operations. For example:

```
# Non-vectorized "and" operator
(-2:2 >= 0) && (-2:2 <= 0)
```

```
## [1] FALSE
```

Keep in mind that as of R 4.3.0, these operators must be given inputs of length 1.

6.1.3 Conclusion

Understanding boolean operators and logical operations is fundamental in programming with R. These operators enable you to make decisions based on comparisons, creating more dynamic and powerful code. Whether you are comparing values or combining conditions, boolean operators are essential tools in your programming toolkit. Experiment with different comparisons and logical combinations to gain a deeper understanding of their versatility and practicality in R.

6.2 Conditional statements (`if`, `else`)

In the world of programming, making decisions based on data is crucial. Imagine you're analyzing gene expression data, and you want to process it differently depending on whether a gene is highly expressed or not. In R, this decision-making ability is known as control flow, and it's essential for creating flexible and adaptive programs. In this tutorial, we'll explore the power of the `if` statement, which allows us to control the flow of our R code based on specific conditions. We'll walk through practical examples, starting with gene expression analysis.

6.2.1 The `if` Statement

The `if` statement is a fundamental tool for controlling program flow in R. It checks whether a specified condition evaluates to `TRUE` or `FALSE`. Depending on the result, either the code within the `if` block or the `else` block (if defined) gets executed.

Imagine you're working with DNA sequences, and you want to check if a given sequence contains the sequence motif "ATG" which is a start codon in genetics. Here's a basic example using the `if` statement:

```
# Bioinformatics example: DNA sequence
sequence <- "GCTAGTGTAGCGT"

# Check if the sequence contains the start codon "ATG"
if (grep("ATG", sequence)) {
  print("Start codon found")
} else {
  print("Start codon not found")
}

## [1] "Start codon not found"
```

In this scenario, the if statement checks whether the sequence contains “ATG” using the grepl function. If it’s found, it prints a message indicating that the start codon is present; otherwise, it prints a message indicating that the start codon is not found.

If you want to check whether both start codon and stop codon are in the DNA sequence:

```
sequence <- "GCTAGTGTAGCGT"

# Check if the sequence contains the start codon "ATG" or the stop codon "TAA"
if (grepl("ATG", sequence) || grepl("TAA", sequence)) {
  print("Start codon or Stop Codon are found")
} else {
  print("No Start codon or Stop codon are found")
}

## [1] "No Start codon or Stop codon are found"
```

Note we use `||`, the non-vectorized version or for condition checking.

6.2.2 else statement

Now, let’s explore the `else` statement in the context of DNA sequence analysis. Suppose you want to perform a different action if the sequence doesn’t contain the start codon. For instance, you might want to check for a stop codon. Here’s a simplified example:

```
# Bioinformatics example: DNA sequence without the start codon
sequence <- "CGTACTAGCGT"

# Check if the sequence contains the start codon "ATG"
if (grepl("ATG", sequence)) {
  print("Start codon found")
} else {
  print("Start codon not found, checking for stop codon")

  # Check if the sequence contains the stop codon "TAA"
  if (grepl("TAA", sequence)) {
    print("Stop codon found")
  } else {
    print("No start or stop codon found")
  }
}
```

```
## [1] "Start codon not found, checking for stop codon"
## [1] "No start or stop codon found"
```

In this example, the if statement first checks for the presence of the start codon “ATG.” If it’s not found, it enters the else block and checks for the stop codon “TAA.” Depending on the outcome, it prints the corresponding message.

6.2.3 else if

If you have multiple conditions and want to test them one by one, the pseudo-code is:

```
if (condition1) {
    expr1
} else if (condition2) {
    expr2
} else if (condition3) {
    expr3
} else {
    expr4
}
```

6.2.4 Exercise

In real biology, there are multiple stop codons: UAA, UAG, and UGA (T is converted to U in RNA). Use the multiple else if clauses to find the stop codon: TAA, TAG, and TGA in a DNA sequence.

		Second Position								Third Position			
		U		C		A		G					
		code	amino acid	code	amino acid	code	amino acid	code	amino acid				
First Position	U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U			
		UUC		UCC		UAC		UGC		C			
		UUA	leu	UCA		UAA	STOP	UGA	STOP	A			
		UUG		UCG		UAG	STOP	UGG	trp	G			
	C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U			
		CUC		CCC		CAC		CGC		C			
		CUA		CCA		CAA	gln	CGA		A			
		CUG		CCG		CAG		CGG		G			
	A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U			
		AUC		ACC		AAC		AGC		C			
		AUA		ACA		AAA	lys	AGA	arg	A			
		AUG	met START	ACG		AAG		AGG		G			
	G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U			
		GUC		GCC		GAC		GGC		C			
		GUA		GCA		GAA	glu	GGA		A			
		GUG		GCG		GAG		GGG		G			

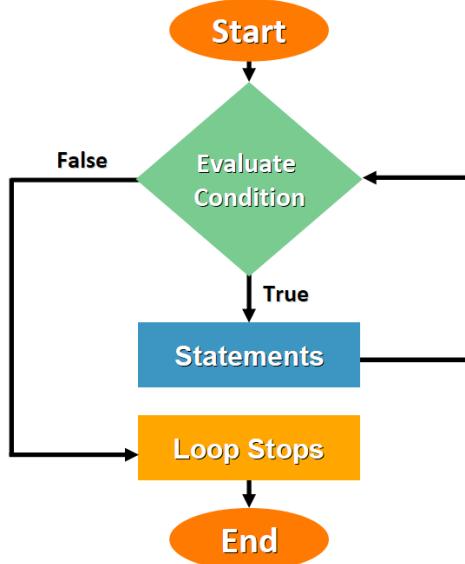
6.2.5 Conclusion

In conclusion, the if and else statements in R provide bioinformaticians with the means to adapt their analyses and make data-driven decisions. By employing these control flow structures, researchers can enhance the reproducibility and adaptability of their bioinformatics workflows, ultimately advancing our understanding of biological systems. Whether you're searching for genetic elements or classifying sequences, mastering control flow is an essential skill in the bioinformatics toolbox.

6.3 Loops

In the world of data analysis and programming, loops are indispensable tools for executing repetitive tasks efficiently. They allow us to automate processes like processing multiple files or iterating through steps in an analysis pipeline. In R, we have two main types of loops: the `while` loop and the `for` loop. In this section, we'll delve into their usage with real-world examples and explore when to employ each type.

6.3.1 The `while` Loop



The `while` loop repeatedly runs a block of code as long as a specified condition remains true. A practical scenario could involve quality-controlling sequencing files one by one until we encounter a file that fails a test. Here's an example:

```

expression_vec <- c(0, 4, 8, 16, 32)
new_expression_vec <- c()
i <- 1

while (i <= length(expression_vec)) {
  expression_value <- expression_vec[i]
  new_expression_vec[i] <- log2(expression_value) # Calculate the base-2 logarithm
  i <- i + 1 # Increment the index to process the next element
}

new_expression_vec
## [1] -Inf     2      3      4      5
  
```

In the given code, the loop counter `i` is initially set to 1. The while loop iterates as long as `i` is less than or equal to the length of the `expression_vec` vector. In each iteration, it calculates the base-2 logarithm of the current element in `expression_vec` and stores it in `new_expression_vec`, while incrementing the value of `i` by 1. This incrementing of `i` ensures that the loop processes the

next element in the vector during each iteration until all elements have been processed.

Note, the calculation in R is vectorized, you can use:

```
log2(expression_vec)
```

```
## [1] -Inf    2     3     4     5
```

to get the same results.

6.3.2 While Loops in Real Life

Researchers often work with large datasets generated from DNA sequencing machines. These machines produce files containing vast amounts of genetic information, and it's crucial to ensure the quality of this data before using it in further analyses. Imagine these files as a collection of books, each representing genetic information from a different sample.

To ensure that the data is reliable, scientists perform a process called quality control (QC). It's similar to checking books for errors, missing pages, or smudged ink before studying their contents. One important aspect of QC in sequencing data is assessing the quality of the readings from the sequencing machine. This quality is often represented as a numerical score, with higher scores indicating better data. Researchers typically set a threshold value, like a score of 30, which they consider acceptable quality.

The code snippet below (this is just a pseudo-code) illustrates how to iterate through a vector of file names, read each file, and check if the mean quality score falls below a specified threshold:

```
files <- c("sample1.fq", "sample2.fq", "sample3.fq")
i <- 1

# Start a `while` loop with the condition: while `i` is less than or equal to the length of `files`
while (i <= length(files)) {

  # Read the current file
  fq <- readFastq(files[i])

  # Check if the mean quality score is below 30
  if (meanQual(fq) < 30) {
    # Print a failure message if the quality check fails
    print(paste(files[i], "failed QC"))
    # Exit the loop using `break`
  }
}
```

```

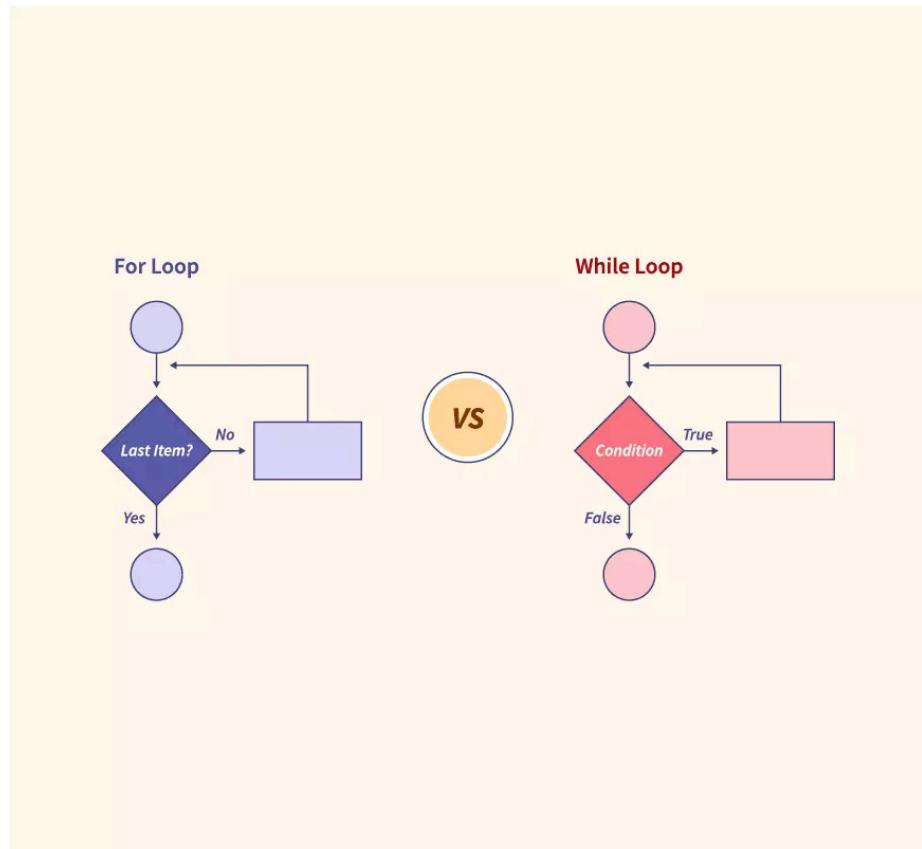
    break
}

# Increment the index `i` to move to the next file
i <- i + 1
}

```

In this case, the loop iterates through files, reads each one, and performs quality control. If a file fails the quality check, the loop prints a failure message and exits.

6.3.3 The `for` Loop



On the other hand, the `for` loop iterates through a predefined sequence of values. Consider a scenario where we want to standardize gene expression across all samples using Z-scores:

```

# Create a matrix of gene expression data
expression_mat <- matrix(1:12, nrow = 3, ncol = 4)

# Define the row names (gene names) and column names (sample names)
rownames(expression_mat) <- c("gene1", "gene2", "gene3")
colnames(expression_mat) <- c("sample1", "sample2", "sample3", "sample4")

# Get the gene names
genes <- rownames(expression_mat)

# Start a for loop that iterates through each gene name 'g' in 'genes'
for (g in genes) {

  # Calculate the mean expression for the current gene 'g'
  mean_expr <- mean(expression_mat[g, ])

  # Calculate the standard deviation of expression for the current gene 'g'
  sd_expr <- sd(expression_mat[g, ])

  # Standardize the expression values for the current gene 'g' using Z-scores
  expression_mat[g, ] <- (expression_mat[g, ] - mean_expr) / sd_expr
}

# Print the resulting standardized expression matrix 'expression_mat'
expression_mat

```

	sample1	sample2	sample3	sample4
## gene1	-1.161895	-0.3872983	0.3872983	1.161895
## gene2	-1.161895	-0.3872983	0.3872983	1.161895
## gene3	-1.161895	-0.3872983	0.3872983	1.161895

In this example, the for loop efficiently iterates through each gene, calculates the mean and standard deviation of expression, and then standardizes that gene's row. This process repeats for all genes, allowing for consistent normalization.

Of course, you can use the `scale` function in R to do this directly without using a `for` loop. Note, `scale` works by columns. To get the same result, you need to first transpose the matrix, scale it and then transpose it back.

```
t(scale(t(expression_mat)))
```

	sample1	sample2	sample3	sample4
## gene1	-1.161895	-0.3872983	0.3872983	1.161895
## gene2	-1.161895	-0.3872983	0.3872983	1.161895

```
## gene3 -1.161895 -0.3872983 0.3872983 1.161895
## attr(,"scaled:center")
## gene1 gene2 gene3
##      0      0      0
## attr(,"scaled:scale")
## gene1 gene2 gene3
##      1      1      1
```

6.3.4 Conclusion

It's crucial to understand that loops can provide fine-grained control for accessing and transforming data at an element level. However, in many cases, R offers vectorized operations that simplify code and make it more readable, as demonstrated with the Z-score calculation using the scale function.

Remember, `while` loops and `for` loops are valuable tools in your data analysis toolkit, but it's essential to choose the most suitable method for the task at hand. By mastering these loop structures, you can streamline your data analysis and automation processes, making your work more efficient and precise. Computers are good at repetitive work. Whenever you are manually doing the same task multiple times, think of the loops!

6.4 Gene Expression Annotation using Loops and Control Structures

We often encounter scenarios where we need to categorize data based on specific criteria. In this example, we'll use a combination of loops and control structures in the R programming language to add custom annotations to gene expression measurements for a group of patients. We'll categorize their expression levels into different classes: "not-expressed," "low," "medium," and "high."

6.4.1 The Data

Suppose we have gene expression measurements for 5 patients stored in a vector called `expression_vec`. These measurements represent the expression levels of a specific gene.

```
expression_vec <- c(0, 5, 10, 6, 22)
```

6.4.2 Creating Annotations

We want to annotate each patient's expression status based on the range of their expression values. To do this, we'll initiate an empty vector called `annotations` to store our annotations.

```
annotations <- c()
```

Now, let's go through the process step by step.

6.4.3 The Loop

```
for (expression_value in expression_vec) {
  if (expression_value == 0) {
    annotation <- "not-expressed"
  } else if (expression_value > 0 & expression_value < 5) {
    annotation <- "low"
  } else if (expression_value >= 5 & expression_value < 20) {
    annotation <- "medium"
  } else {
    annotation <- "high"
  }
  annotations <- c(annotations, annotation)
}
```

Here's what's happening in the code:

- We use a for loop to iterate through each `expression_value` in the `expression_vec` vector.
- Inside the loop, we use a series of if and else if statements to categorize each `expression_value` based on its range.
- If the value is exactly 0, we assign the annotation “not-expressed.”
- If the value is greater than 0 but less than 5, we assign the annotation “low.”
- If the value is greater than or equal to 5 but less than 20, we assign the annotation “medium.”
- If none of the above conditions are met, we assign the annotation “high.”
- Finally, we append each annotation to the `annotations` vector.

6.4.4 Output

Let's see the results of our annotations:

```
annotations
```

```
## [1] "not-expressed" "medium"      "medium"      "medium"
## [5] "high"
```

6.4.5 Putting it all together

```
expression_vec<- c(0, 5, 10, 6, 22)

# initiate an empty vector
annotations <- c()

for (expression_value in expression_vec){
  if ( expression_value ==0 ){
    annotation <- "not-expressed"
  } else if (expression_value >0 & expression_value < 5) {
    annotation<- "low"
  } else if (expression_value >=5 & expression_value <20) {
    annotation<- "medium"
  } else {
    annotation<- "high"
  }
  annotations<- c(annotations, annotation)
}

annotations

## [1] "not-expressed" "medium"      "medium"      "medium"
## [5] "high"
```

In R, everything is vectorized. There is a much better way to achieve the same thing using the `case_when` function in the `dplyr` package. We will cover it in the later lecture.

6.4.6 Conclusion

This approach to categorizing gene expression data is essential in various biological and medical research contexts. For example:

- Drug Development: When studying the impact of a drug on gene expression, researchers need to categorize gene expression levels to assess the drug's effectiveness.
- Cancer Research: Identifying genes with high or low expression levels can provide insights into cancer progression and potential therapeutic targets.
- Disease Biomarker Discovery: Categorizing gene expression in patients with a specific disease can help identify biomarkers for early diagnosis.

By combining loops and control structures as shown in this example, scientists and analysts can efficiently handle and interpret complex biological data.

6.5 Let's solve a Challenge

You're given a vector of daily average temperatures (in Celsius) for a month. Your task is to analyze the temperature data to find out the following:

- The number of days with temperatures above the monthly average.
- Whether any day's temperature exceeds 30°C (considering it as a threshold for a very hot day).
- The number of days with temperatures below 15°C (considering it as a threshold for a cold day).

Given Data:

```
temperatures <- c(12, 14, 16, 20, 22, 24, 26, 28, 30, 32, 18, 16, 14, 22, 24, 26, 20, 18, 17, 15)
```

Tasks:

1. Calculate the monthly average temperature.
2. Use a loop to iterate through the temperatures vector.
 - For each temperature, check if it's above the monthly average and count these occurrences.
 - Check if there's any day with a temperature exceeding 30°C.
 - Count the number of days with temperatures below 15°C.
3. Print the results:
 - Total number of days above the monthly average.
 - Whether there was a very hot day (temperature > 30°C).
 - Number of cold days (temperature < 15°C).

6.6 Solution

Before diving into the solution, I encourage all students to take a moment to challenge themselves and attempt to solve the problem independently. This coding exercise provides a valuable opportunity to practice essential programming concepts in R, such as loops, conditional statements, and basic data manipulation. Start by considering how you would calculate the monthly average temperature from a list of daily temperatures and how you might track the number of days above the average, very hot days (above 30°C), and cold days (below 15°C). Once you've given it a try, feel free to compare your approach with the provided solution to deepen your understanding and refine your coding skills. Happy coding!

6.6.1 using a for loop

```
# Given data: Daily average temperatures for a month (in Celsius)
temperatures <- c(12, 14, 16, 20, 22, 24, 26, 28, 30, 32, 18, 16, 14, 22, 24, 26, 20, 15)

# Manually calculate the monthly average temperature
total_temperature <- 0 # Initialize a variable to hold the sum of all temperatures
for (temp in temperatures) {
    total_temperature <- total_temperature + temp # Accumulate the total temperature
}
monthly_average <- total_temperature / length(temperatures) # Divide by the number of temperatures
print(paste("Monthly average temperature:", monthly_average, "C"))

## [1] "Monthly average temperature: 17.3 C"

# Initialize counters for the conditions
days_above_average <- 0
very_hot_days <- 0
cold_days <- 0

# Use a loop to iterate through the temperatures vector
for (temp in temperatures) {
    # Check if the temperature is above the monthly average
    if (temp > monthly_average) {
        days_above_average <- days_above_average + 1
    }

    # Check if the temperature exceeds 30°C (very hot day)
    if (temp > 30) {
        very_hot_days <- very_hot_days + 1
    }
}
```

```

}

# Check if the temperature is below 15°C (cold day)
if (temp < 15) {
  cold_days <- cold_days + 1
}
}

# Print the results
print(paste("Number of days above the monthly average:", days_above_average))

## [1] "Number of days above the monthly average: 13"

print(paste("Number of very hot days (temperature > 30°C):", very_hot_days))

## [1] "Number of very hot days (temperature > 30°C): 1"

print(paste("Number of cold days (temperature < 15°C):", cold_days))

## [1] "Number of cold days (temperature < 15°C): 12"

```

In this solution, we start with a list of daily average temperatures for a month, stored in the ‘temperatures’ vector. The first part of the code calculates the monthly average temperature by adding up all the daily temperatures and then dividing the total by the number of days in the month. We use a loop to go through each temperature, adding it to the ‘total_temperature’ variable. Once we have the sum, we divide it by the number of days to find the average and print it using the ‘cat’ function.

The second part of the code uses loops and conditional statements to analyze the temperatures. It tracks three things: the number of days with temperatures above the monthly average, the number of very hot days (where the temperature is above 30°C), and the number of cold days (where the temperature is below 15°C). For each temperature in the ‘temperatures’ vector, the code checks if it’s above the monthly average, above 30°C, or below 15°C, and increments the corresponding counter if the condition is met. Finally, the code prints out the results using ‘paste’, displaying the count of days for each condition.

6.6.2 vectorized solution

The solution shown here is how you usually solve the problem in python. However, as I introduced earlier, R is vectorized, many of the calculations can be simplified.

```

temperatures <- c(12, 14, 16, 20, 22, 24, 26, 28, 30, 32, 18, 16, 14, 22, 24, 26, 20, 22)

monthly_average<- mean(temperatures)

monthly_average

## [1] 17.3

# get a logical vector
very_hot_days_lgl<- temperatures > 30

very_hot_days_lgl

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE
## [13] FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE

# remember in R, FALSE is 0 and TRUE is 1 under the hood. you can sum it up to
# find how many are TRUE
very_hot_days<- sum(very_hot_days_lgl)
very_hot_days

## [1] 1

```

Similarly, you can

```

cold_days<- sum(temperatures < 15)
cold_days

## [1] 12

```

You see how powerful R is when combining the logical vector and its vectorized feature!

6.7 Section complete

Congratulations on completing this section of our course! You've made significant progress in understanding the essentials of boolean operators, control flow with if statements, and the power of loops. These foundational skills are crucial for analyzing data, automating tasks, and making logical decisions in your

code. It's impressive how much you've learned and can now apply to real-world problems, from gene expression analysis to data quality control and beyond.

Keep this momentum going as you move forward. The concepts you've mastered here will serve as building blocks for more advanced programming techniques and analytical methods. Remember, practice is key to deepening your understanding and honing your skills. Let's move on to the next section.

Chapter 7

Going more in-depth with R

7.1 Handling Missing Values

In bioinformatics, dealing with missing data is a common challenge. Missing values can arise from various sources, including experimental limitations or data collection errors. It's crucial to understand how to identify and handle these missing values effectively to ensure the accuracy of your analyses. In R, missing values are represented by `NA`, indicating the absence of a value. In this tutorial, we will explore key properties of `NA`, learn how to identify missing values, and discover techniques to handle them in your gene expression datasets.

7.1.1 Understanding NA in R

7.1.1.1 Propagation of NA:

When performing operations on vectors or data frames, most functions in R will propagate `NA` values. This means that if even a single element within a vector is `NA`, the result of the operation will also be `NA`. For example:

```
# Example vector with missing values
gene_expression <- c(10, 15, NA, 25, 18, NA, 30)

# Performing a calculation on the vector
result <- gene_expression * 2

result
```

```
## [1] 20 30 NA 50 36 NA 60
```

As you can see, the presence of `NA` in the `gene_expression` vector leads to `NA` values in the result vector.

7.1.1.2 Handling NA in Summary Statistics

By default, `NA` values are omitted from summary statistics like the mean, median, or standard deviation. However, specialized functions often allow you to handle `NA` values explicitly using the `na.rm` parameter. For instance:

```
# Calculate the mean, ignoring NA values
mean_expression <- mean(gene_expression, na.rm = TRUE)
mean_expression
```

```
## [1] 19.6
```

```
# NA propagates
mean_expression <- mean(gene_expression)
mean_expression
```

```
## [1] NA
```

7.1.2 Identifying Missing Values

To identify missing values in your data, you can use the `is.na()` function. It returns a logical vector where `TRUE` indicates a missing value. For example:

```
# Identify missing values
missing_values <- is.na(gene_expression)
```

```
missing_values
```

```
## [1] FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE
```

7.1.3 remove missing values

Now, let's explore how to handle missing values in your gene expression dataset. Suppose you want to remove missing values to clean your data. You can do this using subsetting with the logical vector we created earlier:

```
# Remove missing values
clean_data <- gene_expression[!missing_values]

clean_data

## [1] 10 15 25 18 30
```

The `clean_data` vector now contains only the non-missing values from the original `gene_expression` vector.

7.1.4 real-life note

In real life, the data are usually messy. People use different ways to represent missing values. It can be -, N/A, NULL etc. see below

Null Values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently	R, Python, SQL	Best Option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good Option
N/A	Alternative form of NA, but often not compatible with software		Avoid
NULL	Can cause problem with data type		Avoid
None	Uncommon. Can cause problem with data type.	Python	Avoid
No data	Uncommon. Can cause problem with data type, Contains a space		Avoid
Missing	Uncommon. Can cause problem with data type.		Avoid
-+,..	Uncommon. Can cause problem with data type.		Avoid

More often, you have data in a dataframe. You can use the `table(df$column_name, useNA="ifany")` function to check all the possible values and you will spot the NAs.

See this old post from me https://rpubs.com/crazyhottommy/when_NA_is_not_NA

7.1.5 Conclusion

Handling missing values is a crucial skill in bioinformatics, as it ensures the reliability of your analyses. Whether you're calculating summary statistics or performing complex analyses, understanding how to work with missing data is an essential part of your bioinformatics toolkit.

7.2 Introduction to Statistical Tests and P-Values

We often need to determine whether there are significant differences between groups of data. Let's consider a scenario where we have two sets of cells: a control group and a treatment group (which could represent various treatments like chemical or radiation exposure). Our goal is to assess if a particular gene, let's call it Gene A, exhibits differential expression under treatment conditions. Each group has 12 replicates.

We typically start with a null hypothesis (H_0), which suggests that there is no difference in gene expression for Gene A after treatment, and an alternative hypothesis (H_1), which suggests that Gene A's expression changes after treatment.

Now, we perform a statistical test, like the t-test, on the averages of the two groups. If the test yields a p-value, say $p = 0.035$, and we've set a significance threshold (alpha) of 0.05, we compare these values. A p-value of 0.035 is less than 0.05, which leads us to reject the null hypothesis, concluding that Gene A's expression significantly changes after treatment.

But what does a p-value of 0.035 really mean?

A p-value starts with the assumption that the null hypothesis is true. In this context, a p-value of 0.035 means that under the null hypothesis, the probability of observing the observed difference in gene expression after treatment is 0.035, which is quite low. By selecting a significance level of 0.05, we establish a threshold for significance. When the p-value falls below this threshold, we reject the null hypothesis in favor of the alternative hypothesis. Thus, understanding the null hypothesis is crucial for interpreting the p-value's significance.

7.2.1 Practical Application of statistical test with R

Let's dive deeper into t-tests and their practical application using R:

7.2.1.1 Two-Sample t-Test

In bioinformatics, the two-sample t-test is a valuable tool for comparing the means of two groups when you have continuous data, such as gene expression levels. It assesses whether the difference between the two group means is statistically significant.

Let's consider an example with two sets of gene expression data, 'condition1' and 'condition2':

```
# Gene expression data for two conditions
condition1 <- c(12, 15, 20, 25)
condition2 <- c(8, 10, 7, 9)

# Perform a two-sample t-test
t_test_result <- t.test(condition1, condition2)
t_test_result

## 
## Welch Two Sample t-test
##
## data: condition1 and condition2
## t = 3.2426, df = 3.3053, p-value = 0.04159
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.6450826 18.3549174
## sample estimates:
## mean of x mean of y
##      18.0       8.5
```

In this example, the t-test yields a p-value of 0.04159. This p-value represents the probability of observing the difference in gene expression between 'condition1' and 'condition2' if the null hypothesis were true (i.e., no difference). Since 0.04159 is less than the typical significance level of 0.05, we reject the null hypothesis, indicating that there is a statistically significant difference in gene expression between the two conditions.

7.2.1.2 One-Sided t-Test

In some cases, you may be interested in whether one group's mean is greater than the other. This is where one-sided t-tests come into play.

```
# Perform a one-sided t-test (condition1 > condition2)
t.test(condition1, condition2, alternative = "greater")
```

```
## Welch Two Sample t-test
##
## data: condition1 and condition2
## t = 3.2426, df = 3.3053, p-value = 0.0208
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## 2.857331      Inf
## sample estimates:
## mean of x mean of y
##          18.0          8.5
```

In this example, the p-value is 0.0208. This test specifically checks if the gene expression in ‘condition1’ is greater than in ‘condition2’. By specifying the ‘alternative’ parameter as ‘greater,’ we focus our test on this direction. Again, the p-value is compared to the significance level to make a determination.

7.2.1.3 Non-Parametric Test

The t-test assumes that the data follows a normal distribution. If this assumption is not met, you can use non-parametric tests like the Wilcoxon rank-sum test.

```
wilcox.test(condition1, condition2)
```

```
## Wilcoxon rank sum exact test
##
## data: condition1 and condition2
## W = 16, p-value = 0.02857
## alternative hypothesis: true location shift is not equal to 0
```

In this example, the p-value is 0.02857. The Wilcoxon rank-sum test is valuable when your data doesn’t meet the normality assumption, making it a robust choice for analyzing gene expression or other biological data.

These t-tests are essential tools in bioinformatics for assessing the significance of differences between groups, helping researchers make data-driven decisions in various experimental scenarios.

7.2.2 Conclusion

Understanding statistical tests and p-values is fundamental in the field of bioinformatics. These tools empower researchers to determine whether observed

differences in data are statistically significant, enabling informed decisions and discoveries in the world of life sciences. The t-test, with its variations and non-parametric alternatives, is a powerful ally when comparing groups and assessing changes in gene expression or other biological phenomena. By grasping the significance of p-values and the interplay with null and alternative hypotheses, you can confidently interpret the results of your analyses, paving the way for meaningful insights and breakthroughs in bioinformatics.

7.3 Understanding R Packages

In R, packages are similar to complementary toolboxes that can significantly enhance your data analysis capabilities. Think of R as your basic toolkit, but packages are the specialized instruments that make complex tasks easier. In this lesson, we'll embark on a journey to understand what R packages are, how to find them, install them, and put them to use in your data analysis.

7.3.1 What is an R Package?

At its core, R provides a set of fundamental functions and features. However, when you dive into more complex tasks like genomics workflows or advanced statistical analysis, you'll quickly realize that creating everything from scratch is neither efficient nor practical. That's where R packages come in! These are like pre-made modules created by the R community, containing specialized functions and tools for various tasks.

We have learned functions in R. An R package is a collection of R functions, data, and code organized in a specific directory structure. It allows users to bundle related functionality, making it easier to distribute, share, and reuse code.

- Purpose: Packages provide a way to extend R's capabilities. They can contain functions, datasets, documentation, and more. Using packages enhances code organization, collaboration, and efficiency.

7.3.2 Why Use R Packages?

Imagine you need to perform complex statistical analysis, visualize data, or carry out genomics-related tasks. Instead of writing extensive code from scratch, R packages allow you to leverage the expertise of other researchers and developers. These packages encapsulate data processing routines, saving you time and effort.

7.3.3 How to use a package?

7.3.3.1 Installation

To use an R package, you first need to install it. Most packages are hosted on CRAN (Comprehensive R Archive Network), which is the primary repository for R packages. You can install a package using the `install.packages()` function. For example, if you want to install the popular `ggplot2` package for data visualization:

```
install.packages("ggplot2")
```

Once installed, you need to load the package into your current R session using the `library()` function:

```
library(ggplot2)
```

This action makes all the functions and tools from the `ggplot2` package available for use in your environment.

7.3.3.2 Installing specialized packages

For specialized fields like biology or bioinformatics, Bioconductor is a valuable resource. Bioconductor provides a vast array of genomics and statistics packages tailored to these domains. To install Bioconductor packages, you first need to check if you have the `BiocManager` package installed and then use it to install other packages. For instance, to install the `DESeq2` package for differential expression analysis:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

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

Remember, you typically only need to install a package once on your machine.

Sometimes, people host their R packages on `github`. To install an package from `github` you need an R package called `devtools`. Install it first

```
install.packages("devtools")
```

Then we can use it to install packages from `github`:

e.g., this package <https://github.com/immunogenomics/presto>

```
library(devtools)
install_github("immunogenomics/presto")
```

7.3.3.3 Putting Packages to Work

Once you've installed and loaded a package, you can harness its power. For example, let's say you want to perform differential expression analysis on your count data:

```
library(DESeq2)
counts <- matrix(c(10, 15, 5, 20, 8, 12, 18, 25, 30), nrow = 3, byrow = TRUE)

rownames(counts) <- c("Gene1", "Gene2", "Gene3")

colnames(counts) <- c("Sample1", "Sample2", "Sample3")

# create a 'samples' dataframe
samples <- data.frame(condition = c("Control", "Treatment", "Control"),
                      misc1 = c(1, 2, 1),
                      misc2 = c("A", "B", "C"))

dds <- DESeqDataSetFromMatrix(countData = counts,
                               colData = samples,
                               design = ~ condition)

res <- DESeq(dds)

res

## class: DESeqDataSet
## dim: 3 3
## metadata(1): version
## assays(4): counts mu H cooks
## rownames(3): Gene1 Gene2 Gene3
## rowData names(22): baseMean baseVar ... deviance maxCooks
## colnames(3): Sample1 Sample2 Sample3
## colData names(4): condition misc1 misc2 sizeFactor
```

In this example, we use the `DESeq2` package to handle the analysis. This package takes care of normalization and statistical calculations behind the scenes.

7.3.4 Embracing Errors

As you explore new packages, don't be alarmed if you encounter errors during installation. Error messages can be your allies, providing valuable information. If the message seems mysterious, copy it and try googling it, or seek help from online forums like biostars.org, seqanswers.com, or support.bioconductor.org. Error messages are part of the learning process, and you'll soon become skilled at deciphering them.

7.3.5 How to look for package docs within R studio?

Once a package is loaded, you can access its documentation using the `help()` function or the `?` operator. Simply provide the name of the function from the package as the argument to `help()` or follow `?` with the function name. For instance:

```
library(dplyr)
?select

## Help on topic 'select' was found in the following packages:
##          Package      Library
##    dplyr            /Library/Frameworks/R.framework/Versions/4.1/Resources/library/dplyr
##    AnnotationDbi   /Library/Frameworks/R.framework/Versions/4.1/Resources/library/AnnotationDbi
##
## Using the first match ...

help("select")

## Help on topic 'select' was found in the following packages:
##          Package      Library
##    dplyr            /Library/Frameworks/R.framework/Versions/4.1/Resources/library/dplyr
##    AnnotationDbi   /Library/Frameworks/R.framework/Versions/4.1/Resources/library/AnnotationDbi
##
## Using the first match ...
```

This command opens the documentation for the `select()` function from the `dplyr` package, providing details on its usage and arguments.

7.4. EXPLORING FUNCTIONS IN DIFFERENT PACKAGES: AVOIDING COLLISIONS AND ACCESSING THEM

7.3.6 Exploring Package Vignettes

Many R packages include vignettes, which are comprehensive documents detailing package functionalities, use cases, and examples. You can access these vignettes using the `browseVignettes()` function. Syntax:

```
browseVignettes(package_name)
```

For example:

```
browseVignettes("dplyr")
```

Executing this command opens a list of available vignettes for the `dplyr` package, allowing you to explore specific topics in detail.

7.3.7 Practical Tip

When looking for packages, you can search for specific ones related to your data type or analysis task. For instance, if you're working with DNA methylation data, you can search for packages like "DNA methylation bioconductor" in google.

7.3.8 Conclusion

In conclusion, R packages are your companions on the journey of data analysis. They allow you to stand on the shoulders of the community and streamline your work. As you advance in your data science or bioinformatics endeavors, you'll discover that these packages play a pivotal role in making your tasks more efficient and effective.

7.4 Exploring Functions in Different Packages: Avoiding Collisions and Accessing Them

One of the key skills you'll develop is harnessing the power of packages or libraries in R to extend its functionality. However, it's crucial to understand how to navigate and use functions when multiple packages offer functions with the same name, as collisions can occur. In this tutorial, we'll demystify this concept and show you how to access functions from different packages without loading them, enabling you to choose which one to use.

Imagine you're working on a project that requires both `dplyr` for data frame wrangling and `AnnotationDbi` for mapping gene IDs. You know that both

packages offer a `select` function. Here's where the challenge arises: if you load both libraries, the function in the later-loaded library will override the previous one. So, how do you access the `select` function from the desired package?

7.4.1 Using Double Colon (::)

The answer lies in the double colon (::) operator. You can specify the package along with the function name to avoid ambiguity. Let's look at some examples:

7.4.1.1 Data Selection

Suppose you are working with data frames and need to select specific columns. The `dplyr` package offers a `select` function for this task, but there is also a `select` function in the `AnnotationDbi` package. To avoid confusion, use the following approach:

```
# Select data columns using 'select' from 'dplyr' and 'AnnotationDbi' packages
selected_data_dplyr <- dplyr::select(data_frame, column1, column2)
selected_data_AnnotationDbi <- AnnotationDbi::select(x, keys, columns, keytype)
```

Here, we illustrate how to select specific columns using ‘`select`’ functions from both the `dplyr` and `AnnotationDbi` packages.

7.4.1.2 Data Reduction

Imagine you need to reduce a dataset to a single value. The `purrr` package offers a `reduce` function for this purpose, and so does the `GenomicRanges` package. Here's how to differentiate between them:

```
# Reduce data using 'reduce' from 'purrr' and 'GenomicRanges' packages
reduced_data_purrr <- purrr::reduce(data_list, reducer_function)
reduced_data_GenomicRanges <- GenomicRanges::reduce(gr)
```

For data reduction, we show how to use `reduce` functions from `purrr` and `GenomicRanges` packages.

7.4.1.3 Set Operations

In some cases, you may need to find the differences between two sets of data. The `GenomicRanges` package offers a `setdiff` function, but base R also has a `setdiff` function. Here's how to use them separately:

```
# Perform set difference using 'setdiff' from 'GenomicRanges' and base R
set_diff_GenomicRanges <- GenomicRanges::setdiff(gr1, gr2)
set_diff_base_R <- base::setdiff(set1, set2)
```

Lastly, for set operations, we demonstrate how to perform set differences using `setdiff` functions from `GenomicRanges` and base R.

These examples might seem abstract, but in the real world, you'll encounter situations where different packages offer functions with the same name but cater to distinct needs. By mastering the ‘`::`’ operator, you gain the ability to choose the right tool for the job, ensuring that your data analysis and manipulation are precise and tailored to your specific requirements.

7.4.2 Conclusion

Accessing functions from packages without loading them is a powerful technique that allows you to resolve function collisions and use the right function for your specific needs. It's a valuable skill for any data analyst or programmer working with R packages, ensuring that your code runs smoothly and produces accurate results.

7.5 Writing Custom Scripts

Writing organized and modular code is crucial for improving code readability, maintainability, and reusability. In this lesson, we will explore the process of creating, organizing, and loading scripts in R, with a focus on modular programming. We assume no prior computer science experience, making this lesson beginner-friendly.

This will be our hypothetical project structure:

```
my_awesome_project
  data
    data.csv
  scripts
    data_loading.R
    main_script.R
  results
    intermediate_table.csv
    figure.pdf
```

7.5.1 Your Working Directory

In R, the working directory is like a current location on your computer where R will look for files and where it will save files by default. It's important to set the working directory correctly because it helps R know where to find and store your data and scripts.

People tend to use `setwd()` to set their working directory.

```
setwd("/path/to/your/project")
```

However, this makes the analysis not reproducible. If people take your code and run in their own computer, the file paths are different.

Tip: For more advanced users, you may want to stay away from `setwd()` and use the `here()` function for reproducibility. Read this blog post for more information <https://www.tidyverse.org/blog/2017/12/workflow-vs-script/>

In RStudio, you can create a new R script by following these steps:

- Click on “File” in the top menu.
- Select “New File” and then “R Script.”

This will open a new R script file where you can write your code.

7.5.2 Writing Functions in Scripts

To make your code modular, you can define functions within your script files. Functions allow you to encapsulate specific tasks, making your code more readable and reusable. For example:

```
# data_loading.R
load_data <- function(file_path) {
  data <- read.csv(file_path)
  return(data)
}
```

This function, `load_data` reads data from a CSV file and returns it. Save this script as `data_loading.R` in the `scripts` folder.

7.5.3 Importing Functions from Scripts

Now that you've defined a function in a script, you can import it into your main script for use. To do this, use the `source()` function:

```
# main_script.R
source("scripts/data_loading.R")

# Now you can use the imported function
my_data <- load_data("data/my_data.csv")
```

The `source()` function reads and evaluates the code in the specified script file, making the `load_data` function available for use in `main_script.R`.

In this dummy example, the `data_loading` function is very simple, but in real life analysis, it can be more complicated: search on the web, crawl the files, download them to the local computer and then read into R.

7.6 Data I/O

In this lesson, we'll dive into the fundamental skills of handling data in R. Being proficient in inputting and outputting data is essential for automating data analysis workflows. We'll explore how to import, explore, and export files using R, all while cultivating good coding habits.

7.6.1 Built-In Functions

Built-in functions are pre-made tools in a programming language or software that do specific jobs without you having to create them from scratch or find extra tools elsewhere. They simplify common tasks in coding and are readily available for use.

7.6.1.1 Importing Data

When working with data analysis, one of the common challenges is reading data into R. Two common file formats for data storage are tab-separated/delimited values (TSV) and comma-separated values (CSV).

R comes to the rescue with functions designed to understand these languages - `read.csv()` and `read.table()`. These functions are like translators for R, helping it understand the data you provide.

Let's begin by importing an example gene expression file. You can download it from this link. We will use `read.csv()` to load the data into R.

```
# Import data from a CSV file
# by default the downloaded file will be in your Downloads folder
tcga_data <- read.csv("~/Downloads/TCGA_cancer_genes_expression.csv")
```

This code reads the data from the CSV file and stores it in the tcga_data dataframe. Now, we can manipulate and analyze this dataset.

7.6.1.2 Exploring Data

To understand the data, it's crucial to explore its structure. One way to do this is by examining the unique values in a specific column. In the following code, we count the occurrences of each cancer study type:

```
# Count the occurrences of each study type
table(tcga_data$study)
```

```
## 
##   ACC BLCA BRCA CESC CHOL COAD DLBC ESCA   GBM HNSC KICH KIRC KIRP LAML   LGG LIHC
##   79  433 1256  309   45  546   48  198   175  548   91  618  323  178  532  424
##   LUAD LUSC MESO    OV PAAD PCPG PRAD READ SARC SKCM STAD TGCT THCA THYM UCEC  UCS
##   601  555    87  430   183  187   558   177  265   473  453   156  572   122  589   57
##   UVM
##   80
```

7.6.1.3 Data Inspection

To get a quick look at the data, we can use the head() function, which displays the first six rows of the dataframe:

```
# Display the first 6 rows of the dataframe
head(tcga_data)
```

```
##                                     X TACSTD2 VTCN1 MUC1
## 1 43e715bf-28d9-4b5e-b762-8cd1b69a430e 0.7035937 0.00000000 0.67502205
## 2 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872 25.4360736 0.00000000 2.01525394
## 3 93b382e4-9c9a-43f5-bd3b-502cc260b886 1.5756197 0.00000000 0.90784666
## 4 1f39dadd-3655-474e-ba4c-a5bd32c97a8b 0.2702156 0.09099681 0.04293345
## 5 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb 0.4122814 0.00000000 0.11484380
## 6 85a86b91-4f24-4e77-ae2d-520f8e205efc 4.5469193 4.85973690 0.04208195
##   NECTIN4 FOLH1 FOLR1 CD276 MSLN CLDN6 ERBB2
## 1 0.08620727 7.213342 0.00000000 52.75981 0.06674445 0.09704962 1.879518
## 2 0.07279804 23.552286 0.12154673 78.78551 0.95554610 0.25458796 7.777976
```

```

## 3 0.69905270 2.853812 1.01000271 145.84399 0.04563568 0.25701910 2.905926
## 4 0.01652257 1.157070 0.27942068 48.45022 0.03154912 0.24746913 4.914280
## 5 0.03168398 2.408137 0.04922458 42.25592 0.26968788 0.12576720 1.494744
## 6 0.06828305 1.010411 0.02248965 20.63795 0.01336404 0.01823883 13.474689
##           MUC16        DLL3 CEACAM5      PVR      EPCAM      PROM1      CD24
## 1 0.0011479879 0.49589978          0 52.08113 4.521984 0.025311008 0.55036003
## 2 0.0008049670 2.52244014          0 40.87926 9.530414 0.023576862 9.67272890
## 3 0.0026190288 0.77074712          0 33.26727 42.358567 0.000000000 0.06939934
## 4 0.0051705741 0.10636402          0 28.26457 16.316524 0.007783431 0.84522244
## 5 0.0004894306 0.04483123          0 41.66776 12.529742 0.019204339 0.21369023
## 6 0.0000000000 0.01184285          0 30.18711 2.430109 0.043719865 4.95506593
##           EGFR        MET TNFRSF10B      tcga_tcga_barcode
## 1 1.286481 0.9320235 12.80547 TCGA-OR-A5KU-01A-11R-A29S-07
## 2 5.373307 8.0610999 31.46289 TCGA-P6-A50G-01A-22R-A29S-07
## 3 4.600918 0.1295387 65.57967 TCGA-OR-A5K5-01A-11R-A29S-07
## 4 3.010374 2.9728030 24.31636 TCGA-OR-A5K4-01A-11R-A29S-07
## 5 16.476552 19.7360055 21.11014 TCGA-OR-A5LP-01A-11R-A29S-07
## 6 2.010338 8.6087283 37.91574 TCGA-PK-A5H9-01A-11R-A29S-07
##   tcga_cgc_sample_sample_type study sample_type
## 1             Primary Tumor    ACC     cancer
## 2             Primary Tumor    ACC     cancer
## 3             Primary Tumor    ACC     cancer
## 4             Primary Tumor    ACC     cancer
## 5             Primary Tumor    ACC     cancer
## 6             Primary Tumor    ACC     cancer

```

This visual check helps us spot any potential issues early in the analysis. You may notice that the first column name is ‘X,’ which can happen when the data file has no column header. The output will display a table with the study abbreviations and their respective counts, providing insights into the dataset.

7.6.1.4 Data Anomalies

In some cases, data anomalies may arise, such as missing column names. For example, when opening the data in Excel, the first column may appear empty. In R, it defaults to using ‘X’ for that missing column name.

To inspect the last six rows of the data:

```
# Display the last 6 rows of the dataframe
tail(tcga_data)
```

```

##           X      TACSTD2      VTCN1      MUC1
## 11343 9506723f-9193-4d8e-bd97-8a0062ab2f9c 0.08154275 0.06388402 0.2653041
## 11344 3ee533bd-5832-4007-8f1f-439166256eb0 0.09602460 0.00000000 0.5017033

```

```

## 11345 a2c71c07-af0c-4016-808c-dfef458c91c7 0.11766953 0.06015085 1.0880740
## 11346 98b2a7f8-a7bd-4da2-8541-950e44d9acd7 0.00000000 0.00000000 0.6843479
## 11347 d3fc3968-b263-4756-bf7f-1941f70b04da 0.23376567 0.00000000 0.4993842
## 11348 f5b9b89b-6821-43ee-bcf8-623689d03af9 0.04404158 0.00000000 0.5674195
##          NECTIN4      FOLH1      FOLR1     CD276      MSLN      CLDN6
## 11343 0.027466444 0.55495200 0.00000000 16.30346 0.00000000 0.000000000
## 11344 0.000000000 0.28053151 0.00000000 30.29320 0.00000000 0.000000000
## 11345 0.070299290 1.48065011 0.39058212 52.87115 0.0000000 0.328483815
## 11346 0.011930801 0.09355531 1.69415825 28.02235 0.0000000 0.002753009
## 11347 0.011465502 0.38806433 0.04872839 26.70443 0.0000000 0.000000000
## 11348 0.006823029 0.20336014 0.37653301 27.12231 0.1554201 0.003539148
##          ERBB2      MUC16      DLL3      CEACAM5      PVR      EPCAM      PROM1
## 11343 9.768257      0 5.0197156 0.000000000 14.792782 0.67698452 0.04564542
## 11344 16.990612      0 13.6913305 0.000000000 13.244923 0.53035459 0.12887000
## 11345 17.496578      0 6.9048415 0.001147923 38.883462 0.32233902 0.07192491
## 11346 14.669438      0 3.8565185 0.000000000 8.523324 0.00000000 0.07007871
## 11347 8.645711      0 0.7416064 0.000000000 12.697792 0.00000000 0.10824816
## 11348 21.718607      0 5.5137596 0.000000000 11.473157 0.04096058 0.01835981
##          CD24      EGFR      MET TNFRSF10B      tcga_tcga_barcode
## 11343 0.11377251 0.1559625 67.49615 22.528273 TCGA-VD-A8K7-01B-11R-A405-07
## 11344 0.03138053 0.1998664 89.53855 5.072147 TCGA-VD-A8KB-01A-11R-A405-07
## 11345 0.08860096 1.0542988 200.75676 10.206628 TCGA-V4-A9EI-01A-11R-A405-07
## 11346 0.06745506 0.3034171 10.81844 3.908425 TCGA-V4-A9EY-01A-11R-A405-07
## 11347 0.10317872 0.1465101 49.09478 2.825917 TCGA-VD-AA8N-01A-11R-A405-07
## 11348 0.34742178 0.2299831 19.00013 3.644837 TCGA-V4-A9ET-01A-11R-A405-07
##          tcga_cgc_sample_sample_type study sample_type
## 11343             Primary Tumor    UVM    cancer
## 11344             Primary Tumor    UVM    cancer
## 11345             Primary Tumor    UVM    cancer
## 11346             Primary Tumor    UVM    cancer
## 11347             Primary Tumor    UVM    cancer
## 11348             Primary Tumor    UVM    cancer

```

This helps identify any anomalies or inconsistencies towards the end of the dataset.

7.6.1.5 Exporting Data

Once we've processed and analyzed the data, it's essential to save our results. We can export edited datasets using the `write.csv()` function. For instance, if we want to save the first six rows of the dataframe to a new CSV file:

```

# Export the first 6 rows to a new CSV file
write.csv(head(tcga_data), "~/Downloads/top_6_tcga.csv")

```

This code creates a new CSV file containing the selected data.

7.6.2 Real-World Applications

- Importing Data: Imagine you work in a research lab and need to analyze experimental results stored in CSV files. You can use R to import and process this data efficiently.
- Data Exploration: If you are a data analyst, you might use R to explore datasets, count occurrences of specific values, and gain insights to guide your analysis.
- Data Cleaning: Data often comes with anomalies or missing values. R can help you identify and address these issues before conducting statistical analyses.
- Data Export: Whether you're conducting research or generating reports, exporting your analysis results to share with colleagues or stakeholders is a common requirement.

7.6.3 Conclusion

In this lesson, we've covered the basics of handling data in R. You've learned how to import, explore, and export data, which are crucial skills for automating data analysis workflows. These skills will serve as a strong foundation for your journey into data science and analysis.

For more detailed information on these functions or any other aspect of R, you can always refer to the documentation by typing `?function_name` in R, such as `?read.table`, to access the help page for that function.

7.7 Best Practices in Modular Programming and Project Management

When using the R programming language, it is crucial to write code that is well-organized, reusable, and scalable. This guide will introduce you to best practices for modular programming and project management, making it accessible even if you have no prior experience in computer science. We will cover the following topics:

7.7.1 Introduction to Modular Programming

When you find yourself copying and pasting the same code more than twice, it's a clear signal that it's time to write a function. Functions in R allow you to encapsulate a set of instructions into a reusable block of code. This not only enhances code readability but also promotes reusability and maintainability.

7.7.1.1 Why Use Functions?

Imagine you are working on a data analysis project that involves multiple steps like loading data, data manipulation, and visualization. Instead of having one long script that combines all these tasks, you can break it down into smaller, more manageable functions. Here's an example:

```
get_count_data <- function() {
  counts <- read.csv("expression.csv")
  counts <- filter_zeros(counts)
  normalize(counts)
}
```

In this code, `get_count_data` is a function responsible for loading data, filtering out zeros, and normalizing the data. Now, when you need to access the data in your analysis, you can simply call this function:

```
counts <- get_count_data()
```

7.7.1.2 Benefits of Modular Functions

- **Code Reusability:** You can reuse these functions across different projects, saving time and effort. If you need to modify data loading or normalization logic, you only have to do it in one place.
- **Code Clarity:** Functions make your code more readable and maintainable. Each function encapsulates a specific part of your workflow, making it easier to understand.
- **Easy Debugging:** Smaller functions are easier to debug than long scripts. If there's an issue with data loading, you only need to focus on the `get_count_data` function.
- **Scalability:** As your project grows, you can easily add new functions to handle additional tasks without disrupting the existing code.

7.7.2 Breaking Down Your Project

To further enhance modularity in your data analysis projects, it's advisable to break down your project into separate files and directories. This structure simplifies project management and encourages a more organized workflow.

A typical project structure might look like this:

```
project
  data
  results
  scripts
    data_loading.R
    downstream.R
    preprocessing.R
    visualization.R
```

Here's what each component represents:

- data: This directory contains your data files or datasets.
- results: This is where you can store the results of your analysis, such as plots, tables, or reports.
- scripts: This directory is divided into separate files, each responsible for a specific part of your analysis.

Watch the video to understand how to set up a project fold structure and use the package `here` to avoid using `setwd()`.

In this Real-World RNAseq analysis Example, we

1. Loading Data and Preprocessing: By creating a `data_loading` function and `preprocessing` function, you can easily load, clean, and merge data from various sources in a consistent manner.
2. Exploratory Data Analysis (EDA): During EDA, you might need to create various visualizations for different aspects of your data. Separating visualization code into a dedicated script makes it easier to experiment with different plots and ensures a consistent look and feel across your project.
3. Statistical Modeling: When building predictive models, you can encapsulate the modeling process into a function called `downstream.R`. This allows you to apply the same model to different datasets or update the model with ease when new data becomes available.

Even better if you apply a consistent naming convention to all your scripts. In this case, you know the order of the scripts you used for each project.

```
project
  data
  results
  scripts
    01_data_loading.R
    02_preprocessing.R
    03_visualization.R
    04_downstream.R
```

Advantages of Project Structure:

- Clear Organization: By segregating your code into different files, you have a clear view of which script handles what aspect of your analysis.
- Focused Files: Each script file can focus on a specific task, making it easier to work on and understand.
- Easy Collaboration: When collaborating with others, this structure allows team members to work on different parts of the project simultaneously.
- Version Control: If you use version control systems like Git, having separate files for different tasks facilitates tracking changes and collaboration.

You should take advantage of the R project feature in Rstudio.

7.7.3 Conclusion

In summary, adopting modular programming practices and organizing your data analysis projects effectively in R not only enhances code quality but also simplifies project management. Functions serve as reusable building blocks, while a well-structured project directory keeps your work organized and maintainable. Embracing these principles will help you become a more efficient and effective data analyst, regardless of your level of experience.

7.8 Section complete

Congratulations on completing this section of the course! You've tackled complex topics such as handling missing values, explored statistical tests and P-values, and expanded your knowledge of R packages. Your understanding of functions within different packages has equipped you with strategies for efficient

access and avoidance of collisions. Not only have you improved your scripting skills, but you've also mastered effective data input/output management.

Your are making great progress, and these skills are essential for advanced data analysis. As you continue your journey in R, remember that these concepts form the foundation for further growth in modular programming and beyond.

Maintain your momentum as you move into the next section, where we'll build upon this foundation. If you have any questions, our community's Q&A section and lesson-specific comments are available for support.

Let's continue exploring together with curiosity and determination!

Chapter 8

Fundamental Data Structures in R

8.1 Named Vectors

A named vector allows you to assign names to individual elements within the vector. This may seem like a small feature, but it can greatly enhance your ability to organize, manipulate, and analyze data effectively.

8.1.1 What is a Named Vector?

In R, a vector can be named, meaning that each element within the vector can have a descriptive name associated with it. Think of it as a way to label your data. You can use the `names()` function to create a named vector by assigning names to the elements within the vector.

Let's start with a simple example:

```
# Create a numeric vector
expression_vec <- c(10, 25, 30, 12, 20)

# Assign names to the vector elements
names(expression_vec) <- c("gene1", "gene2", "gene3", "gene4", "gene5")

# View the named vector
expression_vec
```



```
## gene1 gene2 gene3 gene4 gene5
##    10    25    30    12    20
```

As you can see, each element now has a corresponding name, making it easier to identify and work with specific values in the vector.

8.1.2 Using Names to Subset a Vector

One of the key advantages of named vectors is the ability to subset them based on the names. This allows you to access specific elements of the vector easily.

For example, if you want to select only the values associated with “gene1” and “gene3,” you can do so like this:

```
# Subset the vector using names
selected_genes <- expression_vec[c("gene1", "gene3")]

# View the selected genes
selected_genes

## gene1 gene3
##      10     30
```

8.1.3 Named Vectors as Dictionary-Like Structures

Unlike some programming languages like Python, R doesn’t have a built-in dictionary data structure. However, named vectors can serve as a similar tool because they establish a one-to-one mapping between names and values.

Imagine you have gene expression measurements across various conditions that you want to analyze. You can create a named vector like this:

```
# Create a named vector from scratch
expression <- c(normal = 2.3, treated = 5.1, resistant = 3.8, sensitive = 1.2)

# View the named vector
expression

##      normal    treated   resistant   sensitive
##        2.3       5.1       3.8       1.2
```

Now, if you want to retrieve the expression value for the “resistant” condition, you can do so by using the name as follows:

```
# Access the expression value for "resistant" condition
expression["resistant"]

## resistant
##       3.8
```

8.1.4 Real-life example

In single-cell RNAseq analysis, you have tens of thousands of cells in the experiment and you cluster the cells into different clusters (cell types). Usually, you get the cluster id as numeric numbers: 1, 2, 3, 4, 5, ...

After you annotate the clusters with marker genes, you can give the clusters a specific name.

```
cells<- c("cell1", "cell2", "cell3", "cell4", "cell5", "cell6")

# the clustering algorithm assign each cell to cluster 1-4
clusters<- c(1, 2, 3, 2, 1, 4)

# we have annotated the clusters 1-4 as follows:
annotation<- c("CD4T", "CD8T", "NK", "B cell")

# create a named vector
names(annotation)<- c(1,2,3,4)

annotation

##      1      2      3      4
##  "CD4T"  "CD8T"  "NK"  "B cell"
```

We can use this named vector to re-annotate the original cells

```
annotation[clusters]

##      1      2      3      2      1      4
##  "CD4T"  "CD8T"  "NK"  "CD8T"  "CD4T"  "B cell"
```

We can then combine the cells and the new annotation as a data frame (which we will cover later).

```
data.frame(cell= cells, cell_annotation= annotation[clusters])

##   cell cell_annotation
## 1 cell1          CD4T
## 2 cell2          CD8T
## 3 cell3           NK
## 4 cell4          CD8T
## 5 cell5          CD4T
## 6 cell6          B cell
```

You see that the named vector can be very useful as a dictionary.

8.1.5 Reordering and Subsetting with Names

One of the remarkable features of named vectors is their flexibility. Even if the order of names differs from another vector, you can still use the names to reorder or subset the vector effectively.

Let's say you have another vector representing fold changes in gene expression:

```
# Create a vector for fold change of gene expression
folds <- c(resistant = 1.1, sensitive = 4.2, normal = 1.3, treated = 2.1)

# View the fold change vector
folds
```

```
## resistant sensitive    normal   treated
##        1.1         4.2         1.3         2.1
```

Notice that the order of conditions is different from the expression vector. However, you can use the names from the expression vector to reorder or subset the fold change vector:

```
# Reorder the fold change vector using names from the expression vector
folds[names(expression)]
```

```
##      normal   treated resistant sensitive
##        1.3         2.1         1.1         4.2
```

8.1.6 Conclusion

Named vectors may seem simple, but they offer a valuable way to organize and manipulate your data, serving as a powerful tool for tasks like subsetting, indexing, and organizing information. This beginner-friendly guide has introduced you to the concept of named vectors and demonstrated their practical use in real-world scenarios. As you delve deeper into R, you'll find that mastering this fundamental feature will greatly enhance your data analysis capabilities.

8.2 Lists

Lists in R are versatile data structures that can hold various types of elements, including both numeric values and character strings, and even elements of different lengths. In this section, we will explore the concept of lists, understand how to create and manipulate them, and learn different ways to access their elements.

8.2.1 Introduction to Lists

Unlike matrices and vectors, which can only store elements of the same type, lists can accommodate a mix of numeric and character elements with different lengths. This flexibility makes lists a powerful tool for organizing and storing complex data structures.

8.2.2 Creating a List

Imagine you are conducting a series of biological experiments over multiple days. For each day, you collect various data, including numeric measurements (e.g., gene expression levels) and descriptive summaries (e.g., experimental conditions). Lists can be used to store this data efficiently. Each list element can represent a day's data, containing both numeric vectors and character strings to capture the diverse information for each experiment:

```
results <- list(p_values = c(0.01, 0.56),
                 summary = "Non-significant",
                 experiments = c("day1", "day2"))
results

## $p_values
## [1] 0.01 0.56
##
## $summary
## [1] "Non-significant"
##
## $experiments
## [1] "day1" "day2"
```

When we print the `results` list, you will notice that each element is preceded by a `$` sign, indicating its name.

8.2.3 Accessing List Elements

You can access list elements using various methods:

8.2.3.1 Using `$` Notation

To access an element by its name, you can use the `$` notation:

```
results$experiments

## [1] "day1" "day2"

results$summary

## [1] "Non-significant"
```

8.2.3.2 Using Single Brackets

Using single brackets [] will return a list:

```
results[1]
```

```
## $p_values
## [1] 0.01 0.56
```

####Slicing the List You can slice a list by specifying a range of indices:

```
results[1:2]
```

```
## $p_values
## [1] 0.01 0.56
##
## $summary
## [1] "Non-significant"
```

8.2.3.3 Using [[]] Double Brackets

To access the actual element rather than a list, use double brackets [[]]:

```
results[[1]]
```

```
## [1] 0.01 0.56
```

```
results[["p_values"]]
```

```
## [1] 0.01 0.56
```

8.2.3.4 Using Names

You can also access elements by their names using single brackets and double brackets:

```
# returns a list
results["p_values"]
```

```
## $p_values
## [1] 0.01 0.56
```

```
# return the element
results[["p_values"]]
```

```
## [1] 0.01 0.56
```

Tip: In RStudio, when you type \$ after the list, it will display a dropdown menu showing all the available names of the list elements, making it easier to access them.

8.2.4 Conclusions

Lists are essential for managing diverse data structures efficiently in R. They allow you to store and organize data with different characteristics, making them a valuable tool for data manipulation and analysis in various applications.

8.3 Dataframes

Data frames are one of the fundamental data structures in R, and they play a pivotal role in various data analysis tasks. In the realm of biology research, data frames are exceptionally useful for organizing and manipulating biological data efficiently. In this tutorial, we will explore the basics of data frames, including their creation, column access, and subsetting.

8.3.1 What is a Data Frame?

A data frame is a two-dimensional tabular data structure in R. It is similar to a spreadsheet or a database table, where rows and columns intersect to form a grid-like structure. In a data frame, each row can hold different types of data, such as numbers, text, or factors, but all columns must have the same number of rows and each column should have the same data type.

8.3.2 Creating a Data Frame

To create a data frame, you can use the `data.frame()` function. Here's an example of creating a simple data frame for patient data in a biology study.

```
patient_data <- data.frame(
  id = c("P1", "P2"),
  age = c(42, 36),
  status = c("normal", "tumor")
)
```

In this example, we have three vectors (`id`, `age`, and `status`) that are combined into a data frame. The variable names of the vectors become the column names of the data frame. You can see the resulting data frame by simply typing `patient_data`.

```
patient_data

##   id age status
## 1 P1  42 normal
## 2 P2  36 tumor
```

8.3.3 Accessing Columns

You can access individual columns within a data frame using the `$` sign, which is similar to accessing elements in a list. For instance:

```
patient_data$id
```

```
## [1] "P1" "P2"
```

This command will retrieve the `id` column from the `patient_data` data frame.

Similarly, you can access the `status` column by using `patient_data$status`.

8.3.4 Subsetting a Data Frame

Subsetting a data frame means extracting specific rows and columns based on your analysis needs. You can use square brackets `[row_indices, column_indices]` to subset or slice the data frame, similar to subsetting a matrix.

For example, to select the first row and the first and third columns from the `patient_data` data frame:

```
patient_data[1, c(1, 3)]
```

```
##   id status
## 1 P1 normal
```

This command returns a subset of the data frame with the `id` and `status` columns for the first row.

8.3.5 Conclusion

In biology research, data frames are invaluable for organizing and analyzing diverse datasets. Imagine you have a dataset with patient information, where each row represents a patient and columns contain attributes like age, gender, and diagnosis. You could use data frames to filter patients by specific criteria, calculate summary statistics, or create visualizations to gain insights into your biological data.

Data frames in R serve as the backbone for handling and exploring data in a structured and meaningful way, making them an essential tool for any biologist or data scientist working in the field of life sciences.

8.4 How to Represent Categorical Data in R? Understanding Factors

In R, factors are a fundamental data type used to represent categorical data. Factors are akin to biological categories, such as gene types, experimental conditions, or mutation classes. They provide a structured way to handle and analyze qualitative data. In this tutorial, we will explore factors in R, starting from their creation, manipulation, and practical applications.

8.4.1 Creating Factors

Let's begin with a hypothetical scenario involving phenotypes: wild type (WT), mutant (Mut), and heterozygote (Het). We can create a factor called `phenotypes` to represent these categories:

```
phenotypes <- factor(c("WT", "Mut", "Mut", "WT", "Het"))
```

When you print `phenotypes`, you'll see the categories along with their associated levels:

```
phenotypes
```

```
## [1] WT  Mut Mut WT  Het
## Levels: Het Mut WT
```

In this example, the factor phenotypes has three levels: “Het,” “Mut,” and “WT.”

8.4.2 Changing levels

By default, the levels are ordered alphabetically. You can use the `relevel` to change it

```
relevel(phenotypes, ref = "WT")
```

```
## [1] WT  Mut Mut WT  Het
## Levels: WT Het Mut
```

Now, the WT becomes the base level.

8.4.3 Converting Factors

Factors can be converted to characters or numeric values. When you convert them to numbers, R assigns a numerical value to each category based on their order in the factor levels. The lowest level gets assigned the number 1, and the highest level gets the highest number.

8.4.3.1 As character

This R command converts the phenotypes factor into a character vector. In other words, it changes the factor’s categorical labels into text.

```
as.character(phenotypes)
```

```
## [1] "WT"  "Mut" "Mut" "WT"  "Het"
```

8.4.3.2 Converting to numbers

This R command converts the `phenotypes` factor into a numeric vector. It assigns a numerical value to each category based on their order in the factor levels.

```
as.numeric(phenotypes)
```

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

In this case, “WT” corresponds to 3, “Mut” to 2, and “Het” to 1.

8.4.4 Customizing Factor Levels

You can customize the order of factor levels to control how they are converted to numbers. Suppose you want “Mut” to be assigned the lowest value. You can specify the desired order when creating the factor:

```
phenotypes <- factor(c("WT", "Mut", "Mut", "WT", "Het"),
                      levels= c("Mut", "Het", "WT"))
```

Now, when you convert it to numbers, “Mut” will be 1:

```
as.numeric(phenotypes)
```

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

8.4.5 Creating Factors from Character Vectors

You can also create a factor from a character vector and explicitly specify the desired factor levels. For example:

```
phenotypes <- c("WT", "Mut", "Mut", "WT", "Het")
phenotypes <- factor(phenotypes, levels= c("Mut", "Het", "WT"))
```

```
phenotypes
```

```
## [1] WT  Mut Mut WT  Het
## Levels: Mut Het WT
```

This approach allows you to control the factor levels directly.

8.4.6 Subsetting Factors

When you subset a factor, the factor levels are preserved. For instance:

```
phenotypes[1:2]
```

```
## [1] WT  Mut
## Levels: Mut Het WT
```

Even though we only extracted elements 1 and 2, the factor levels “Mut,” “Het,” and “WT” are still present.

8.4.7 Removing Unused Levels

Sometimes, you may want to remove unused factor levels to keep your data clean. You can achieve this using the `droplevels()` function. For example:

```
droplevels(phenotypes[1:2])
```

```
## [1] WT  Mut
## Levels: Mut WT
```

In this case, the “Het” level was removed because it was not present in the subset.

8.4.8 Conclusion

Factors are invaluable in various fields of research, including biology and genetics, where they are used to represent and analyze categorical data like gene types, mutation classes, or experimental conditions. Understanding factors and their manipulation is essential for accurate data analysis and interpretation.

Statistical modeling uses factors heavily. For example, when doing differential gene expression analysis, the factor levels determine which group is used as the baseline. Also, when plotting figures, the order of the geometric objects (e.g, the boxes in the boxplot) is ordered by the factor levels. Changing the factor levels will change the look of the plots. We will see an example in the later data visualization lecture.

In summary, factors in R provide a structured way to work with categorical data, allowing you to control the order of levels and efficiently manage and analyze information. Mastery of factors is a crucial skill for any data scientist or researcher working with categorical data in R.

8.5 Section Complete

Congratulations on completing the section!

You've learned about named vectors, lists, data frames, and factors. Understanding these basics is essential for working with data effectively. This knowledge will be valuable as you tackle more advanced data analysis tasks.

As you move on to the next section, remember to use our Q&A section and comments for any questions you have. Let's keep exploring and learning together in our R programming journey!

Chapter 9

Introduction to the tidyverse ecosystem

9.1 What is the Tidyverse?

In this lesson, we will explore the Tidyverse, a powerful collection of R packages designed to simplify and streamline the data science workflow. The Tidyverse is particularly well-suited for beginners and professionals alike, as it encourages a structured and intuitive approach to data manipulation and visualization.



The Tidyverse is a comprehensive ecosystem of R packages that share common principles for data manipulation and visualization. It encourages the use of tidy data, a specific data structure that makes data analysis more straightforward. Tidy data arranges observations in rows and variables in columns, making it easier to work with and visualize.

9.2 Key Tidyverse Packages

Let's dive into some of the essential packages within the Tidyverse and understand how they can be beneficial in data analysis.

9.2.1 dplyr

Explore `dplyr` docs here.

The core of the Tidyverse is the `dplyr` package. It provides a set of functions for data manipulation, including filtering, summarizing, and transforming data frames. Here's a simple example of how you might use `dplyr` to filter data:

```
# install.packages("dplyr")
# Load the dplyr package
library(dplyr)

# Create a data frame
data <- data.frame(
  Name = c("Alice", "Bob", "Charlie"),
  Age = c(25, 30, 22)
)

data

##      Name Age
## 1    Alice 25
## 2     Bob 30
## 3 Charlie 22

# Filter the data to select individuals older than 25
filtered_data <- data %>%
  filter(Age > 25)

# View the filtered data
filtered_data

##      Name Age
## 1     Bob 30
```

In this example, we used `filter()` to select rows where the “Age” column is greater than 25.

9.2.2 A note on the pipe

You just saw the `%>%` operator. It is also called a pipe.

`%>%` is from the `magrittr` package and when you load the `dplyr` package, the `%>%` will be available for you to use.

The R language has a new, built-in pipe operator as of R version 4.1: `|>`.

```
mtcars %>%
  head()
```

```
##          mpg cyl disp hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6 160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6 160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4 108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive 21.4   6 258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360 175 3.15 3.440 17.02  0  0    3    2
## Valiant       18.1   6 225 105 2.76 3.460 20.22  1  0    3    1
```

```
mtcars |>
  head()
```

```
##          mpg cyl disp hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6 160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6 160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4 108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive 21.4   6 258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360 175 3.15 3.440 17.02  0  0    3    2
## Valiant       18.1   6 225 105 2.76 3.460 20.22  1  0    3    1
```

Think of `%>%` and `|>` in R the same as `|` in the unix commands: the output of the previous command is the input of the next command.

The pipe operator `%>%` is a very useful tool for writing efficient, easy-to-read code in R. You can use as many pipes as you want. I usually build the pipes one by one by looking at the output.

```
mtcars %>%
  head() %>%
  tail(n=3)
```

```
##          mpg cyl disp hp drat    wt  qsec vs am gear carb
## Hornet 4 Drive 21.4   6 258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360 175 3.15 3.440 17.02  0  0    3    2
## Valiant       18.1   6 225 105 2.76 3.460 20.22  1  0    3    1
```

This means print out the first 6 rows and then take the last 3 rows.

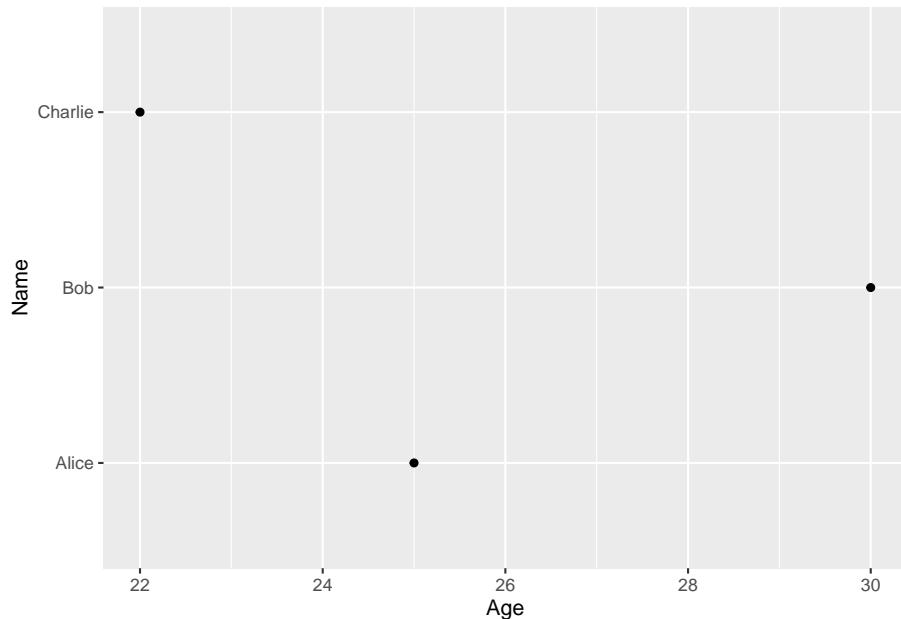
9.2.3 ggplot2

Explore `ggplot2` docs here.

The `ggplot2` package is the go-to choice for data visualization in the Tidyverse. It follows the “grammar of graphics” approach, allowing you to build complex plots layer by layer. Here’s a basic example:

```
# install.packages("ggplot2")
# Load the ggplot2 package
library(ggplot2)

# Create a scatter plot
ggplot(data, aes(x = Age, y = Name)) +
  geom_point()
```



In this code, we used `ggplot()` to set up the plot and `geom_point()` to add scatterplot points to the canvas.

9.2.4 tidyverse

Explore `tidyverse` docs here.

The `tidyverse` package assists in reshaping data between **wide** and **long** formats, which is often necessary for analysis and visualization. Suppose you have a

dataset with columns representing different years and want to convert it to a long format for easier analysis. `tidyverse` can help with this transformation.

Wide Format				Long Format		
Team	Points	Assists	Rebounds	Team	Variable	Value
A	88	12	22	A	Points	88
B	91	17	28	A	Assists	12
C	99	24	30	A	Rebounds	22
D	94	28	31	B	Points	91
				B	Assists	17
				B	Rebounds	28
				C	Points	99
				C	Assists	24
				C	Rebounds	30
				D	Points	94
				D	Assists	28
				D	Rebounds	31

9.2.5 `readr`

Explore `readr` docs here.

`readr` is a fast and efficient package for reading **tabular** data into R. It's faster than the base R functions like `read.csv()`. When working with large datasets, this can significantly speed up your data loading process.

Tabular data are rectangular:

Tabular Data

columns = attributes for those observations

The diagram illustrates a tabular dataset with a red bracket on the left labeled "Rows = observations" pointing to the vertical axis of the table. A red bracket at the top labeled "columns = attributes for those observations" points to the horizontal axis of the table. The table itself has a light blue header row and contains nine data rows, each representing a player's statistics.

Player	Minutes	Points	Rebounds	Assists
A	41	20	6	5
B	30	29	7	6
C	22	7	7	2
D	26	3	3	9
E	20	19	8	0
F	9	6	14	14
G	14	22	8	3
I	22	36	0	9
J	34	8	1	3

9.2.6 tibble

Explore `tibble` docs here.

the rectangular data will be read into R as a data frame. `tibble` provides an improved data frame structure that addresses some issues with traditional R data frames. It displays data more neatly and provides better compatibility with the Tidyverse functions.

9.2.7 stringr

Explore `stringr` docs here.

The `stringr` package offers a range of string manipulation functions that are easier to use and more consistent than the base R functions. It's particularly handy when dealing with text data, such as cleaning and formatting strings.

9.2.8forcats

Explore `forcats` docs here.

`forcats` is designed for handling factor variables, which are categorical variables with predefined levels. It provides tools for changing the order of levels and combining levels when necessary.

9.2.9 Real-World Applications

To illustrate the practical utility of the Tidyverse, consider the following scenarios:

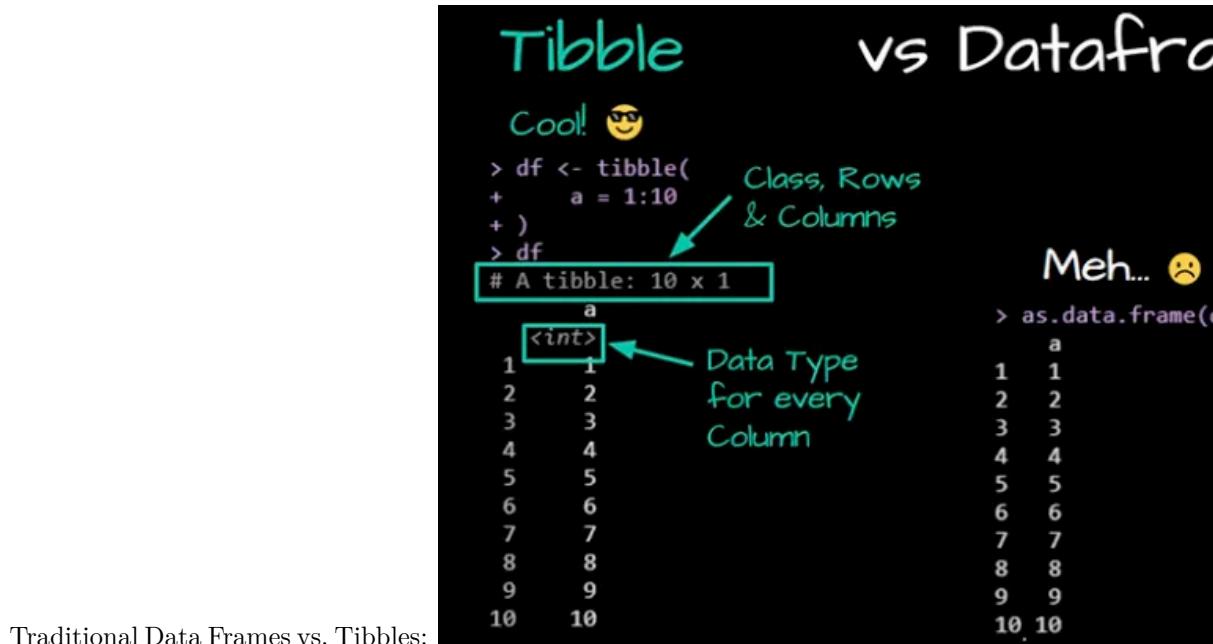
1. Data Import: `readr` helps you efficiently import data from various sources, such as CSV files, Excel spreadsheets, or even web-based data.
2. Data Cleaning: When working with messy data, `tidyverse` can assist in reshaping and cleaning data for analysis.
3. Data Exploration: When you receive a dataset, you can use `dplyr` to quickly filter, summarize, and explore the data, helping you understand its characteristics.
4. Data Visualization: `ggplot2` allows you to create stunning visualizations, making it easier to convey insights to others. For instance, you can create bar charts, scatter plots, and histograms.
5. Working with Factors: `forcats` simplifies tasks like reordering factor levels for better visual representation.

By mastering the Tidyverse, you'll have a powerful set of tools at your disposal for all stages of data analysis, from data preprocessing to visualization and modeling. Happy coding!

9.3 Tibble and Readr - Modern Data Structures in R

In this lesson, we will explore two essential concepts in data manipulation with R: tibbles and the `readr` package. These tools are designed to make working with data more efficient and user-friendly, especially when dealing with large datasets. We'll see why tibbles are preferred over traditional data frames and how to read and manipulate data using the `readr` package.

9.3.1 Tibbles: A Better Way to Store Data



Traditional Data Frames vs. Tibbles:

In R, data frames are widely used for storing and manipulating data. However, tibbles offer several advantages over traditional data frames:

1. Cleaner Printing: Tibbles offer a more organized way to present your data. When you print a tibble, it only displays the first 10 rows and as many columns as can fit on your screen. This ensures a concise overview of your data, which is particularly helpful for large datasets. In contrast, traditional data frames tend to print all rows and columns by default, leading to overwhelming output.
2. Structured Indexing: When you subset a tibble using [], it returns another tibble that retains the original data's structure. This means you still have clear column names and data types. In contrast, data frames return vectors, which provide less information about the structure of the data subset. You need to specify `drop = FALSE` to retain as a data frame and we see examples in previous lectures.
3. Preservation of Variable Types: Tibbles respect the original data types of your variables, even after subsetting. This is important because it ensures that your data remains consistent. In some cases, data frames may convert variables to factors or matrices when you subset them, potentially causing unexpected issues.

4. String Handling: In tibbles, character vectors remain as characters, maintaining the integrity of your data. However, data frames may automatically convert character vectors to factors by default. This behavior can lead to unexpected changes in your data and, subsequently, your analysis.
5. List-Columns: One unique feature of tibbles is their ability to directly contain list-columns. List-columns allow you to store lists within your tibble, providing a convenient way to work with complex data structures. Data frames do not offer this feature, which can limit your ability to represent certain types of data effectively.
6. Enhanced Output: Tibbles enhance the print display by showing data types, highlighting missing values, and truncating output for better readability. This additional information helps you understand your data at a glance, making it easier to spot potential issues or trends.

9.3.2 Rownames and Tibbles

In regular data frames, you can assign and work with rownames, which are helpful for labeling and referencing rows. However, tibbles do not support rownames, so you want to add another column that contains the row ids.

9.3.3 Reading Data with readr

To work with data effectively, you first need to import it into R. The `readr` package provides a powerful toolset for reading various data file formats. Let's see how to read data using `readr`.

9.3.3.1 Reading CSV Data

Download the TCGA gene expression data (CSV file) to your working directory or specify the correct file path.

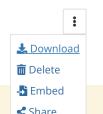
Download the file at <https://osf.io/yeun5>

Click the Download:

[master_real_world_bioinformatics_in_R_files](#)

TCGA_cancer_genes_expression.csv

Table is too large to render.



Suppose we have a CSV file named “TCGA_cancer_genes_expression.csv” in the Downloads folder, to read it using `readr::read_csv`, follow these steps:

```
# Load the readr package (if not already loaded)
library(readr)

# Read the CSV data into a tibble
tcga_data <- readr::read_csv("~/Downloads/TCGA_cancer_genes_expression.csv")

# Display the first few rows of the tibble
head(tcga_data)

## # A tibble: 6 x 25
##   ...1     TACSTD2    VTCN1    MUC1 NECTIN4 FOLH1  FOLR1 CD276   MSLN CLDN6 ERBB2
##   <chr>     <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 43e715bf~  0.704 0     0.675  0.0862 7.21 0     52.8 0.0667 0.0970 1.88
## 2 1a5db9fc~  25.4 0     2.02   0.0728 23.6 0.122 78.8 0.956 0.255 7.78
## 3 93b382e4~  1.58 0     0.908  0.699  2.85 1.01 146. 0.0456 0.257 2.91
## 4 1f39dadd~  0.270 0.0910 0.0429 0.0165 1.16 0.279 48.5 0.0315 0.247 4.91
## 5 8c8c09b9~  0.412 0     0.115  0.0317 2.41 0.0492 42.3 0.270 0.126 1.49
## 6 85a86b91~  4.55 4.86  0.0421 0.0683 1.01 0.0225 20.6 0.0134 0.0182 13.5
## # i 14 more variables: MUC16 <dbl>, DLL3 <dbl>, CEACAM5 <dbl>, PVR <dbl>,
## # EPCAM <dbl>, PROM1 <dbl>, CD24 <dbl>, EGFR <dbl>, MET <dbl>,
## # TNFRSF10B <dbl>, tcga_tcga_barcode <chr>,
## # tcga_cgc_sample_sample_type <chr>, study <chr>, sample_type <chr>
```

Here, we load the `readr` package, use `read_csv` to read the CSV file, and store the data in the `tcga_data` tibble. Finally, we use `head` to display the first few rows of the tibble.

note that the first column’s name changes to “...1,” which is a default behavior when column names are missing in the source data.

9.3.3.2 Understanding the Output

The output shows the data in a neat tabular format. Column names are displayed at the top, followed by rows of data. Each column’s data type is indicated (e.g., `dbl` for double-precision floating point numbers or `chr` for character).

9.3.4 Using Tibbles with Imported Data

When working with tibbles, imported data maintains its format and advantages over data frames. Here’s how you can convert the imported data into a tibble:

```
# Load the dplyr package for tibble conversion
library(dplyr)

# read in the data using built-in read.csv
tcga_data<- read.csv("~/Downloads/TCGA_cancer_genes_expression.csv")

# regular dataframe, and the first column name is X if it is empty
head(tcga_data)
```

	X	TACSTD2	VTCN1	MUC1				
## 1	43e715bf-28d9-4b5e-b762-8cd1b69a430e	0.7035937	0.00000000	0.67502205				
## 2	1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872	25.4360736	0.00000000	2.01525394				
## 3	93b382e4-9c9a-43f5-bd3b-502cc260b886	1.5756197	0.00000000	0.90784666				
## 4	1f39dadd-3655-474e-ba4c-a5bd32c97a8b	0.2702156	0.09099681	0.04293345				
## 5	8c8c09b9-ec83-45ec-bc4c-0ba92de60acb	0.4122814	0.00000000	0.11484380				
## 6	85a86b91-4f24-4e77-ae2d-520f8e205efc	4.5469193	4.85973690	0.04208195				
##	NECTIN4	FOLH1	FOLR1	CD276	MSLN	CLDN6	ERBB2	
## 1	0.08620727	7.213342	0.00000000	52.75981	0.06674445	0.09704962	1.879518	
## 2	0.07279804	23.552286	0.12154673	78.78551	0.95554610	0.25458796	7.777976	
## 3	0.69905270	2.853812	1.01000271	145.84399	0.04563568	0.25701910	2.905926	
## 4	0.01652257	1.157070	0.27942068	48.45022	0.03154912	0.24746913	4.914280	
## 5	0.03168398	2.408137	0.04922458	42.25592	0.26968788	0.12576720	1.494744	
## 6	0.06828305	1.010411	0.02248965	20.63795	0.01336404	0.01823883	13.474689	
##	MUC16	DLL3	CEACAM5	PVR	EPCAM	PROM1	CD24	
## 1	0.0011479879	0.49589978		0	52.08113	4.521984	0.025311008	0.55036003
## 2	0.0008049670	2.52244014		0	40.87926	9.530414	0.023576862	9.67272890
## 3	0.0026190288	0.77074712		0	33.26727	42.358567	0.000000000	0.06939934
## 4	0.0051705741	0.10636402		0	28.26457	16.316524	0.007783431	0.84522244
## 5	0.0004894306	0.04483123		0	41.66776	12.529742	0.019204339	0.21369023
## 6	0.0000000000	0.01184285		0	30.18711	2.430109	0.043719865	4.95506593
##	EGFR	MET	TNFRSF10B			tcga.tcga_barcode		
## 1	1.286481	0.9320235	12.80547	TCGA-OR-A5KU-01A-11R-A29S-07				
## 2	5.373307	8.0610999	31.46289	TCGA-P6-A50G-01A-22R-A29S-07				
## 3	4.600918	0.1295387	65.57967	TCGA-OR-A5K5-01A-11R-A29S-07				
## 4	3.010374	2.9728030	24.31636	TCGA-OR-A5K4-01A-11R-A29S-07				
## 5	16.476552	19.7360055	21.11014	TCGA-OR-A5LP-01A-11R-A29S-07				
## 6	2.010338	8.6087283	37.91574	TCGA-PK-A5H9-01A-11R-A29S-07				
##	tcga.cgc_sample_sample_type	study	sample_type					
## 1		Primary	Tumor	ACC	cancer			
## 2		Primary	Tumor	ACC	cancer			
## 3		Primary	Tumor	ACC	cancer			
## 4		Primary	Tumor	ACC	cancer			
## 5		Primary	Tumor	ACC	cancer			
## 6		Primary	Tumor	ACC	cancer			

```
# Convert the data to a tibble
tcga_data <- tcga_data %>%
  tibble::as_tibble()

# Display the first few rows of the tibble
head(tcga_data)

## # A tibble: 6 x 25
##   X          TACSTD2  VTCN1    MUC1 NECTIN4 FOLH1   FOLR1 CD276   MSLN  CLDN6 ERBB2
##   <chr>      <dbl>    <dbl>    <dbl>  <dbl>    <dbl>    <dbl>  <dbl>    <dbl>  <dbl>  <dbl>
## 1 43e715bf~  0.704    0       0.675  0.0862   7.21    0      52.8   0.0667  0.0970  1.88
## 2 1a5db9fc~  25.4     0       2.02   0.0728  23.6    0.122   78.8   0.956   0.255   7.78
## 3 93b382e4~  1.58     0       0.908  0.699    2.85    1.01    146.    0.0456  0.257   2.91
## 4 1f39dadd~  0.270    0.0910  0.0429  0.0165   1.16    0.279   48.5   0.0315  0.247   4.91
## 5 8c8c09b9~  0.412    0       0.115   0.0317  2.41    0.0492  42.3   0.270   0.126   1.49
## 6 85a86b91~  4.55     4.86   0.0421  0.0683  1.01    0.0225  20.6   0.0134  0.0182  13.5
## # i 14 more variables: MUC16 <dbl>, DLL3 <dbl>, CEACAM5 <dbl>, PVR <dbl>,
## # EPCAM <dbl>, PROM1 <dbl>, CD24 <dbl>, EGFR <dbl>, MET <dbl>,
## # TNFRSF10B <dbl>, tcga_tcga_barcode <chr>,
## # tcga_cgc_sample_sample_type <chr>, study <chr>, sample_type <chr>
```

By using the `%>%` pipe operator and `tibble::as_tibble()`, we convert the data into a tibble while preserving its structure and advantages.

The output remains in a similar format as before, with clear column names and data types.

9.3.5 Note on reading Excel files

We still deal with a lot of spreadsheets, to read in Excel files, take a look at the `readxl` package.

9.3.6 Conclusion

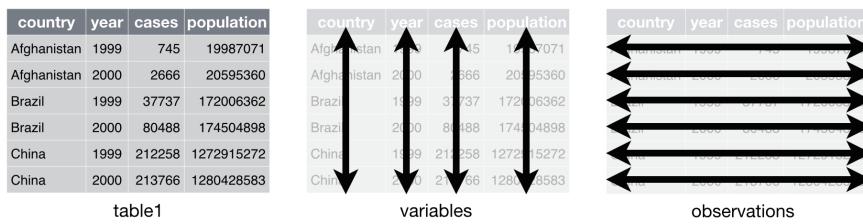
In this lesson, we learned about `tibbles`, a modern data structure in R that offers several advantages over traditional data frames. We also explored how to read and manipulate data using the `readr` package, converting imported data into tibbles for efficient data analysis. Tibbles and `readr` are valuable tools for data scientists and analysts working with R, making data manipulation and exploration more user-friendly and efficient.

9.4 The tidy data format

In this lesson, we will delve into the essential data cleaning and tidying operations in R using the powerful `dplyr` package. Tidying data is a crucial step in data analysis, especially when dealing with real-world data that can be messy and inconsistent. We will explore the concept of tidy data and learn about the distinction between long and wide dataset formats.

9.4.1 The Importance of Tidy Data

Real-world data is often messy, scattered across various files, missing metadata, and inputted in inconsistent formats. Before we can effectively analyze this data, we need to bring it all together into one structured table and resolve any issues. This process of data ingestion and standardization is known as data tidying.



The tidyverse packages, including `dplyr`, work seamlessly with tidy data. As defined by Hadley Wickham in R for Data Science, tidy data adheres to three interrelated rules:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

Why should we make our data tidy? There are both general and specific advantages:

- General Advantage: Using a consistent data structure makes it easier to learn and work with data manipulation tools because they follow a uniform pattern.
- Specific Advantage: Placing variables in columns allows R's vectorized nature to shine. Most built-in R functions work efficiently with vectors of values. This natural alignment between tidy data and R's capabilities makes data transformation feel particularly intuitive.

9.4.2 Long Format vs. Wide Format

Before we dive into the practical aspects of tidying data, let's understand the difference between long format and wide format for a dataframe. We'll use an example with the dplyr package:

```
library(tidyverse)

head(table4a)

## # A tibble: 3 x 3
##   country    `1999` `2000`
##   <chr>      <dbl>   <dbl>
## 1 Afghanistan  745    2666
## 2 Brazil       37737   80488
## 3 China        212258  213766
```

The `table4a` dataframe is in the wide format, which is a common way to enter data into software like Excel. In this format, each column often represents a year, and the values are spread across columns for each country.

To convert this data into the long format, we'll use `tidyverse::pivot_longer()`. This function reshapes the data, making it easier to work with:

```
table4a %>%
  pivot_longer(c(`1999`, `2000`), names_to = "year", values_to = "cases")

## # A tibble: 6 x 3
##   country   year   cases
##   <chr>     <chr>  <dbl>
## 1 Afghanistan 1999    745
## 2 Afghanistan 2000   2666
## 3 Brazil      1999  37737
## 4 Brazil      2000  80488
## 5 China       1999 212258
## 6 China       2000 213766
```

Now, the data fulfills the three rules for tidy data:

1. Each variable (country and year) has its own column.
2. Each observation (combinations of country and year) has its own row.
3. Each value (the number of cases) has its own cell.

Understanding the difference between long and wide formats is essential because it determines how we structure our data for analysis. Once you grasp these concepts, reshaping and tidying data becomes a more comfortable and intuitive process.

Take your time to study and understand these data formats. It will significantly boost your confidence in reshaping and tidying data for your analytical tasks.

9.5 Introducing dplyr: Your Data Wrangling Toolkit

In this lesson, we'll delve into the world of the `dplyr` package in R, which offers a powerful and concise set of tools for working with data frames. As a biology student new to programming, `dplyr` will be your trusty companion for effortlessly managing, filtering, and summarizing biological data.

To master `dplyr`, it's crucial to grasp its five core functions: `mutate()`, `select()`, `filter()`, `summarise()`, and `arrange()`. Each function serves a specific purpose in data manipulation.

9.5.1 Selecting Specific Columns

We will be working with the same tibble called `tcga_data` that has numerous columns, but you're interested in extracting the “EPCAM” column along with other metadata columns. EPCAM, short for epithelial cellular adhesion molecule, is highly relevant in the context of epithelial cancer cells.

```
tcga_data %>%
  select(EPCAM, tcga.tcgbarcode:sample_type)

## # A tibble: 11,348 x 5
##   EPCAM tcga.tcgbarcode      tcga.cgc_sample_sample~1 study sample_type
##   <dbl> <chr>                  <chr>                <chr> <chr>
## 1 4.52 TCGA-OR-A5KU-01A-11R-A29S-07 Primary Tumor      ACC  cancer
## 2 9.53 TCGA-P6-A50G-01A-22R-A29S-07 Primary Tumor      ACC  cancer
## 3 42.4 TCGA-OR-A5K5-01A-11R-A29S-07 Primary Tumor      ACC  cancer
## 4 16.3 TCGA-OR-A5K4-01A-11R-A29S-07 Primary Tumor      ACC  cancer
## 5 12.5 TCGA-OR-A5LP-01A-11R-A29S-07 Primary Tumor      ACC  cancer
## 6 2.43 TCGA-PK-A5H9-01A-11R-A29S-07 Primary Tumor      ACC  cancer
## 7 3.74 TCGA-OR-A5LD-01A-11R-A29S-07 Primary Tumor      ACC  cancer
## 8 4.08 TCGA-OR-A5JX-01A-11R-A29S-07 Primary Tumor      ACC  cancer
## 9 2.84 TCGA-PK-A5H8-01A-11R-A29S-07 Primary Tumor      ACC  cancer
## 10 3.61 TCGA-OR-A5J3-01A-11R-A29S-07 Primary Tumor     ACC  cancer
```

```
## # i 11,338 more rows
## # i abbreviated name: 1: tcga.cgc_sample_sample_type
```

This code utilizes the `select()` function to pick specific columns by their names, creating a tidy subset of your data. We used `:` to select a range of columns. Note that the columns are reordered putting EPCAM to the first column.

You can also use index to select the columns

```
colnames(tcga_data)
```

```
## [1] "X"                      "TACSTD2"
## [3] "VTCN1"                  "MUC1"
## [5] "NECTIN4"                "FOLH1"
## [7] "FOLR1"                  "CD276"
## [9] "MSLN"                   "CLDN6"
## [11] "ERBB2"                  "MUC16"
## [13] "DLL3"                   "CEACAM5"
## [15] "PVR"                    "EPCAM"
## [17] "PROM1"                  "CD24"
## [19] "EGFR"                   "MET"
## [21] "TNFRSF10B"              "tcga.tcga_barcode"
## [23] "tcga.cgc_sample_sample_type" "study"
## [25] "sample_type"
```

The EPCAM column is the 16th column; `tcga.tcga_barcode` is the 23rd column; `sample_type` is 25th column:

```
tcga_data %>%
  select(16, 23:25)
```

```
## # A tibble: 11,348 x 4
##   EPCAM tcga.cgc_sample_sample_type study sample_type
##   <dbl> <chr>                  <chr> <chr>
## 1 4.52 Primary Tumor               ACC   cancer
## 2 9.53 Primary Tumor               ACC   cancer
## 3 42.4 Primary Tumor              ACC   cancer
## 4 16.3 Primary Tumor              ACC   cancer
## 5 12.5 Primary Tumor              ACC   cancer
## 6 2.43 Primary Tumor              ACC   cancer
## 7 3.74 Primary Tumor              ACC   cancer
## 8 4.08 Primary Tumor              ACC   cancer
## 9 2.84 Primary Tumor              ACC   cancer
## 10 3.61 Primary Tumor             ACC   cancer
## # i 11,338 more rows
```

We can even mix the index with column names:

```
tcga_data %>%
  select(EPCAM, 23:25)

## # A tibble: 11,348 x 4
##   EPCAM tcga.cgc_sample_sample_type study sample_type
##   <dbl> <chr>                <chr> <chr>
## 1 4.52 Primary Tumor             ACC    cancer
## 2 9.53 Primary Tumor             ACC    cancer
## 3 42.4 Primary Tumor            ACC    cancer
## 4 16.3 Primary Tumor            ACC    cancer
## 5 12.5 Primary Tumor            ACC    cancer
## 6 2.43 Primary Tumor            ACC    cancer
## 7 3.74 Primary Tumor            ACC    cancer
## 8 4.08 Primary Tumor            ACC    cancer
## 9 2.84 Primary Tumor            ACC    cancer
## 10 3.61 Primary Tumor           ACC    cancer
## # i 11,338 more rows

# save it to a new variable

tcga_data<- tcga_data %>%
  select(EPCAM, 23:25)
```

9.5.2 Adding New Columns

The `mutate()` function allows you to introduce new variables that are derived from existing ones. For instance, you can create a new column, “`log2EPCAM`,” containing the logarithm base-2 of the `EPCAM` values:

```
tcga_data<- tcga_data %>%
  mutate(log2EPCAM = log2(EPCAM))

tcga_data

## # A tibble: 11,348 x 5
##   EPCAM tcga.cgc_sample_sample_type study sample_type log2EPCAM
##   <dbl> <chr>                <chr> <chr>        <dbl>
## 1 4.52 Primary Tumor             ACC    cancer       2.18
## 2 9.53 Primary Tumor             ACC    cancer       3.25
## 3 42.4 Primary Tumor            ACC    cancer       5.40
## 4 16.3 Primary Tumor            ACC    cancer       4.03
```

```

## 5 12.5 Primary Tumor          ACC cancer 3.65
## 6 2.43 Primary Tumor          ACC cancer 1.28
## 7 3.74 Primary Tumor          ACC cancer 1.90
## 8 4.08 Primary Tumor          ACC cancer 2.03
## 9 2.84 Primary Tumor          ACC cancer 1.51
## 10 3.61 Primary Tumor         ACC cancer 1.85
## # i 11,338 more rows

```

9.5.3 Reordering Columns

If you want to rearrange the order of columns, you can again use `select()`. Here, we move the “log2EPCAM” column to the front while keeping all other columns intact:

```

tcga_data %>%
  select(EPCAM, log2EPCAM, everything())

```



```

## # A tibble: 11,348 x 5
##   EPCAM log2EPCAM tcga.cgc_sample_sample_type study sample_type
##   <dbl>     <dbl> <chr>                <chr> <chr>
## 1 4.52      2.18 Primary Tumor             ACC cancer
## 2 9.53      3.25 Primary Tumor             ACC cancer
## 3 42.4       5.40 Primary Tumor            ACC cancer
## 4 16.3       4.03 Primary Tumor            ACC cancer
## 5 12.5       3.65 Primary Tumor            ACC cancer
## 6 2.43       1.28 Primary Tumor            ACC cancer
## 7 3.74       1.90 Primary Tumor            ACC cancer
## 8 4.08       2.03 Primary Tumor            ACC cancer
## 9 2.84       1.51 Primary Tumor            ACC cancer
## 10 3.61      1.85 Primary Tumor           ACC cancer
## # i 11,338 more rows

```

The `everything()` helper function denotes all other columns, ensuring your numeric columns are at the front.

9.5.4 Filtering Data

`filter()` is your go-to function for extracting observations that meet specific criteria. To isolate data only related to glioblastoma (GBM), you can apply the following filter:

```
tcga_data %>%
  filter(study == "GBM")

## # A tibble: 175 x 5
##   EPCAM tcga.cgc_sample_sample_type study sample_type log2EPCAM
##   <dbl> <chr> <chr> <chr> <dbl>
## 1 0.329 Primary Tumor             GBM  cancer    -1.60
## 2 0.152 Primary Tumor             GBM  cancer    -2.72
## 3 0.0814 Primary Tumor            GBM  cancer    -3.62
## 4 0.367 Recurrent Tumor           GBM  cancer    -1.45
## 5 0.0614 Primary Tumor            GBM  cancer    -4.02
## 6 0.350 Primary Tumor             GBM  cancer    -1.51
## 7 0.165 Primary Tumor             GBM  cancer    -2.60
## 8 0.0989 Primary Tumor            GBM  cancer    -3.34
## 9 0.466 Primary Tumor             GBM  cancer    -1.10
## 10 0.707 Recurrent Tumor          GBM  cancer    -0.500
## # i 165 more rows
```

This code snippet retains only the rows corresponding to GBM in your dataset.

9.5.5 Summarizing Data

Suppose you want to calculate the average EPCAM expression for each cancer type in your dataset. You can utilize `summarise()` in conjunction with `group_by()`:

```
tcga_data %>%
  group_by(study) %>%
  summarise(average_EPCAM = mean(EPCAM))

## # A tibble: 33 x 2
##   study average_EPCAM
##   <chr>      <dbl>
## 1 ACC        11.2
## 2 BLCA       102.
## 3 BRCA       177.
## 4 CESC       166.
## 5 CHOL       223.
## 6 COAD       792.
## 7 DLBC        2.61
## 8 ESCA       273.
## 9 GBM        0.623
## 10 HNSC      50.3
## # i 23 more rows
```

Here, the data is grouped by the “study” variable, and the `summarise()` function calculates the mean EPCAM value for each group.

9.5.6 Sorting Data

To sort your data frame based on a specific column, employ `arrange()`. For instance, you can order your dataset by the median level of EPCAM:

```
tcga_data %>%
  group_by(study) %>%
  summarise(average_EPCAM = mean(EPCAM),
            median_EPCAM = median(EPCAM)) %>%
  arrange(median_EPCAM)
```

```
## # A tibble: 33 x 3
##   study average_EPCAM median_EPCAM
##   <chr>      <dbl>        <dbl>
## 1 UVM       0.765       0.0583
## 2 SKCM      0.980       0.133
## 3 SARC      2.87        0.224
## 4 GBM       0.623       0.323
## 5 DLBC      2.61        0.578
## 6 LAML      3.14        0.595
## 7 LGG       0.906       0.681
## 8 MESO      11.3        1.14
## 9 LIHC      31.2        1.22
## 10 ACC      11.2        4.83
## # i 23 more rows
```

The default is `arrange` from the smallest to the biggest. Let’s reverse it by using the helper descending function `desc`:

```
tcga_data %>%
  group_by(study) %>%
  summarise(average_EPCAM = mean(EPCAM),
            median_EPCAM = median(EPCAM)) %>%
  arrange(desc(median_EPCAM))
```

```
## # A tibble: 33 x 3
##   study average_EPCAM median_EPCAM
##   <chr>      <dbl>        <dbl>
## 1 READ      834.        808.
## 2 COAD      792.        787.
```

```

## 3 THCA      350.    345.
## 4 UCEC      351.    337.
## 5 STAD      335.    306.
## 6 LUAD      307.    289.
## 7 PAAD      278.    265.
## 8 CHOL      223.    236.
## 9 OV        228.    207.
## 10 PRAD     199.    182.
## # i 23 more rows

```

We see READ and COAD colon cancers have the highest EPCAM expression.

This code sorts the dataset from the smallest to the largest median EPCAM values.

9.5.7 Conclusion

In summary:

- `mutate()` adds new columns.
- `filter()` extracts specific observations.
- `select()` picks/reorders columns.
- `summarise()` reduces multiple values to summaries.
- `arrange()` reorders rows.

These four fundamental functions empower you to efficiently manipulate, analyze, and summarize biological data frames, providing a more concise and readable approach compared to traditional R methods. As your programming skills grow, `dplyr` will remain an indispensable tool in your data science toolkit.

9.6 stringr: your essential toolkit to manipulate strings

xxxx

9.7 purrr: ditch your for loops

In this lesson, we'll learn about the `purrr` package in R. `Purrr` provides a set of tools for working with lists and other recursive data structures in a functional programming style.

A recursive data structure in R refers to a data object that can contain other objects of the same type as its components. For example, a list in R can be recursive because it can contain other lists within it, creating a nested or hierarchical structure.

As a biology student, you'll likely need to apply the same operation to multiple data sets or columns. That's where `purrr` becomes really handy! The key functions we'll cover are:

- `map()` and its variants `map_chr()`, `map_dbl()` - Applies a function to each element of a list or vector. For example, you could calculate the mean of every column in a data frame:

In the previous section, we learned about loops. There is nothing wrong with for loops. However, with `purrr::map()`, I find myself writing less and less for loops.

9.7.1 Nesting Data with `nest()`

The `nest()` function in R, when used with tibbles, groups your data based on a specific variable and creates a new column containing nested data frames for each group. It's like putting similar data into separate containers, making it easier to work with and analyze them as a whole or individually.

Imagine you have a large table of information about different types of cancer samples. You want to organize this data in a way that groups all the information related to each type of cancer separately. One way to do this is by using the `tidyverse::nest()` function along with the `purrr` package in R.

Here's how you can achieve this:

```
# read in the data again
tcga_data <- readr::read_csv("~/Downloads/TCGA_cancer_genes_expression.csv")

# Group and nest the data
tcga_nest <- tcga_data %>%
  filter(sample_type == "cancer") %>%
  select(EPCAM, tcga_barcode:sample_type) %>%
```

```
group_by(study) %>%
tidyr::nest()

tcga_nest

## # A tibble: 32 x 2
## # Groups:   study [32]
##   study      data
##   <chr>     <list>
## 1 ACC     <tibble [79 x 4]>
## 2 BLCA    <tibble [414 x 4]>
## 3 BRCA    <tibble [1,127 x 4]>
## 4 CESC     <tibble [304 x 4]>
## 5 CHOL    <tibble [36 x 4]>
## 6 COAD    <tibble [504 x 4]>
## 7 DLBC    <tibble [48 x 4]>
## 8 ESCA    <tibble [184 x 4]>
## 9 GBM     <tibble [170 x 4]>
## 10 HNSC   <tibble [502 x 4]>
## # i 22 more rows
```

The `tidyr::nest()` function creates a list-column within your tibble, where each element of the list is a nested data frame. This is a powerful feature of tibbles, as they can contain tables within the table.

You can access the nested data using the `$` sign, just like you would with a regular data frame, and use the double bracket to access the element. For example:

```
# Accessing the first nested data frame
first_nested_df <- tcga_nest$data[[1]]

first_nested_df

## # A tibble: 79 x 4
##   EPCAM tcga_tcga_barcode      tcga_cgc_sample_sample_type sample_type
##   <dbl> <chr>                  <chr>                      <chr>
## 1 4.52 TCGA-OR-A5KU-01A-11R-A29S-07 Primary Tumor            cancer
## 2 9.53 TCGA-P6-A50G-01A-22R-A29S-07 Primary Tumor            cancer
## 3 42.4 TCGA-OR-A5K5-01A-11R-A29S-07 Primary Tumor            cancer
## 4 16.3 TCGA-OR-A5K4-01A-11R-A29S-07 Primary Tumor            cancer
## 5 12.5 TCGA-OR-A5LP-01A-11R-A29S-07 Primary Tumor            cancer
## 6 2.43 TCGA-PK-A5H9-01A-11R-A29S-07 Primary Tumor            cancer
## 7 3.74 TCGA-OR-A5LD-01A-11R-A29S-07 Primary Tumor            cancer
```

```

##  8 4.08 TCGA-OR-A5JX-01A-11R-A29S-07 Primary Tumor      cancer
##  9 2.84 TCGA-PK-A5H8-01A-11R-A29S-07 Primary Tumor      cancer
## 10 3.61 TCGA-OR-A5J3-01A-11R-A29S-07 Primary Tumor      cancer
## # i 69 more rows

```

In this example, `first_nested_df` contains the first nested data frame, which corresponds to one of the “study” groups.

You can add the names to the list column, and now you can access it by cancer type:

```

names(tcga_nest$data) <- tcga_nest$study

tcga_nest

## # A tibble: 32 x 2
## # Groups:   study [32]
##       study data
##       <chr> <named list>
## 1 ACC    <tibble [79 x 4]>
## 2 BLCA   <tibble [414 x 4]>
## 3 BRCA   <tibble [1,127 x 4]>
## 4 CESC   <tibble [304 x 4]>
## 5 CHOL   <tibble [36 x 4]>
## 6 COAD   <tibble [504 x 4]>
## 7 DLBC   <tibble [48 x 4]>
## 8 ESCA   <tibble [184 x 4]>
## 9 GBM    <tibble [170 x 4]>
## 10 HNSC   <tibble [502 x 4]>
## # i 22 more rows

tcga_nest$data[["ACC"]]

## # A tibble: 79 x 4
##   EPCAM tcga.tcgBarcode tcga.cgcSample.sample_type sample_type
##   <dbl> <chr>           <chr>                  <chr>
## 1 4.52 TCGA-OR-A5KU-01A-11R-A29S-07 Primary Tumor      cancer
## 2 9.53 TCGA-P6-A50G-01A-22R-A29S-07 Primary Tumor      cancer
## 3 42.4 TCGA-OR-A5K5-01A-11R-A29S-07 Primary Tumor      cancer
## 4 16.3 TCGA-OR-A5K4-01A-11R-A29S-07 Primary Tumor      cancer
## 5 12.5 TCGA-OR-A5LP-01A-11R-A29S-07 Primary Tumor      cancer
## 6 2.43 TCGA-PK-A5H9-01A-11R-A29S-07 Primary Tumor      cancer
## 7 3.74 TCGA-OR-A5LD-01A-11R-A29S-07 Primary Tumor      cancer
## 8 4.08 TCGA-OR-A5JX-01A-11R-A29S-07 Primary Tumor      cancer
## 9 2.84 TCGA-PK-A5H8-01A-11R-A29S-07 Primary Tumor      cancer

```

```
## 10 3.61 TCGA-OR-A5J3-01A-11R-A29S-07 Primary Tumor cancer
## # i 69 more rows
```

9.7.2 map() and Its Variants

Let's calculate the median value of EPCAM for each cancer type using `map()`.

`map()` takes in a vector or a list, and a function to be applied to every element of the vector or the list.

```
map(tcga_nest$data, function(x) (median(x$EPCAM)))
```

```
## $ACC
## [1] 4.830195
##
## $BLCA
## [1] 81.1488
##
## $BRCA
## [1] 161.9358
##
## $CESC
## [1] 75.93449
##
## $CHOL
## [1] 251.4888
##
## $COAD
## [1] 777.5484
##
## $DLBC
## [1] 0.5783636
##
## $ESCA
## [1] 189.3882
##
## $GBM
## [1] 0.3073375
##
## $HNSC
## [1] 24.70459
##
## $KICH
## [1] 93.78075
##
```

```
## $KIRC
## [1] 24.10957
##
## $KIRP
## [1] 71.4589
##
## $LGG
## [1] 0.6812875
##
## $LIHC
## [1] 0.7269132
##
## $LUAD
## [1] 305.5941
##
## $LUSC
## [1] 138.6225
##
## $MESO
## [1] 1.144048
##
## $OV
## [1] 206.6447
##
## $PAAD
## [1] 267.3574
##
## $PCPG
## [1] 6.188397
##
## $PRAD
## [1] 194.4213
##
## $READ
## [1] 808.5985
##
## $SARC
## [1] 0.2179763
##
## $SKCM
## [1] 0.2021555
##
## $STAD
## [1] 320.1896
##
## $TGCT
```

```
## [1] 116.1016
##
## $THCA
## [1] 342.4736
##
## $THYM
## [1] 13.63511
##
## $UCEC
## [1] 348.4112
##
## $UCS
## [1] 94.77784
##
## $UVM
## [1] 0.05832845
```

In this example, the function takes each element of the list of the data frame and return the median of the EPCAM.

9.7.3 Note on Anonymous functions

There are three ways to specify a function

```
# full function with a function name
calculate_median<- function(x){
  return(median(x$EPCAM))
}

# base R anonymous function
function(x) (median(x$EPCAM))

# purrr anonymous function using formula ~, note you use .x instead of x
~ median(.x$EPCAM)
```

The following will have the same results

```
map(tcga_nest$data, calculate_median)
map(tcga_nest$data, function(x) (median(x$EPCAM)))
map(tcga_nest$data, ~ median(.x$EPCAM))
```

read more at https://jennybc.github.io/purrr-tutorial/ls03_map-function-syntax.html#anonymous_function,_conventional.

map always returns a list, it returns a list of median values in this case. If you want to return a vector, use `map_dbl`:

```
map_dbl(tcga_nest$data, function(x) (median(x$EPCAM)))
```

```
##      ACC       BLCA       BRCA       CESC       CHOL       COAD
## 4.83019522 81.14880204 161.93580869 75.93448764 251.48875310 777.54836210
##      DLBC       ESCA       GBM       HNSC       KICH       KIRC
## 0.57836358 189.38816785 0.30733745 24.70459340 93.78075352 24.10956585
##      KIRP       LGG       LIHC       LUAD       LUSC       MESO
## 71.45889552 0.68128748 0.72691319 305.59410291 138.62247973 1.14404760
##      OV        PAAD       PCPG       PRAD       READ       SARC
## 206.64472617 267.35742947 6.18839726 194.42130740 808.59850470 0.21797629
##      SKCM       STAD       TGCT       THCA       THYM       UCEC
## 0.20215548 320.18957706 116.10162785 342.47358682 13.63510740 348.41123728
##      UCS       UVM
## 94.77783740 0.05832845
```

Let's save the output to a new variable

```
median_EPCAM<- map_dbl(tcga_nest$data, function(x) (median(x$EPCAM)))

# returns a list of log2 values
map(median_EPCAM, function(x) log2(x))

## $ACC
## [1] 2.272081
##
## $BLCA
## [1] 6.342498
##
## $BRCA
## [1] 7.339278
##
## $CESC
## [1] 6.246683
##
## $CHOL
## [1] 7.97435
##
## $COAD
## [1] 9.602789
##
## $DLBC
```

```
## [1] -0.7899514
##
## $ESCA
## [1] 7.565202
##
## $GBM
## [1] -1.702105
##
## $HNSC
## [1] 4.626707
##
## $KICH
## [1] 6.55122
##
## $KIRC
## [1] 4.591534
##
## $KIRP
## [1] 6.159042
##
## $LGG
## [1] -0.5536644
##
## $LIHC
## [1] -0.460145
##
## $LUAD
## [1] 8.255473
##
## $LUSC
## [1] 7.115017
##
## $MESO
## [1] 0.1941471
##
## $OV
## [1] 7.691009
##
## $PAAD
## [1] 8.062626
##
## $PCPG
## [1] 2.629566
##
## $PRAD
## [1] 7.603043
```

```

## 
## $READ
## [1] 9.65928
##
## $SARC
## [1] -2.197757
##
## $SKCM
## [1] -2.306463
##
## $STAD
## [1] 8.322783
##
## $TGCT
## [1] 6.859244
##
## $THCA
## [1] 8.419849
##
## $THYM
## [1] 3.769254
##
## $UCEC
## [1] 8.444647
##
## $UCS
## [1] 6.566478
##
## $UVM
## [1] -4.099656

# returns a vector
map_dbl(median_EPCAM, function(x) log2(x))

##      ACC       BLCA       BRCA       CESC       CHOL       COAD       DLBC
## 2.2720815 6.3424979 7.3392782 6.2466834 7.9743501 9.6027886 -0.7899514
##      ESCA       GBM       HNSC       KICH       KIRC       KIRP       LGG
## 7.5652024 -1.7021045 4.6267074 6.5512200 4.5915338 6.1590417 -0.5536644
##      LIHC       LUAD       LUSC       MESO       OV       PAAD       PCPG
## -0.4601450 8.2554729 7.1150174 0.1941471 7.6910087 8.0626260 2.6295658
##      PRAD       READ       SARC       SKCM       STAD       TGCT       THCA
## 7.6030425 9.6592797 -2.1977569 -2.3064627 8.3227825 6.8592444 8.4198489
##      THYM       UCEC       UCS       UVM
## 3.7692542 8.4446473 6.5664778 -4.0996564

```

We can stay in the original tibble by just adding the median to a new column

with `mutate()`:

```
tcga_nest %>%
  mutate(median_EPCAM = map_dbl(data, function(x) (median(x$EPCAM))))
```

	study	median_EPCAM
1	ACC	4.83
2	BLCA	81.1
3	BRCA	162.
4	CESC	75.9
5	CHOL	251.
6	COAD	778.
7	DLBC	0.578
8	ESCA	189.
9	GBM	0.307
10	HNSC	24.7
	# i 22 more rows	

of course, we can use `group_by` followed by `summarise` as shown above to get the same thing, but it demonstrates how we can combine `map()` function and list column to do powerful data analysis within a data frame.

```
tcga_data %>%
  filter(sample_type == "cancer") %>%
  select(EPCAM, tcga_barcode:sample_type) %>%
  group_by(study) %>%
  summarise(median_EPCAM = median(EPCAM))
```

	study	median_EPCAM
1	ACC	4.83
2	BLCA	81.1
3	BRCA	162.
4	CESC	75.9
5	CHOL	251.
6	COAD	778.
7	DLBC	0.578
8	ESCA	189.
9	GBM	0.307
10	HNSC	24.7
	# i 22 more rows	

Read this <https://dplyr.tidyverse.org/reference/summarise.html> for more examples using group_by with summarise.

You can even nest by two columns:

```
tcga_data %>%
  select(EPCAM, tcga.tcgbarcode:sample_type) %>%
  group_by(study, sample_type) %>%
  tidyverse::nest()
```

```
## # A tibble: 73 x 3
## # Groups:   study, sample_type [73]
##   study sample_type data
##   <chr>  <chr>      <list>
## 1 ACC    cancer     <tibble [79 x 3]>
## 2 BLCA   cancer     <tibble [414 x 3]>
## 3 BLCA   normal     <tibble [19 x 3]>
## 4 BRCA   cancer     <tibble [1,127 x 3]>
## 5 BRCA   normal     <tibble [112 x 3]>
## 6 BRCA   metastatic <tibble [7 x 3]>
## 7 BRCA   <NA>       <tibble [10 x 3]>
## 8 CESC   cancer     <tibble [304 x 3]>
## 9 CESC   normal     <tibble [3 x 3]>
## 10 CESC  metastatic <tibble [2 x 3]>
## # i 63 more rows
```

You can easily see some cancer types have normal and some have metastatic samples and you can do everything within a dataframe.

The other way to check the information is to use the table function:

```
table(tcga_data$sample_type, tcga_data$study)

##
##          ACC BLCA BRCA CESC CHOL COAD DLBC ESCA GBM HNSC KICH KIRC KIRP
## cancer      79  414 1127  304   36  504   48  184  170  502   66  544  291
## metastatic   0    0    7    2    0    1    0    1    0    2    0    0    0    0
## normal       0   19   112   3    9   41    0   13    5   44   25   72   32
##
##          LAML LGG LIHC LUAD LUSC MESO OV PAAD PCPG PRAD READ SARC SKCM
## cancer      0  532  374  542  504   87  429  178  182  505  167  262  103
## metastatic   0    0    0    0    0    0    0    1    2    1    0    1  368
## normal       0    0   50   59   51    0    0    4    3   52   10    2    1
##
##          STAD TGCT THCA THYM UCEC UCS UVM
```

```
##   cancer      415 156 505 120 554 57 80
##   metastatic 0    0    8    0    0    0    0
##   normal      37   0   59   2   35   0    0
```

The key takeaways is:

- `map()` applies a function over each element of a list/vector

9.7.4 Conclusion

With `purrr`, you get a powerful toolkit for iterating over biological data in a functional programming style. As you advance in bioinformatics, `purrr` will continue to make your code more clear and more concise. This introduction is just the tip of the iceberg. To learn more about `purrr`, read this <https://jennybc.github.io/purrr-tutorial/>

Some key benefits of the tidyverse include:

1. Consistent language and grammar across packages like piping (`%>%`) and verb functions (`filter()`, `mutate()`).
2. Works well with pipe workflows for transforming and visualizing data.
3. Enhances exploratory data analysis and makes it easy to go from raw data to insights.
4. Large community providing learning resources and support.

The tidyverse packages work seamlessly together, allowing you to conduct complete data science projects in R. From import to wrangling to visualization, the tidyverse provides a set of tools that follow common principles. While we can dedicate a full chapter on each package, we only focused on `dplyr`, `tidyverse` and `purrr`.

9.8 Tidying metadata from GEO.

In this lesson, we will learn how to tidy metadata obtained from the **Gene Expression Omnibus** (GEO) database using the Tidyverse package in R. Tidying metadata is an essential step in preparing data for analysis. We will use a real-world example from GEO to demonstrate the process step-by-step.

9.8.1 Prerequisites

Before we begin, make sure you have R and the necessary packages installed. You can install the required packages using the following commands:

```
BiocManager::install("GEOquery")
install.packages("tidyverse")
```

9.8.2 Getting Started

We'll start by loading the necessary libraries and fetching metadata from a GEO dataset called `GSE176021`. GEO is a repository for gene expression data and other omics data, and it often contains metadata in a wide format, which is not suitable for analysis.

```
# Load the necessary libraries
library(GEOquery)
library(tidyverse)

# Fetch metadata from GEO
GSE176021_meta <- getGEO(GEO = "GSE176021", GSEMatrix = FALSE)
```

Let's use the `str` structure command to inspect the `GSE176021_meta@gsms` object

```
str(GSE176021_meta@gsms, max.level = 1)

## List of 110
## $ GSM5352886:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352887:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352888:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352889:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352890:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352891:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352892:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352893:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352894:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352895:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352896:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352897:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352898:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352899:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352900:Formal class 'GSM' [package "GEOquery"] with 2 slots
```



```

## $ GSM5352947:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352948:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352949:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352950:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352951:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352952:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352953:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352954:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352955:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352956:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352957:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352958:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352959:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352960:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352961:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352962:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352963:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352964:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352965:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352966:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352967:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352968:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352969:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352970:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352971:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352972:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352973:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352974:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352975:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352976:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352977:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352978:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352979:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352980:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352981:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352982:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352983:Formal class 'GSM' [package "GEOquery"] with 2 slots
## $ GSM5352984:Formal class 'GSM' [package "GEOquery"] with 2 slots
## [list output truncated]

```

So it is a list of 110 GSM objects. Let's take a look at the first element of the list

```
GSE176021_meta@gsms[[1]]@header
```

```
## $channel_count
```

```
## [1] "1"
##
## $characteristics_ch1
## [1] "patientid: MD01-024"      "cell type: lymphocytes"
## [3] "response status: Non-MPR"
##
## $contact_address
## [1] "1650 Orleans Street, CRB1, 4M"
##
## $contact_city
## [1] "Baltimore"
##
## $contact_country
## [1] "USA"
##
## $contact_email
## [1] "ksmit228@jhmi.edu"
##
## $contact_institute
## [1] "Johns Hopkins University"
##
## $contact_name
## [1] "Kellie,Nicole,Smith"
##
## $contact_state
## [1] "MD"
##
## $`contact_zip/postal_code`
## [1] "21287"
##
## $data_processing
## [1] "Cell Ranger v3.1.0 was used to demultiplex the FASTQ reads, align them to the GRCh38 human genome reference, and generate gene expression and cell type metadata. The resulting RDS files include cell type annotations and quality metrics for each cell." 
## [2] "Supplementary_files_format_and_content: Cell ranger output"
## [3] "Supplementary_files_format_and_content: RDS files include metadata with cell type annotations and quality metrics for each cell." 
##
## $data_row_count
## [1] "0"
##
## $extract_protocol_ch1
## [1] "Cryobanked T cells were thawed and washed twice with pre-warmed RPMI with 20% FBS and genotyped using the Illumina Infinium HumanMethylation450 BeadChip array." 
## [2] "The Single Cell 5' V(D)J and 5' DGE kits (10X Genomics) were used to capture immune receptor repertoires from individual cells." 
##
## $geo_accession
## [1] "GSM5352886"
##
## $instrument_model
```

```
## [1] "Illumina NovaSeq 6000"
##
## $last_update_date
## [1] "Feb 20 2024"
##
## $library_selection
## [1] "cDNA"
##
## $library_source
## [1] "transcriptomic"
##
## $library_strategy
## [1] "RNA-Seq"
##
## $molecule_ch1
## [1] "total RNA"
##
## $organism_ch1
## [1] "Homo sapiens"
##
## $platform_id
## [1] "GPL24676"
##
## $relation
## [1] "BioSample: https://www.ncbi.nlm.nih.gov/biosample/SAMN19514461"
##
## $series_id
## [1] "GSE173351" "GSE176021"
##
## $source_name_ch1
## [1] "tumor"
##
## $status
## [1] "Public on Jul 21 2021"
##
## $submission_date
## [1] "Jun 02 2021"
##
## $supplementary_file_1
## [1] "ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM5352nnn/GSM5352886/suppl/GSM5352886_1"
##
## $supplementary_file_2
## [1] "ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM5352nnn/GSM5352886/suppl/GSM5352886_2"
##
## $taxid_ch1
## [1] "9606"
```

```
##  
## $title  
## [1] "MD01-024_tumor_1"  
##  
## $type  
## [1] "SRA"
```

It will print out a long text. We only need the `characteristics_ch1`

```
GSE176021_meta@gsms[[1]]@header$characteristics_ch1
```

```
## [1] "patientid: MD01-024"      "cell type: lymphocytes"  
## [3] "response status: Non-MPR"
```

This contains the metadata we need.

Now, let's extract all the metadata for all samples and bind them to a dataframe.

```
GSE176021_meta <- purrr::map(GSE176021_meta@gsms,  
                                function(x) x@header$characteristics_ch1) %>%  
  bind_rows() %>%  
  dplyr::slice(c(1, 3))
```

```
GSE176021_meta
```

```
## # A tibble: 2 x 110  
##   GSM5352886   GSM5352887   GSM5352888   GSM5352889   GSM5352890   GSM5352891   GSM5352892  
##   <chr>       <chr>       <chr>       <chr>       <chr>       <chr>  
## 1 patientid: ~ patientid~ patientid~ patientid~ patientid~ patientid~  
## 2 response st~ response ~ response ~ response ~ response ~ response ~  
## # i 103 more variables: GSM5352893 <chr>, GSM5352894 <chr>, GSM5352895 <chr>,  
## #   GSM5352896 <chr>, GSM5352897 <chr>, GSM5352898 <chr>, GSM5352899 <chr>,  
## #   GSM5352900 <chr>, GSM5352901 <chr>, GSM5352902 <chr>, GSM5352903 <chr>,  
## #   GSM5352904 <chr>, GSM5352905 <chr>, GSM5352906 <chr>, GSM5352907 <chr>,  
## #   GSM5352908 <chr>, GSM5352909 <chr>, GSM5352910 <chr>, GSM5352911 <chr>,  
## #   GSM5352912 <chr>, GSM5352913 <chr>, GSM5352914 <chr>, GSM5352915 <chr>,  
## #   GSM5352916 <chr>, GSM5352917 <chr>, GSM5352918 <chr>, GSM5352919 <chr>, ...
```

In this code:

1. We load the `GEOquery` library to access functions related to the Gene Expression Omnibus (GEO) database.
2. We fetch metadata from the GEO dataset “GSE176021” using the `getGEO` function. The `GSEMatrix = FALSE` argument ensures that we retrieve metadata rather than expression data.

3. We use `purrr::map` to extract the “characteristics_ch1” information from each sample within the GEO dataset. This information typically contains details about the samples, such as patient identifiers and response statuses.
 4. Next, we use `bind_rows()` to combine these extracted characteristics into a single data frame.
 5. We use `dplyr::slice(c(1, 3))` to select only the first and third rows of the resulting data frame, essentially keeping a subset of the metadata for demonstration purposes.

Now, let's take a look at the wide-format metadata:

```
# Display the first two rows and first ten columns of the wide-format metadata
GSE176021_meta[1:2, 1:10]

## # A tibble: 2 x 10
##   GSM5352886   GSM5352887   GSM5352888   GSM5352889   GSM5352890   GSM5352891   GSM5352892
##   <chr>         <chr>         <chr>         <chr>         <chr>         <chr>         <chr>
## 1 patientid~ ~ patientid~ ~ patientid~ ~ patientid~ ~ patientid~ ~ patientid~ ~
## 2 response st~ response ~ response ~ response ~ response ~ response ~ response ~
## # i 3 more variables: GSM5352893 <chr>, GSM5352894 <chr>, GSM5352895 <chr>
```

The wide-format metadata is not suitable for analysis. It has many columns, and we need to tidy it before proceeding.

9.8.3 Tidying the Metadata

We will follow these steps to tidy the metadata:

1. Add a “meta” column with labels.
 2. Reorder the columns to have the “meta” column first.

```
# Add a "meta" column and reorder the columns
GSE176021_meta <- GSE176021_meta %>%
  mutate(meta = c("id", "response")) %>%
  select(meta, everything())
```

Now, let's see how it looks like:

```
# Display the first two rows and first ten columns of the tidied metadata
GSE176021_meta[1:2, 1:10]
```

```
## # A tibble: 2 x 10
##   meta      GSM5352886      GSM5352887      GSM5352888      GSM5352889      GSM5352890      GSM5352891
##   <chr>      <chr>      <chr>      <chr>      <chr>      <chr>      <chr>
## 1 id      patientid: MD~ patientid~ patientid~ patientid~ patientid~ patientid~
## 2 response response stat~ response ~ response ~ response ~ response ~ response ~
## # i 3 more variables: GSM5352892 <chr>, GSM5352893 <chr>, GSM5352894 <chr>
```

The metadata is ready to be shaped into a long format.

9.8.4 Converting to a Tidy Data Frame

To fulfill the three tidy data rules, we will convert the metadata into a tidy data frame using the `pivot_longer` and `pivot_wider` functions:

```
# pivot it to long format  
GSE176021_meta %>%  
  pivot_longer(cols = -meta)
```

```
## # A tibble: 220 x 3
##       meta    name     value
##   <chr> <chr>    <chr>
## 1 id    GSM5352886 patientid: MD01-024
## 2 id    GSM5352887 patientid: MD01-010
## 3 id    GSM5352888 patientid: MD01-010
## 4 id    GSM5352889 patientid: MD01-004
## 5 id    GSM5352890 patientid: MD01-004
## 6 id    GSM5352891 patientid: MD01-004
## 7 id    GSM5352892 patientid: MD01-004
## 8 id    GSM5352893 patientid: MD043-011
## 9 id    GSM5352894 patientid: MD043-011
## 10 id   GSM5352895 patientid: MD043-011
## # i 210 more rows
```

```
# pivot to wide format
GSE176021_meta %>%
  pivot_longer(cols = -meta) %>%
  pivot_wider(names_from = meta, values_from = value)
```

```
## # A tibble: 110 x 3
##   name      id             response
##   <chr>    <chr>          <chr>
## 1 GSM5352886 patientid: MD01-024 response status: Non-MPR
## 2 GSM5352887 patientid: MD01-010 response status: MPR
```

```

## 3 GSM5352888 patientid: MD01-010 response status: MPR
## 4 GSM5352889 patientid: MD01-004 response status: Non-MPR
## 5 GSM5352890 patientid: MD01-004 response status: Non-MPR
## 6 GSM5352891 patientid: MD01-004 response status: Non-MPR
## 7 GSM5352892 patientid: MD01-004 response status: Non-MPR
## 8 GSM5352893 patientid: MD043-011 response status: Non-MPR
## 9 GSM5352894 patientid: MD043-011 response status: Non-MPR
## 10 GSM5352895 patientid: MD043-011 response status: Non-MPR
## # i 100 more rows

# put it together
tidy_data <- GSE176021_meta %>%
  pivot_longer(cols = -meta) %>%
  pivot_wider(names_from = meta, values_from = value)

```

Let's take a look at the resulting tidy data frame:

```

# Display the first 10 rows of the tidy data frame
head(tidy_data, 10)

```

```

## # A tibble: 10 x 3
##   name      id      response
##   <chr>     <chr>    <chr>
## 1 GSM5352886 patientid: MD01-024 response status: Non-MPR
## 2 GSM5352887 patientid: MD01-010 response status: MPR
## 3 GSM5352888 patientid: MD01-010 response status: MPR
## 4 GSM5352889 patientid: MD01-004 response status: Non-MPR
## 5 GSM5352890 patientid: MD01-004 response status: Non-MPR
## 6 GSM5352891 patientid: MD01-004 response status: Non-MPR
## 7 GSM5352892 patientid: MD01-004 response status: Non-MPR
## 8 GSM5352893 patientid: MD043-011 response status: Non-MPR
## 9 GSM5352894 patientid: MD043-011 response status: Non-MPR
## 10 GSM5352895 patientid: MD043-011 response status: Non-MPR

```

Now, our data fulfills the three tidy data rules, making it suitable for analysis.

9.8.5 Bonus: Reading Multiple TSV Files

If you have multiple TSV (Tab-Separated Values) files in a directory that you want to read into R and combine, here's how you can do it:

```

# List all TSV files in the current directory
files <- as.list(dir(".", pattern = ".tsv"))

```

```
# Read and combine all TSV files into a single data frame
datlist <- lapply(files, function(f) {
  dat <- read_tsv(f, col_names = TRUE)
  dat$sample <- gsub(".tsv", "", f)
  return(dat)
})

data <- do.call(rbind, datlist)
```

Alternatively, you can use the `bind_rows` function from the `dplyr` package:

```
# Read and combine all TSV files into a single data frame using bind_rows
data <- bind_rows(datlist, .id = "sample")
```

If your files have a common column (e.g., “GeneID”), and you want to create a single data frame with gene IDs and raw counts, you can use the `reduce` function from the `purrr` package:

```
# Combine data frames using reduce and left_join
CCLE_counts <- purrr::reduce(datlist, left_join, by = "GeneID")
```

Watch this video:

9.9 Section Complete

Congratulations on completing this segment of the course!

We’ve covered the Tidyverse, data manipulation, and analysis techniques, including handling GEO metadata and using `purrr` for functional programming. These skills are crucial for efficient data analysis in R, simplifying workflows and cleaning up complex datasets.

As you move forward in the course, remember to utilize the Q&A section and comments for support. Engage actively for any clarifications or assistance you need.

9.9.1 Conclusion

In this lesson, we learned how to tidy metadata from GEO using Tidyverse in R. Tidying data is a crucial step in data preparation for analysis. We also explored how to read and combine multiple TSV files from a directory, which is a common task when dealing with large datasets.

Remember that these skills are valuable in many data analysis and bioinformatics tasks, allowing you to work efficiently with real-world data.

Chapter 10

Data visualization with ggplot2

In this lesson, we'll learn how to create basic data visualizations in R using ggplot2. As a biology student, developing your data visualization skills is crucial for exploring trends and communicating findings.

If you read genomics papers, most of the figures are fell into several categories:

- A bar plot
- A line plot
- A scatter plot
- A boxplot or violin plot
- A histogram
- A heatmap

If you master those six types of figures, you can reproduce 90% of the figures in any genomics paper. Watch this:

ggplot2 is a powerful R package for flexible and professional-quality graphics. The key is thinking about visualization in layers:

- The data layer - the data frame you want to plot
- The aesthetic mapping - how data columns map to visual properties like x/y position, color, size etc.
- Geometric objects like points, lines, boxes - the type of plot

- Facets - panels for subgrouping the data
- Stats transformations like group means, regressions etc.
- Themes to refine the final look

Let's walk through examples of common plot types.

10.1 Creating Scatter Plots

In this lesson, we will delve into data visualization and statistical analysis using R. We will work with a real-world dataset related to gene expression from The Cancer Genome Atlas (TCGA) and learn how to create scatter plots, calculate correlation coefficients, and visualize regression lines. These skills are fundamental for understanding relationships between variables in your data, which can be crucial for making data-driven conclusions.

10.1.1 Prerequisites

Before we begin, make sure you have R installed on your computer, along with the following R packages: `readr`, `dplyr`, and `ggplot2`. You can install these packages using the `install.packages()` function if you haven't already.

```
install.packages("readr")
install.packages("dplyr")
install.packages("ggplot2")
```

Also, download the same TCGA gene expression data (CSV file) to your working directory or specify the correct file path.

Download the file at <https://osf.io/yeun5>

10.1.2 Loading and Exploring the Data

We'll start by loading the TCGA gene expression data into R using the `read_csv` function from the `readr` package. This dataset contains information about various genes' expression levels across different samples, including cancer, metastatic, and normal samples.

```
library(readr)
library(dplyr)

# Load the TCGA gene expression data
tcga_data <- read_csv("~/Downloads/TCGA_cancer_genes_expression.csv")
```

Let's take a look at the first few rows of the dataset to understand its structure.

```
head(tcga_data)

## # A tibble: 6 x 25
##   ...1      TACSTD2  VTCN1   MUC1 NECTIN4 FOLH1  FOLR1 CD276   MSLN CLDN6 ERBB2
##   <chr>     <dbl>  <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <dbl>
## 1 43e715bf~  0.704  0       0.675  0.0862 7.21  0       52.8  0.0667 0.0970 1.88
## 2 1a5db9fc~  25.4   0       2.02   0.0728 23.6   0.122   78.8  0.956  0.255  7.78
## 3 93b382e4~  1.58   0       0.908  0.699   2.85  1.01   146.   0.0456 0.257  2.91
## 4 1f39dadd~  0.270  0.0910 0.0429  0.0165 1.16   0.279   48.5  0.0315 0.247  4.91
## 5 8c8c09b9~  0.412  0       0.115   0.0317 2.41   0.0492  42.3  0.270  0.126  1.49
## 6 85a86b91~  4.55   4.86   0.0421  0.0683 1.01   0.0225  20.6  0.0134 0.0182 13.5
## # i 14 more variables: MUC16 <dbl>, DLL3 <dbl>, CEACAM5 <dbl>, PVR <dbl>,
## # EPCAM <dbl>, PROM1 <dbl>, CD24 <dbl>, EGFR <dbl>, MET <dbl>,
## # TNFRSF10B <dbl>, tcga_tcga_barcode <chr>,
## # tcga_cgc_sample_sample_type <chr>, study <chr>, sample_type <chr>
```

Now, you should see a table with multiple columns, where each row represents a sample. The columns represent different genes' expression levels, and we also have columns indicating the sample type, study, and more.

10.1.3 Filtering the Data

For our analysis, we want to focus on cancer samples only. Let's filter the dataset to include only these samples and exclude normal and metastatic samples.

```
table(tcga_data$sample_type)

##
##      cancer metastatic      normal
##      10021        394        740

# Filter the data to include only cancer samples
tcga_cancer <- tcga_data %>%
  filter(sample_type == "cancer")
```

We use the `%>%` operator to chain commands together. Here, we first filter the dataset to include only rows where `sample_type` is “cancer.”

10.1.4 Creating a Scatter Plot

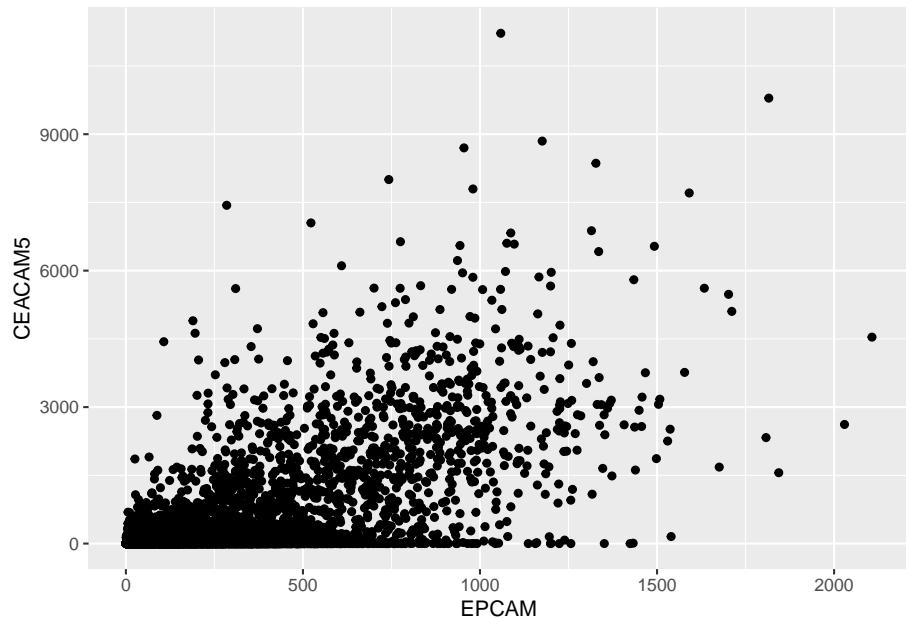
For docs check: <https://ggplot2.tidyverse.org/reference/ggplot.html>

Now, let's create a scatter plot to visualize the relationship between two gene expression levels. We'll plot the expression of the EPCAM gene on the x-axis and the CEACAM5 gene (CEA) on the y-axis.

Docs on geom_point: https://ggplot2.tidyverse.org/reference/geom_point.html

```
library(ggplot2)

# Create a scatter plot
ggplot(tcga_cancer, aes(x = EPCAM, y = CEACAM5)) +
  geom_point()
```



In the code above:

- We use ggplot() to initialize a new ggplot2 plot.
- Inside aes(), we specify the aesthetic mappings, where x represents the EPCAM gene expression and y represents the CEACAM5 gene expression.
- We add ‘geom_point()’ to plot the data points on the graph.

10.1.4.1 Interpretation

The scatter plot visually represents the relationship between the EPCAM and CEACAM5 gene expression levels. Each point on the plot corresponds to a cancer sample, with its EPCAM expression on the x-axis and CEACAM5 expression on the y-axis. If the points tend to fall on the diagonal line, it suggests a relationship between the two gene expressions.

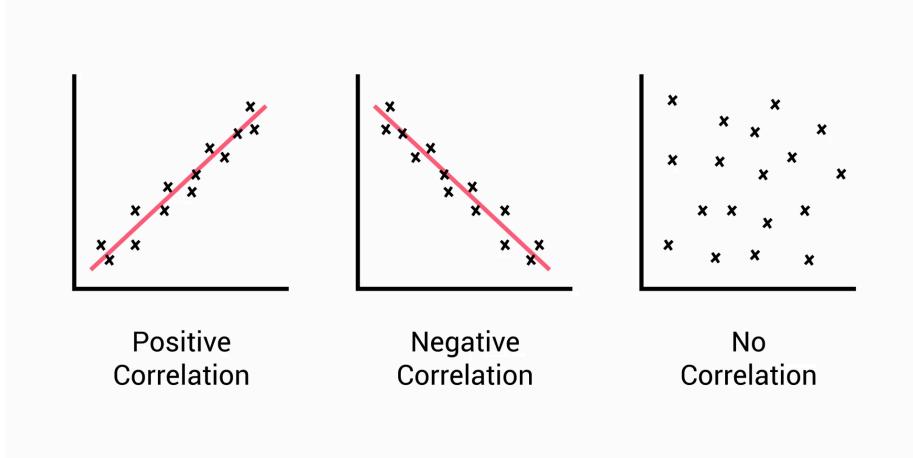
10.1.5 Calculating Correlation

To quantify the relationship between EPCAM and CEACAM5 gene expressions, we can calculate the Pearson correlation coefficient. This coefficient measures the strength and direction of the linear relationship between two variables.

```
# Calculate the Pearson correlation coefficient
correlation <- cor(tcga_cancer$EPCAM, tcga_cancer$CEACAM5)
correlation

## [1] 0.6324328
```

The output will be a value between -1 and 1, where:



- A positive value (closer to 1) indicates a positive correlation (both variables increase together).
- A negative value (closer to -1) indicates a negative correlation (one variable increases as the other decreases).
- A value close to 0 indicates little to no linear correlation.

In our example, the correlation coefficient (Pearson's r) is approximately 0.6324, which suggests a moderately positive correlation between EPCAM and CEACAM5 gene expressions among cancer samples.

10.1.6 Hypothesis Testing

Check docs here: <https://rdrr.io/r/stats/cor.test.html>

To determine if this correlation is statistically significant, we can perform a hypothesis test. In our case, we're interested in testing whether the correlation is significantly different from zero.

```
# Perform a correlation hypothesis test
cor_test_result <- cor.test(tcga_cancer$EPCAM, tcga_cancer$CEACAM5)
cor_test_result
```

```
##
## Pearson's product-moment correlation
##
## data: tcga_cancer$EPCAM and tcga_cancer$CEACAM5
## t = 81.722, df = 10019, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6205372 0.6440374
## sample estimates:
##        cor
## 0.6324328
```

The output will provide various statistics, including the t-value, degrees of freedom, and the p-value.

10.1.6.1 Interpretation

In the results, you'll see:

- The t-value, which measures the number of standard errors the correlation coefficient is away from zero.
- The degrees of freedom (df), which are related to the sample size.
- The p-value, which tells us whether the correlation is statistically significant.

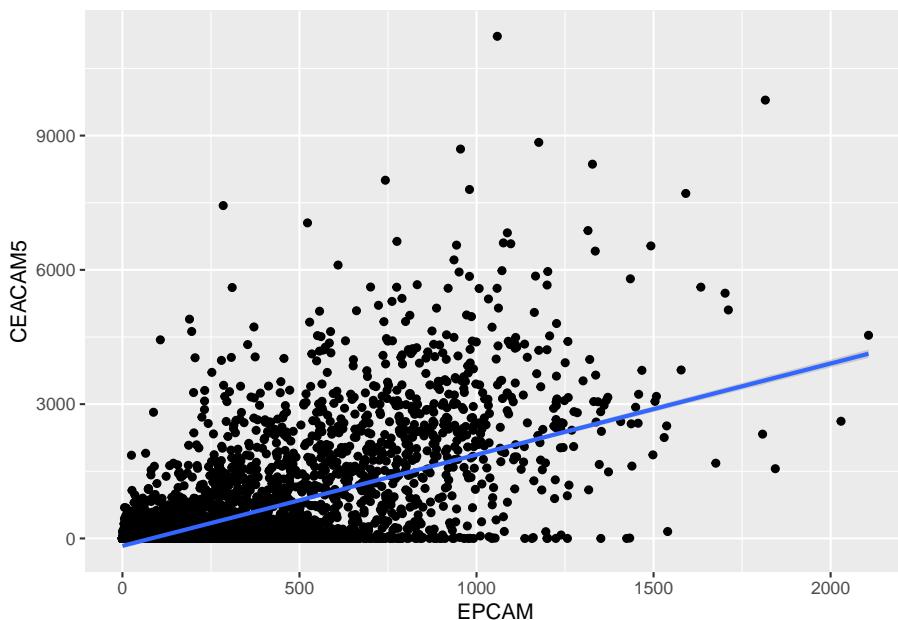
In our case, the p-value is very close to zero ($p\text{-value} < 2.2\text{e-}16$), indicating strong evidence against the null hypothesis (the true correlation is zero). Therefore, we can conclude that the correlation between EPCAM and CEACAM5 gene expressions is statistically significant. You need to keep in mind that in genomics data analysis, typically you have thousands of samples and you will inherently get tiny p values. In our case, focus on the effect size (in this case, the coefficient value which is 0.63).

10.1.7 Adding a Regression Line

Check docs on `geom_smooth` here: https://ggplot2.tidyverse.org/reference/geom_smooth.html

To further explore the relationship between the two gene expressions, we can add a linear regression line to our scatter plot using `geom_smooth()`.

```
# Create a scatter plot with a regression line
ggplot(tcga_cancer, aes(x = EPCAM, y = CEACAM5)) +
  geom_point() +
  geom_smooth(method = "lm")
```



The `geom_smooth()` function with `method = "lm"` fits a linear regression line to the data points, helping us visualize the trend more clearly.

The regression line provides a visual representation of how one gene's expression (EPCAM) changes concerning the other (CEACAM5). If the line slopes upward, it suggests a positive correlation, while a downward slope indicates a negative correlation.

10.1.8 Conclusion

In this lesson, we've covered the basics of creating scatter plots, calculating correlation coefficients, and performing hypothesis tests using R. These skills are essential for exploring relationships between variables in your data, whether you're analyzing gene expressions, financial data, or any other dataset. Remember that correlation does not imply causation, so it's essential to interpret your findings carefully and consider the context of your analysis.

10.2 Understanding Distributions with Histograms

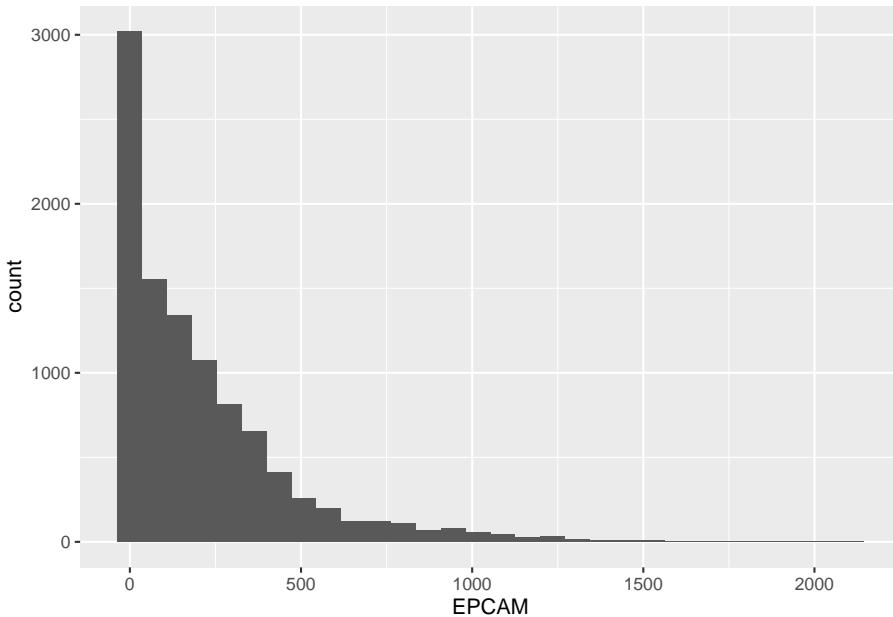
In this section, we will explore another powerful data visualization tool: histograms. Histograms are especially useful for understanding the distribution of a single numerical variable. We will use R and the `ggplot2` package to create histograms and customize them to gain insights into the distribution of gene expression levels.

For this course we're using the data we generated in previous lesson.

10.2.1 Creating a Basic Histogram

Let's start by creating a basic histogram to visualize the distribution of the EPCAM gene expression levels across all cancer samples.

```
# Create a basic histogram for EPCAM gene expression
ggplot(tcga_cancer, aes(x = EPCAM)) +
  geom_histogram()
```



In this code:

- We use `ggplot()` to initialize a new ggplot2 plot.
- Inside `aes()`, we specify that we want to map the EPCAM gene expression values to the x-axis.
- We add `geom_histogram()` to create the histogram.

The resulting plot will display the EPCAM gene expression levels on the x-axis and the count of samples falling into each “bin” on the y-axis. A bin is a range of values, and the height of each bar represents how many samples have expression levels within that range.

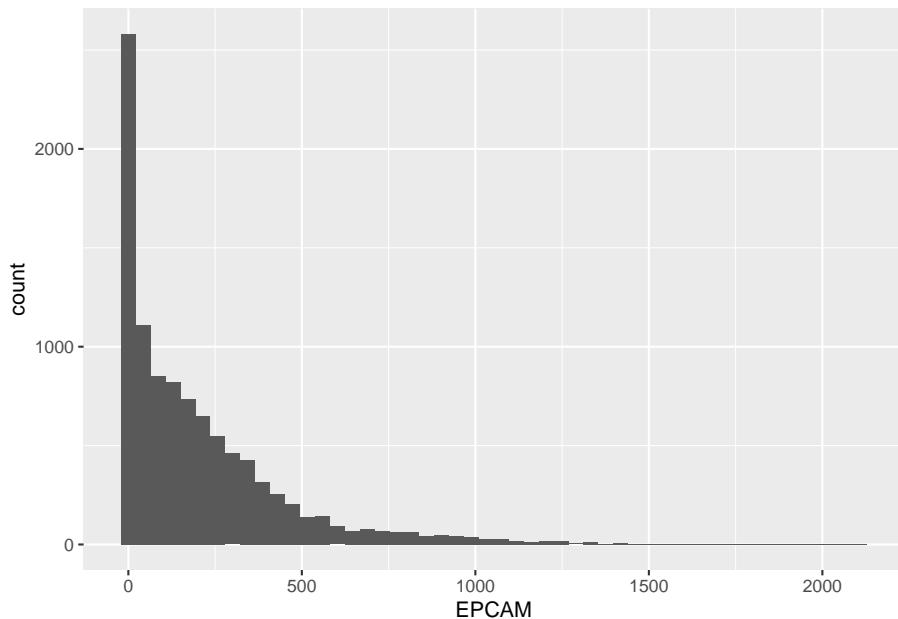
10.2.2 Customizing Histogram Bins

You can customize the granularity of your histogram by changing the number of bins or specifying the bin size. This allows you to get a more detailed or broader view of the data distribution.

10.2.2.1 Changing the Number of Bins

To change the number of bins, you can use the `bins` parameter within `geom_histogram()`. Increasing the number of bins provides more detail.

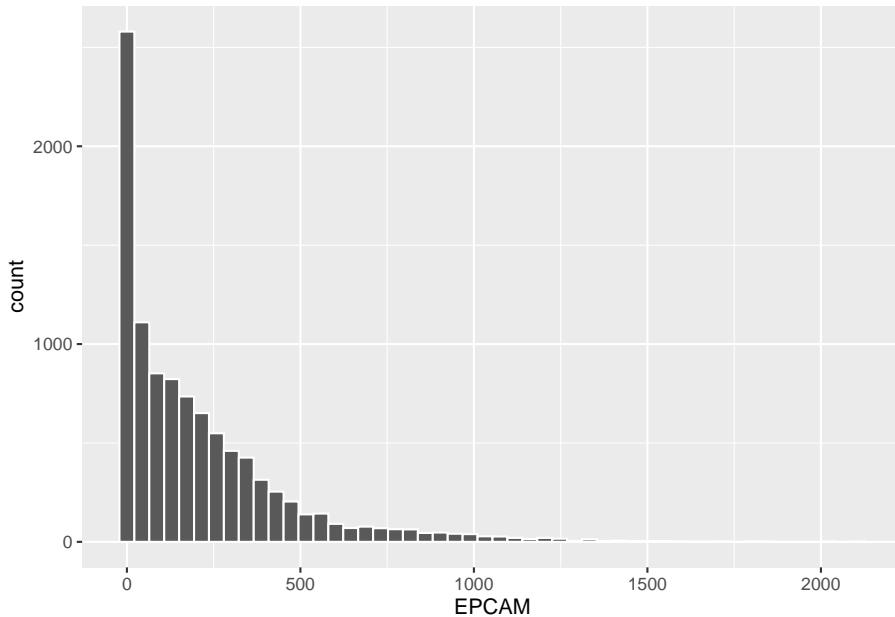
```
# Create a histogram with 50 bins
ggplot(tcga_cancer, aes(x = EPCAM)) +
  geom_histogram(bins = 50)
```



10.2.2.2 Customizing Bin Borders

By default, the borders of the bins in the histogram are not very visible. You can change the border color to white to make them more distinct.

```
# Create a histogram with white bin borders
ggplot(tcga_cancer, aes(x = EPCAM)) +
  geom_histogram(bins = 50, color = "white")
```



Changing the bin border color to white makes it easier to distinguish between adjacent bins.

10.2.3 Conclusion

Histograms are valuable tools for visualizing the distribution of a single numerical variable, helping you understand the underlying data structure. By customizing the number of bins, bin sizes, and bin borders, you can tailor your histograms to reveal the level of detail you need. Whether you are analyzing gene expression data or any other quantitative data, histograms are an essential part of your data exploration toolkit.

10.3 Visualizing Data Distribution with Boxplots and Violin Plots

In this lesson, we will explore how to visualize the distribution of gene expression data across different cancer types using boxplots and violin plots in R. These graphical tools are invaluable for gaining insights into the spread and central tendency of data within different categories.

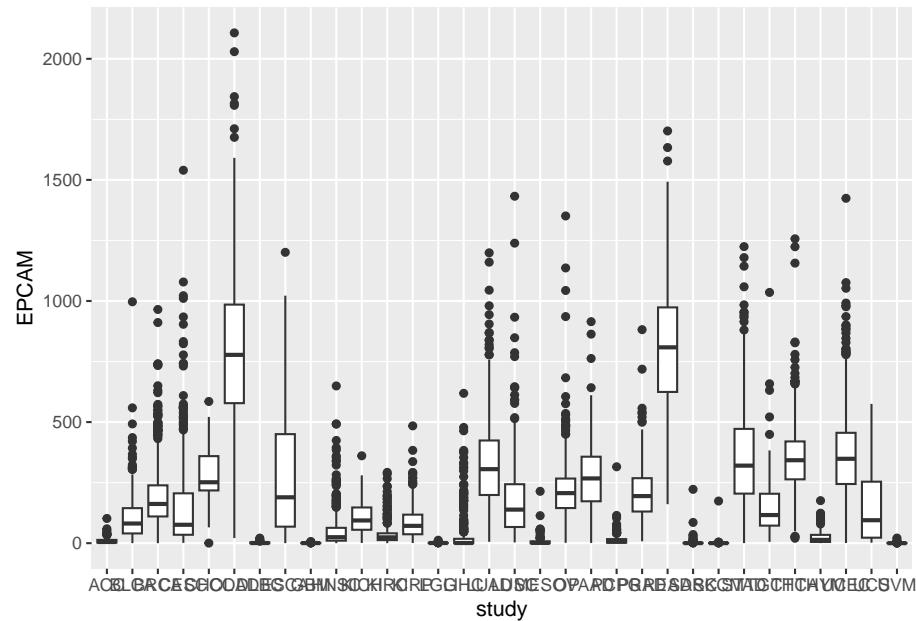
We will continue to work with the TCGA gene expression dataset, specifically focusing on cancer samples (tcga_cancer). This dataset contains various gene expression measurements across different cancer types.

10.3.1 Creating a Basic Boxplot

Let's start by creating a basic boxplot to visualize the distribution of the EPCAM gene expression across different cancer types.

```
library(ggplot2)

ggplot(tcga_cancer, aes(x = study, y = EPCAM)) +
  geom_boxplot()
```



In this code:

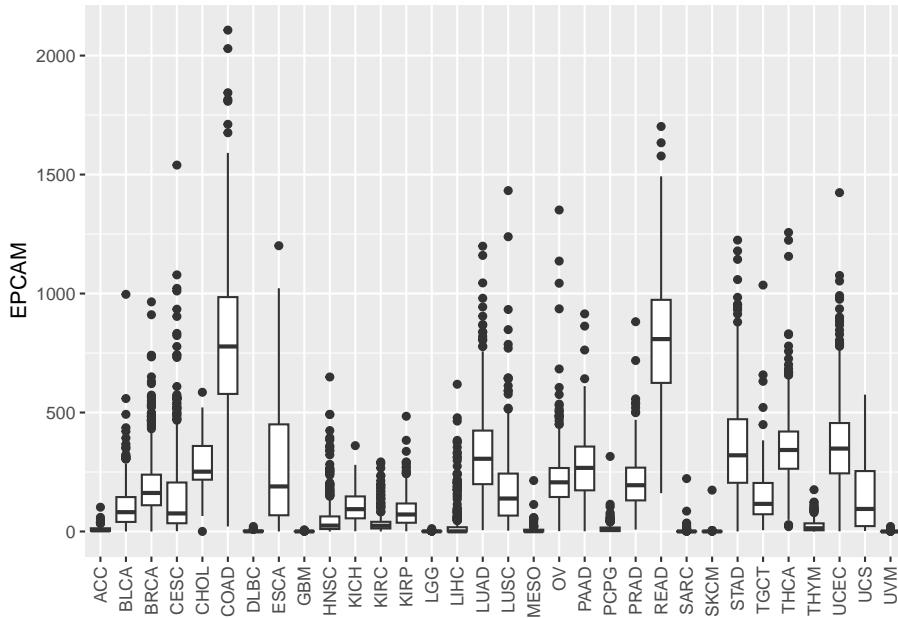
- We use `ggplot()` to initialize the plot.
- We map the x-axis to `study` (representing cancer types) and the y-axis to EPCAM gene expression.
- We add `geom_boxplot()` to create the boxplots.

10.3.2 Rotating X-axis Labels

You may notice that the x-axis labels (cancer types) overlap. To make the plot more readable, we can rotate the x-axis labels by 90 degrees and remove the x-axis label using the `theme` function.

10.3. VISUALIZING DATA DISTRIBUTION WITH BOXPLOTS AND VIOLIN PLOTS 167

```
ggplot(tcga_cancer, aes(x = study, y = EPCAM)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_blank())
```

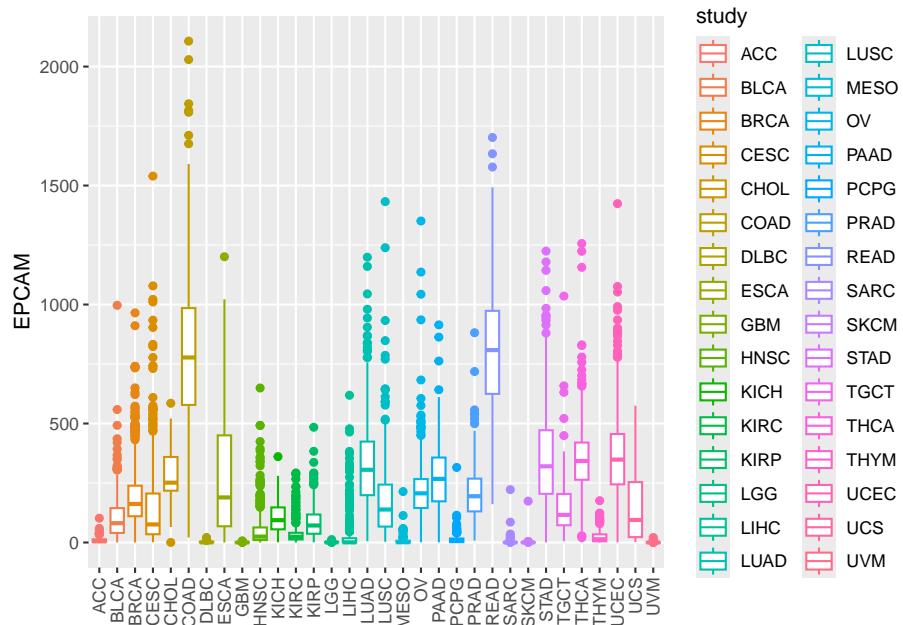


Now, the x-axis labels are more legible.

10.3.3 Adding Color to Boxplots

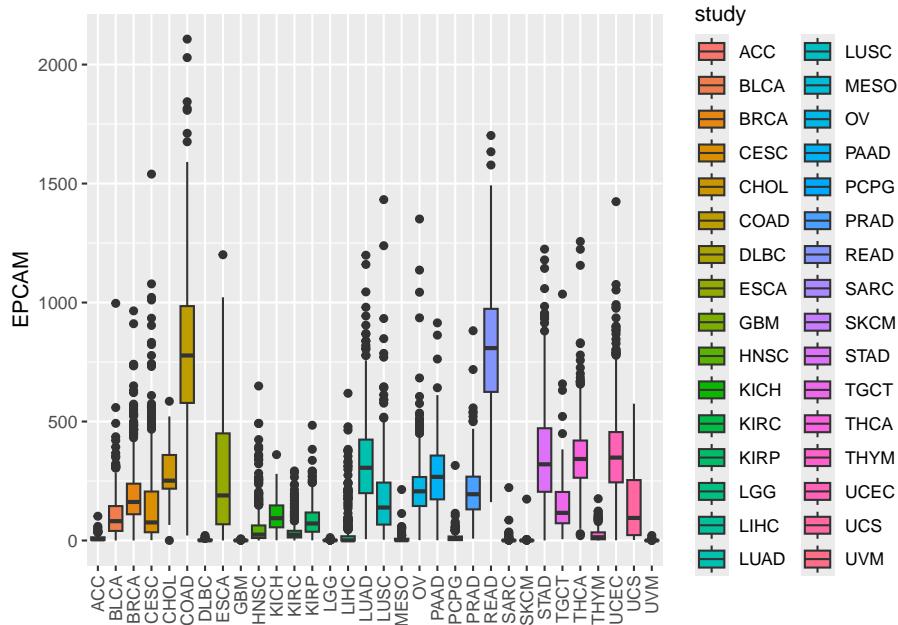
To distinguish between cancer types more effectively, let's add color to the boxplots by mapping the color aesthetic to the study.

```
ggplot(tcga_cancer, aes(x = study, y = EPCAM)) +
  geom_boxplot(aes(color = study)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_blank())
```



Alternatively, you can use `fill` to color the boxes instead of the outlines:

```
ggpplot(tcgacancer, aes(x = study, y = EPCAM)) +
  geom_boxplot(aes(fill = study)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_blank())
```



By default, ggplot2 uses a default color palette, but you can specify colors manually if needed.

10.3.4 Customizing Color Palettes

In case you want to define your color palette for the cancer types, you can use the Polychrome package.

```
# there are total 32 cancer types
length(unique(tcga_cancer$study))
```

```
## [1] 32
```

Here's how to create a custom color palette for 32 cancer types:

```
# install.packages("Polychrome")
library(Polychrome)

# There are a total of 32 cancer types
length(unique(tcga_cancer$study))

## [1] 32
```

```
# Create a custom color palette with Polychrome
P32 <- createPalette(32, c("#FF0000", "#00FF00", "#0000FF"), range = c(30, 80))

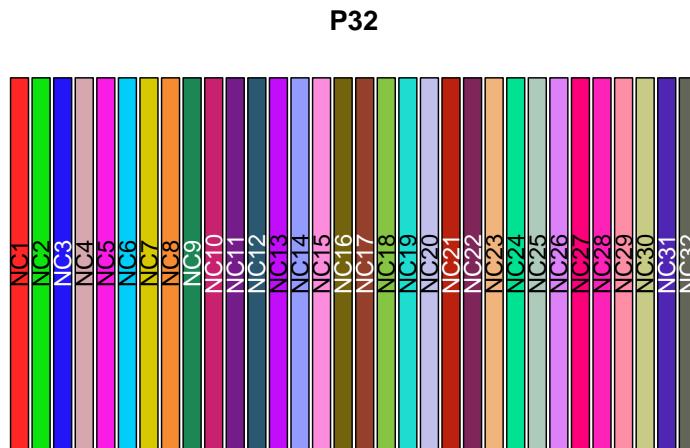
P32
```

```
##      NC1      NC2      NC3      NC4      NC5      NC6      NC7      NC8
## "#FF2626" "#0DE40D" "#2216F8" "#D6A7AE" "#FA1CE5" "#00CDFA" "#D6C900" "#F78B32"
##      NC9      NC10     NC11     NC12     NC13     NC14     NC15     NC16
## "#1C8755" "#C8226E" "#751C8B" "#2A5871" "#C20DFB" "#949BFE" "#FE8CDD" "#73630D"
##      NC17     NC18     NC19     NC20     NC21     NC22     NC23     NC24
## "#94402A" "#86C542" "#1CDDCF" "#C1COED" "#BB220D" "#7E265A" "#F2B47D" "#00E291"
##      NC25     NC26     NC27     NC28     NC29     NC30     NC31     NC32
## "#ADCABB" "#DD81F9" "#FB0078" "#FF22B9" "#FD8EA5" "#C7CB86" "#4F26B4" "#626558"
```

You see the hex code for the colors.

Now, we have a custom color palette P32 with 32 unique colors. You can visualize the colors using `swatch(P32)` in the polychrom package.

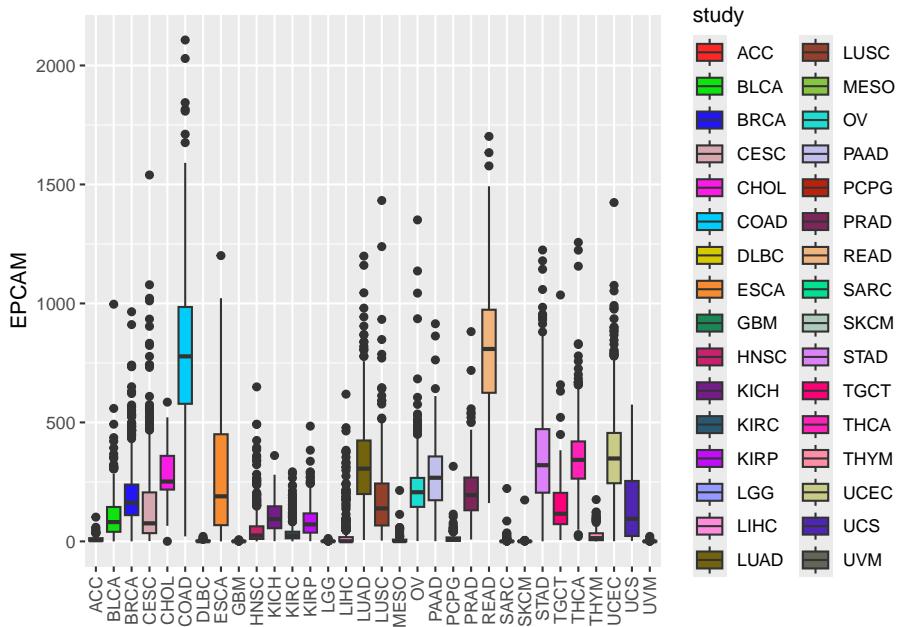
```
swatch(P32)
```



10.3.5 Applying Custom Color Palette

To use the custom color palette with your boxplot, use `scale_fill_manual()` or `scale_color_manual()` to map the colors manually to the study variable.

```
ggplot(tcga_cancer, aes(x = study, y = EPCAM)) +
  geom_boxplot(aes(fill = study)) +
  scale_fill_manual(values = unname(P32)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_blank())
```



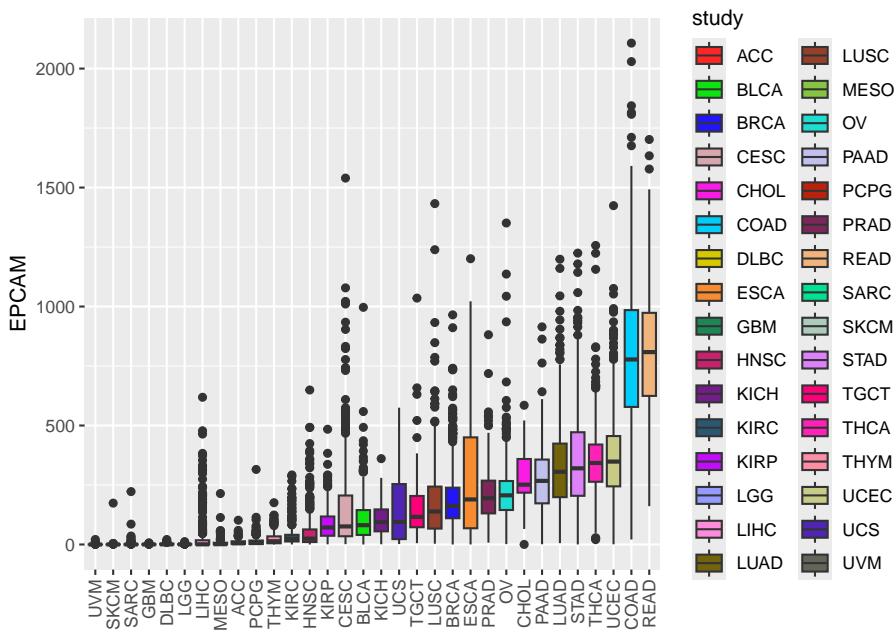
Note, you need to remove the names of the named color vector using `unname`.

Now, the boxplot colors are based on your custom palette, making it more visually appealing and distinct.

10.3.6 Reordering Boxplots

To reorder the boxes according to the median level of EPCAM expression, you can use the `forcats::fct_reorder()` function within your `aes()` mapping.

```
ggplot(tcga_cancer, aes(x = study %>%
                           forcats::fct_reorder(EPCAM, median),
                           y = EPCAM)) +
  geom_boxplot(aes(fill = study)) +
  scale_fill_manual(values = unname(P32)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_blank())
```



The `forcats::fct_reorder()` function takes two arguments:

- The first argument (`study`) is the factor variable whose levels we want to reorder.
- The second argument (`EPCAM`) is the numeric variable based on which the reordering will be done (in this case, the median expression levels of EPCAM).

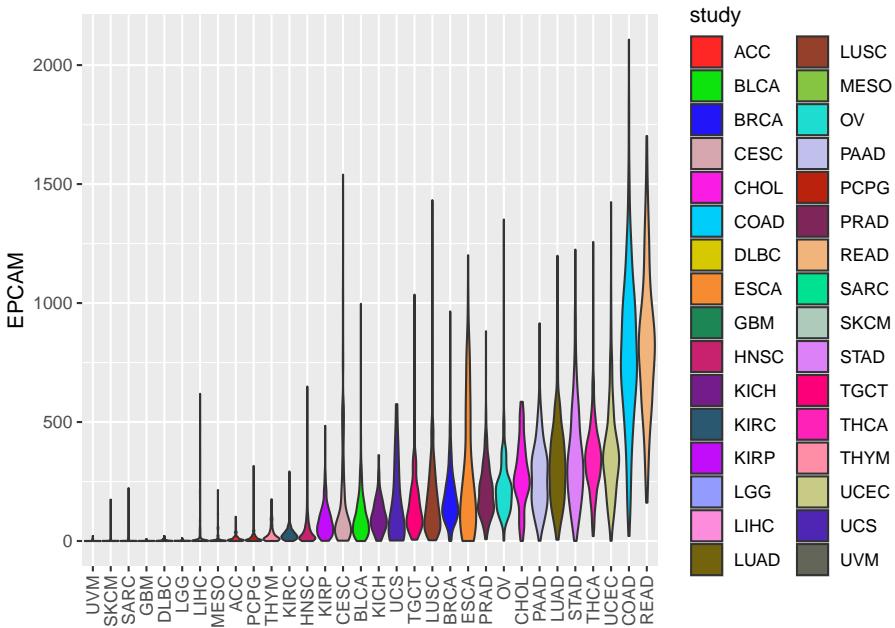
As a result of this reordering, the levels of the study factor will be rearranged from low to high median EPCAM expression levels. Now, the boxes are arranged from low to high EPCAM expression levels, providing a clearer view of the data distribution.

10.3.7 Using Violin Plots

Violin plots combine a boxplot with a kernel density plot, allowing you to visualize both the summary statistics and the entire data distribution.

```
ggplot(tcga_cancer, aes(x = study %>%
                           forcats::fct_reorder(EPCAM, median),
                           y = EPCAM)) +
  geom_violin(aes(fill = study), scale = "width") +
  scale_fill_manual(values = unname(P32)) +
```

```
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
      axis.title.x = element_blank())
```



In this code, we replaced `geom_boxplot` with `geom_violin` to create a violin plot. The width of the violin plots is scaled for better visibility.

10.3.8 Conclusion

In this lesson, we've explored how to visualize data distribution using box-plots and violin plots in R. These techniques are essential for understanding the spread and central tendencies of data within different categories or groups. By customizing colors, reordering boxes, and using violin plots, you can effectively communicate insights from your data to others.

10.4 Creating Bar Plots to Visualize Median EPCAM Levels by Cancer Type

In this section, we will learn how to create informative bar plots using R. Specifically, we'll use bar plots to display the median levels of the EPCAM gene expression for different types of cancer. This visualization will help us understand how EPCAM expression varies across various cancer studies. We will

also explore techniques to reorder and customize our bar plots for better data interpretation.

10.4.1 Calculating Median EPCAM Levels by Cancer Type

The median EPCAM expression levels represent the middle value of EPCAM gene expression measurements for different types of cancer studies, providing a central measure to understand how this gene behaves in each type of cancer.

To begin, we'll calculate the median EPCAM expression levels for each cancer study type. We will create a summary table that includes the cancer study names and their corresponding median EPCAM values.

```
# Calculate median EPCAM levels by cancer type
EPCAM_median <- tcga_cancer %>%
  group_by(study) %>%
  summarize(median_EPCAM = median(EPCAM))

# Display the summary table
EPCAM_median
```

study	median_EPCAM
ACC	4.83
BLCA	81.1
BRCA	162.
CESC	75.9
CHOL	251.
COAD	778.
DLBC	0.578
ESCA	189.
GBM	0.307
HNSC	24.7

A tibble: 32 x 2
study median_EPCAM
<chr> <dbl>
1 ACC 4.83
2 BLCA 81.1
3 BRCA 162.
4 CESC 75.9
5 CHOL 251.
6 COAD 778.
7 DLBC 0.578
8 ESCA 189.
9 GBM 0.307
10 HNSC 24.7
i 22 more rows

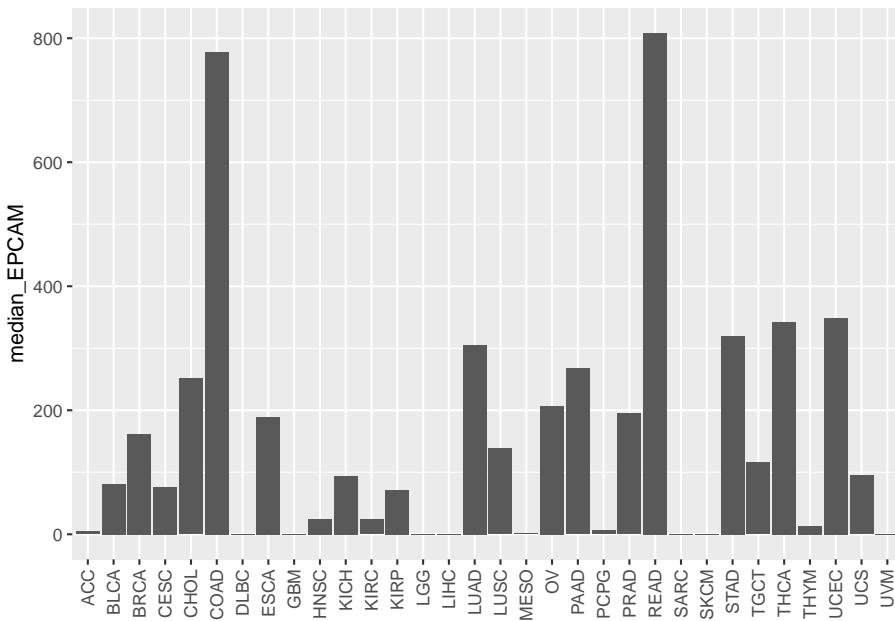
This table now shows the median EPCAM expression levels for each cancer study type, making it easier to compare across different types of cancer.

10.4.2 Creating a Basic Bar Plot

Now, let's create a simple bar plot to visualize these median EPCAM levels. We will use `geom_bar()` to create bars representing the median values.

```
library(ggplot2)

# Create a basic bar plot
ggplot(EPCAM_median, aes(x = study, y = median_EPCAM)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_blank())
```



In this code:

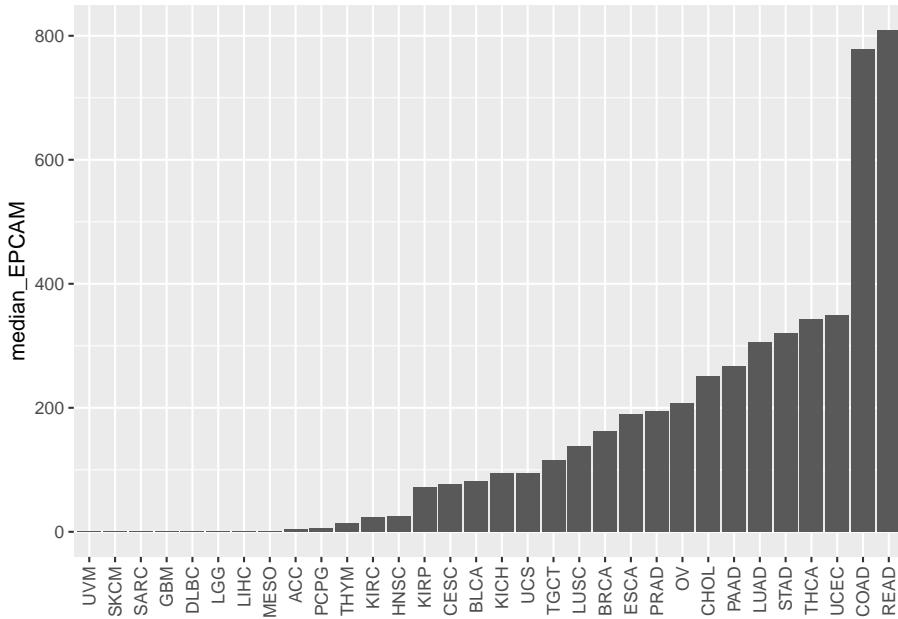
- We use `ggplot()` to initialize the plot.
- `aes()` specifies the aesthetic mappings, where `x` represents the cancer study names, and `y` represents the median EPCAM values.
- `geom_bar(stat = "identity")` creates bars where the height of each bar corresponds to the median EPCAM value.
- `theme()` is used to customize the appearance of the plot, including rotating the x-axis labels for better readability.

The basic bar plot displays the median EPCAM levels for each cancer study type, but the bars are not ordered based on the values. We can improve the plot by reordering the bars according to the median EPCAM levels.

10.4.3 Reordering the Bars

To make our bar plot more informative, let's reorder the bars based on the median EPCAM values. We can use the `forcats::fct_reorder()` function to achieve this.

```
# Reorder the bars based on median EPCAM values
ggplot(EPCAM_median, aes(x = study %>%
                           forcats::fct_reorder(median_EPCAM),
                           y = median_EPCAM)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_blank())
```



In this code:

We use `forcats::fct_reorder()` within `aes()` to reorder the study variable based on the `median_EPCAM` values.

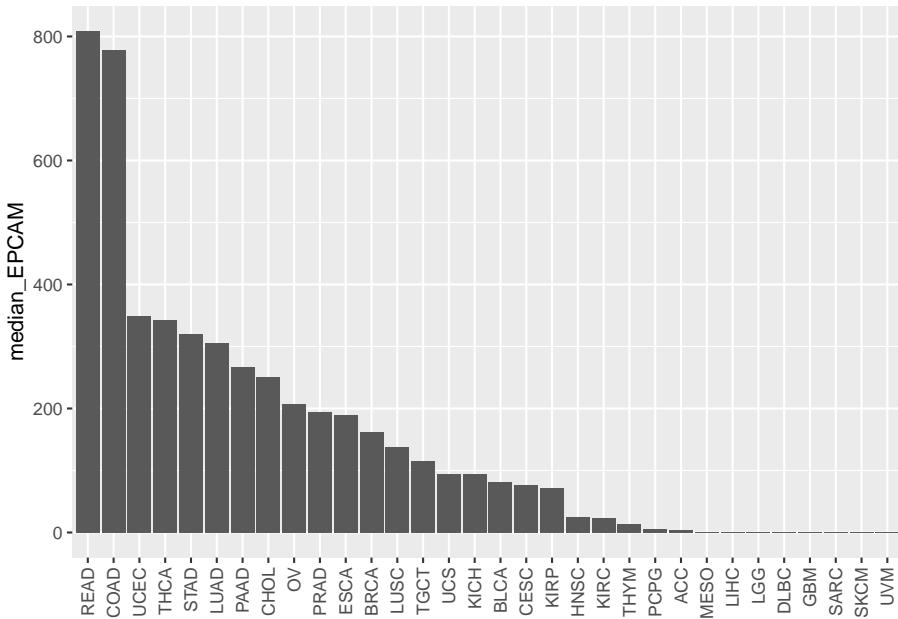
The `median_EPCAM` values determine the order of the bars from lowest to highest.

Now, the bars are reordered based on the median EPCAM levels, allowing for easier comparison between different cancer study types. This visualization provides insights into which studies exhibit higher or lower median EPCAM expression levels.

10.4.4 Reversing the Order

If you want to reverse the order of the bars, you can use the `.desc = TRUE` argument with `forcats::fct_reorder()`.

```
# Reverse the order of the bars
ggplot(EPCAM_median, aes(x = study %>%
                           forcats::fct_reorder(median_EPCAM, .desc = TRUE),
                           y = median_EPCAM)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title.x = element_blank())
```



By specifying `.desc = TRUE`, the bars will be reordered in descending order based on the median EPCAM values.

10.4.5 Note on factor levels

xxxx

10.4.6 Conclusion

In this lesson, we have learned how to create bar plots to visualize median EPCAM expression levels by different cancer study types. We explored the

process of reordering bars based on values, allowing us to better compare and interpret the data. The order of the bar depends on the factor levels of the study and we used `fct_reorder` to accomplish that. Remember that effective data visualization can greatly enhance your ability to communicate insights and make data-driven decisions.

Chapter 11

Introduction to BioConductor

In the ever-evolving field of bioinformatics and computational biology, researchers and scientists rely on specialized tools and packages to analyze and interpret biological data efficiently. **Bioconductor**, a powerful and comprehensive suite of R packages, plays a pivotal role in addressing the unique challenges posed by genomics, transcriptomics, proteomics, and other biological data domains.

This section, “Introduction to Bioconductor in R,” will serve as your gateway to this remarkable resource. We will explore the fundamentals of Bioconductor, its significance in the realm of life sciences, and how you can harness its capabilities to unlock valuable insights from complex biological datasets. Whether you are a biologist, bioinformatician, or data scientist, understanding Bioconductor is a crucial step towards advancing your research and data analysis endeavors in the field of biology. Let’s embark on this exciting journey into the world of Bioconductor, where data meets discovery.

11.1 Introduction to BiocManager

BiocManager is an R package that serves as the primary interface for managing Bioconductor packages, which are extensions of R developed specifically for bioinformatics applications. Bioconductor itself is a project that provides tools for the analysis and comprehension of high-throughput genomic data, including but not limited to next-generation sequencing (NGS), microarrays, and proteomics.

11.1.1 Why Use BiocManager?

The role of BiocManager is ensuring compatibility among Bioconductor packages and between these packages and your version of R. It simplifies the process of installing Bioconductor packages, managing dependencies, and keeping packages up-to-date with the latest releases, thereby fostering a stable and efficient bioinformatics workflow.

11.1.2 Installing BiocManager

To get started with BiocManager, you first need to install it from CRAN (the Comprehensive R Archive Network), which can be done using the following command in R:

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

Once installed, you can load `BiocManager` just like any other R package:

```
library(BiocManager)
```

11.1.3 Installing Bioconductor Packages

With `BiocManager` loaded, installing Bioconductor packages is straightforward. Suppose you want to install the `GenomicRanges` package; you can do so with the following command:

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

`BiocManager` automatically resolves and installs any dependencies, ensuring that all required packages are installed for the `GenomicRanges` package to function correctly.

11.1.4 Updating Bioconductor Packages

Keeping your Bioconductor packages up-to-date is crucial for accessing the latest features, improvements, and bug fixes. `BiocManager` facilitates this through the `install` function, which also checks for and updates any out-of-date packages:

```
BiocManager::install()
```

Running this command without specifying a package name updates all installed Bioconductor packages to their latest versions.

11.1.5 Checking for Valid Bioconductor Versions

Compatibility between your R version and Bioconductor packages is vital for smooth bioinformatics analyses. BiocManager offers a function to validate this compatibility:

```
BiocManager::valid()
```

This command checks that all installed Bioconductor packages are compatible with each other and with your current version of R, providing a report of any inconsistencies.

11.1.6 Conclusion

BiocManager is an indispensable tool for bioinformatics practitioners working with R and **Bioconductor**. It simplifies package management, ensuring that researchers can focus on their analyses without being bogged down by software compatibility issues. By leveraging **BiocManager**, you can maintain a cutting-edge bioinformatics toolkit, fully equipped to tackle the challenges of genomic data analysis.

11.2 Working with Genomic Coordinates and Regions in R

Genomic coordinates are fundamental in the field of genomics. Whether you're dealing with genes, regulatory elements, ChIP-seq peaks, or mutation calling data, they are all represented as genomic coordinates. In R, you can efficiently handle genomic coordinates and regions using the **GenomicRanges** package and related tools. In this guide, we'll explore how to work with genomic ranges, extract relevant information, and perform common operations.

For a complete overview check docs here: <https://bioconductor.org/packages/release/bioc/vignettes/GenomicRanges/inst/doc/GenomicRangesIntroduction.html>

11.2.1 Introduction to GenomicRanges

The **GenomicRanges** package in R provides data structures for storing and manipulating genomic ranges. A genomic range typically includes information about the chromosome (seqname), the start and end positions, and the strand of the sequence.

1. Every gene or regulatory element (promoters, enhancers) in the genome can be represented in chromosome: start-end format.
2. You get a peak file from a ChIP-seq experiment. The peak file is usually represented in a at least 3-column bed format: chromosome, start and end.
3. You get a mutation calling VCF file. You will have the chromosome and position of that single variant.

Let's start by creating a simple genomic range:

```
library(GenomicRanges)

# Create a GenomicRanges object
gr <- GRanges(seqnames = "chr1",
              ranges = IRanges(1:10, width = 3))

gr

## GRanges object with 10 ranges and 0 metadata columns:
##   seqnames      ranges strand
##   <Rle> <IRanges>  <Rle>
## [1] chr1      1-3     *
## [2] chr1      2-4     *
## [3] chr1      3-5     *
## [4] chr1      4-6     *
## [5] chr1      5-7     *
## [6] chr1      6-8     *
## [7] chr1      7-9     *
## [8] chr1      8-10    *
## [9] chr1      9-11    *
## [10] chr1     10-12   *
## -----
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

In this example, we've created a GenomicRanges object for chromosome 1 with ten intervals of width 3. We did not specify the strand, so it is *. Alternatively, we can specify the genomic regions are on the + strand.

```
GRanges(seqnames = "chr1",
        ranges = IRanges(1:10, width = 3),
        strand = "+")
```

```
## GRanges object with 10 ranges and 0 metadata columns:
```

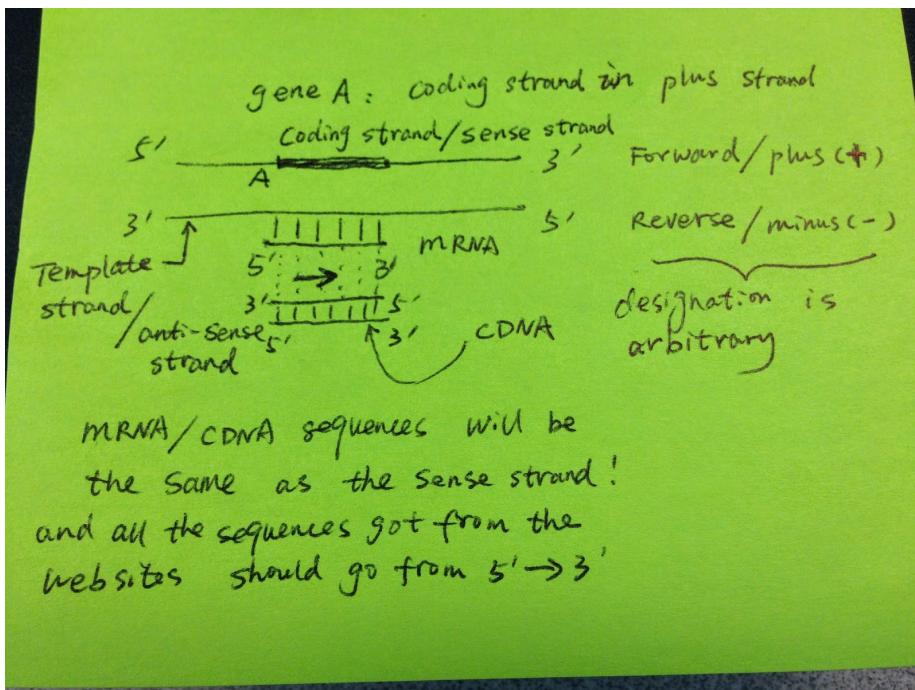
11.2. WORKING WITH GENOMIC COORDINATES AND REGIONS IN R183

```

##      seqnames      ranges strand
##      <Rle> <IRanges>  <Rle>
## [1] chr1      1-3      +
## [2] chr1      2-4      +
## [3] chr1      3-5      +
## [4] chr1      4-6      +
## [5] chr1      5-7      +
## [6] chr1      6-8      +
## [7] chr1      7-9      +
## [8] chr1      8-10     +
## [9] chr1      9-11     +
## [10] chr1     10-12    +
## -----
## seqinfo: 1 sequence from an unspecified genome; no seqlengths

```

To understand the standness, read this blog post by me.



11.2.2 Basic Operations with GenomicRanges

These operations are commonly used in genomics data analysis to perform tasks such as calculating the length of genomic features, extracting specific regions of interest, and modifying intervals for downstream analysis. `GenomicRanges`

provides a flexible and efficient way to work with genomic intervals, which is essential for tasks like annotation, visualization, and statistical analysis of genomics data.

11.2.3 Calculating width of Each Genomic Interval

Genomic intervals can represent various features in a genome, such as genes, exons, or regulatory regions. Knowing the width of these intervals is crucial when analyzing genomic data. For example, you might want to calculate the size of a gene or measure the distance between two regulatory elements. The `width()` function helps you obtain this information quickly.

```
width(gr)
```

```
## [1] 3 3 3 3 3 3 3 3 3
```

This function calculates the width (or length) of each genomic interval in the `GenomicRanges` object `gr`. In genomics, the width typically represents the number of base pairs or genomic coordinates covered by each interval.

11.2.4 Start and End Positions

Genomic intervals are defined by their start and end positions along a chromosome. These functions allow you to extract these positions, which can be essential for tasks like determining the transcription start site of a gene or identifying the boundaries of a specific genomic region.

11.2.4.1 Getting start position

```
start(gr)
```

```
## [1] 1 2 3 4 5 6 7 8 9 10
```

This function retrieves the starting position (or the leftmost coordinate) of each genomic interval in the `GenomicRanges` object `gr`. It tells you where each interval begins along the genome.

11.2.4.2 Getting end position

```
end(gr)
```

```
## [1] 3 4 5 6 7 8 9 10 11 12
```

This function retrieves the ending position (or the rightmost coordinate) of each genomic interval in the `GenomicRanges` object `gr`. It tells you where each interval ends along the genome.

11.2.5 Strand Information

In genomics, it's important to know the orientation of genomic features. The strand information (+ or -) indicates whether a feature is on the forward (+) or reverse (-) strand of DNA. This can be crucial for understanding gene transcription direction, reading frames, and other biological processes.

```
strand(gr)
```

```
## factor-Rle of length 10 with 1 run
##   Lengths: 10
##   Values : *
## Levels(3): + - *
```

Genomic intervals can be associated with a strand information to represent the directionality of a genomic feature. The `strand` function retrieves the strand information for each interval. The strand can be either “+” for the forward strand, “-” for the reverse strand, or “*” for strand-agnostic intervals.

11.2.6 Shifting the Genomic Range

Sometimes, you need to shift genomic intervals to examine neighboring regions. For instance, you might want to find regions that overlap with a gene's promoter, which is typically located upstream of the transcription start site. Shifting intervals allows you to explore nearby genomic areas easily.

```
gr + 1
```

```
## GRanges object with 10 ranges and 0 metadata columns:
##   seqnames      ranges strand
##   <Rle> <IRanges> <Rle>
##   [1] chr1      0-4      *
##   [2] chr1      1-5      *
```

```

## [3] chr1    2-6    *
## [4] chr1    3-7    *
## [5] chr1    4-8    *
## [6] chr1    5-9    *
## [7] chr1    6-10   *
## [8] chr1    7-11   *
## [9] chr1    8-12   *
## [10] chr1   9-13   *
## -----
## seqinfo: 1 sequence from an unspecified genome; no seqlengths

```

This operation demonstrates how you can manipulate the genomic intervals in `gr`. Here, you are adding 1 to each start **AND** end position in the intervals, effectively expanding them by one base left and right .

```
width(gr+1)
```

```
## [1] 5 5 5 5 5 5 5 5 5 5
```

11.2.7 Subsetting

Genomic data can be extensive, and you often need to focus on specific regions of interest. Subsetting helps you extract only the relevant intervals from a larger dataset. This is especially useful when you want to analyze a particular set of genes or genomic regions.

```
gr[1:2]
```

```

## GRanges object with 2 ranges and 0 metadata columns:
##   seqnames      ranges strand
##   <Rle> <IRanges> <Rle>
## [1] chr1     1-3    *
## [2] chr1     2-4    *
## -----
## seqinfo: 1 sequence from an unspecified genome; no seqlengths

```

This code subset the GenomicRanges object `gr` to select the first two intervals. It returns a new GenomicRanges object containing only those intervals.

11.2.8 Flanking Regions

Imagine you have a specific location in the DNA, and you want to study not only that location but also the regions right before and after it. `flank` lets you

11.2. WORKING WITH GENOMIC COORDINATES AND REGIONS IN R187

do this. For example, if you're interested in a particular gene, you can use `flank` to include a bit of the DNA sequence before and after that gene. This helps you see the surrounding context and understand how the gene fits into the bigger picture.

```
flank(gr, 2)
```

```
## GRanges object with 10 ranges and 0 metadata columns:
##   seqnames      ranges strand
##   <Rle> <IRanges> <Rle>
## [1] chr1     -1-0    *
## [2] chr1     0-1    *
## [3] chr1     1-2    *
## [4] chr1     2-3    *
## [5] chr1     3-4    *
## [6] chr1     4-5    *
## [7] chr1     5-6    *
## [8] chr1     6-7    *
## [9] chr1     7-8    *
## [10] chr1    8-9    *
## -----
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

`flank` is useful for tasks like expanding genomic regions of interest to capture nearby regions or creating control regions around known features for downstream analysis. It is commonly used in genomics research to study the context around specific genomic locations.

11.2.9 Resizing

Think of a situation where you have many different pieces of DNA (intervals), and they're all different lengths. Maybe you want to compare them or count something in each of them. It's easier to work with them if they're all the same size. That's what `resize` does. It makes sure that all the pieces of DNA are the same length, so you can compare or analyze them more easily.

```
resize(gr, 10)
```

```
## GRanges object with 10 ranges and 0 metadata columns:
##   seqnames      ranges strand
##   <Rle> <IRanges> <Rle>
## [1] chr1     1-10    *
## [2] chr1     2-11    *
```

```

## [3] chr1    3-12    *
## [4] chr1    4-13    *
## [5] chr1    5-14    *
## [6] chr1    6-15    *
## [7] chr1    7-16    *
## [8] chr1    8-17    *
## [9] chr1    9-18    *
## [10] chr1   10-19   *
## -----
## seqinfo: 1 sequence from an unspecified genome; no seqlengths

```

`resize` is used to standardize the length of genomic intervals, which can be useful for comparing or analyzing regions of interest with consistent sizes. It is often applied to ensure that intervals have the same width, making them suitable for various downstream analyses, such as counting reads or comparing features.

11.2.10 0 based and 1 based coordinate system

One needs to be aware that there are two genomics coordinate systems: 1 based and 0 based. There is really no mystery between these two. You EITHER count start at 0 OR at 1. However, this can make confusions when analyzing genomic data and one may make mistakes if not keep it in mind. Read <https://www.biostars.org/p/84686/>.

The reason that why it matters is that python index starts at 0 while R starts at 1.

Make sure you understand how different bioinformatics format use different coordinate system.

11.2. WORKING WITH GENOMIC COORDINATES AND REGIONS IN R189

Table 9-1. Range types of common bioinformatics formats

Format/library	Type
BED	0-based
GTF	1-based
GFF	1-based
SAM	1-based
BAM	0-based
VCF	1-based
BCF	0-based
Wiggle	1-based
GenomicRanges	1-based
BLAST	1-based
GenBank/EMBL Feature Table	1-based

11.2.11 Other packages

In genomics research, we often work with genomic data in various formats such as GTF (Gene Transfer Format) and BED files. To facilitate this, we have a few essential packages at our disposal:

- **AnnotationHub:** This package provides access to a wide range of genome annotations, including GTF files, which are commonly used to represent gene models. These annotations are invaluable for understanding genomic regions and gene structures.
- **GenomicFeatures:** This package is the powerhouse for working with genomic data in R. It provides functions for creating and manipulating genomic feature objects.
- **rtracklayer:** This package specializes in reading and handling genomic data files, including BED and GTF files.

11.2.12 Accessing Genome Annotations

You can check AnnotationHub docs here: <https://bioconductor.org/packages/release/bioc/vignettes/AnnotationHub/inst/doc/AnnotationHub.html>

Genome annotations provide essential information about the location and structure of genes, which is crucial for understanding how genes function and how they are regulated. For example, knowing the coordinates of exons, introns, and promoters allows us to analyze where specific genetic elements are located in the genome.

```
library(AnnotationHub)

# Initialize the AnnotationHub
ah <- AnnotationHub()

# Query for specific annotations, for example, Homo sapiens (human) in the GRCh37 assembly
annotations <- AnnotationHub::query(ah, c("gtf", "Homo_sapiens", "GRCh37"))

annotations

## AnnotationHub with 7 records
## # snapshotDate(): 2021-10-20
## # $datatypeprovider: Ensembl
## # $species: Homo sapiens
## # $rdataclass: GRanges
## # additional mcols(): taxonomyid, genome, description,
## #   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
## #   rdatapath, sourceurl, sourcetype
## # retrieve records with, e.g., 'object[["AH7558"]]''
##
##           title
## AH7558 | Homo_sapiens.GRCh37.70.gtf
## AH7619 | Homo_sapiens.GRCh37.69.gtf
## AH7666 | Homo_sapiens.GRCh37.71.gtf
## AH7726 | Homo_sapiens.GRCh37.72.gtf
## AH7790 | Homo_sapiens.GRCh37.73.gtf
## AH8753 | Homo_sapiens.GRCh37.74.gtf
## AH10684 | Homo_sapiens.GRCh37.75.gtf

# Select one of the annotations (e.g., GRCh37.gtf)
GRCh37.gtf <- annotations[['AH8753']]
```

Now, we have a `GenomicRanges` object called `GRCh37.gtf`, which contains genomic features from the GRCh37 assembly of the human genome.

11.2.13 Understanding Genomic Biotypes

Genes in the genome can have different biotypes, indicating their functional roles. We can filter our genomic features based on biotypes, such as “protein_coding” and “lincRNA.”

Filtering genes by biotype helps us focus on specific classes of genes, such as protein-coding genes, which are involved in producing proteins, or long intergenic non-coding RNAs (lincRNAs), which play regulatory roles.

```
# what are the available biotypes
table(GRCh37.gtf$gene_biotype)

## # 3prime_overlapping_ncrna antisense IG_C_gene
## 63 28001 228
## IG_C_pseudogene IG_D_gene IG_J_gene
## 27 128 52
## IG_J_pseudogene IG_V_gene IG_V_pseudogene
## 6 747 393
## lincRNA miRNA misc_RNA
## 34236 3361 2174
## Mt_rRNA Mt_tRNA polymorphic_pseudogene
## 2 22 3475
## processed_pseudogene processed_transcript protein_coding
## 1 12720 2106659
## pseudogene rRNA sense_intronic
## 44993 568 1662
## sense_overlapping snoRNA snRNA
## 841 1549 2067
## TR_C_gene TR_D_gene TR_J_gene
## 44 6 164
## TR_J_pseudogene TR_V_gene TR_V_pseudogene
## 4 597 67

# subset
GRCh37.gtf <- GRCh37.gtf[GRCh37.gtf$gene_biotype %in% c("protein_coding", "lincRNA")]
```

11.2.14 Creating a Transcript Database (TxDb)

You can check GenomicFeatures docs here: <https://bioconductor.org/packages/release/bioc/vignettes/GenomicFeatures/inst/doc/GenomicFeatures.html>

To perform more advanced analyses, we'll create a transcript database (**TxDb**) from our genomic features. A **TxDb** is a structured database of transcript information, allowing us to efficiently query and retrieve specific genomic elements for analysis.

```
library(GenomicFeatures)

# Create a TxDb from the filtered genomic features
GRCh37.txdb <- makeTxDbFromGRanges(GRCh37.gtf)

GRCh37.txdb

## TxDb object:
## # Db type: TxDb
## # Supporting package: GenomicFeatures
## # Genome: GRCh37
## # Nb of transcripts: 171683
## # Db created by: GenomicFeatures package from Bioconductor
## # Creation time: 2024-07-28 13:30:22 -0400 (Sun, 28 Jul 2024)
## # GenomicFeatures version at creation time: 1.46.5
## # RSQLite version at creation time: 2.3.1
## # DBSCHEMAVERSION: 1.2
```

11.2.15 Extracting Exons, Introns, and Intergenic Regions

Now that we have our **TxDb**, we can extract various genomic elements for further analysis.

11.2.15.1 Exons by Gene

Analyzing exons by gene is essential for understanding the coding regions of genes and their splicing patterns. Let's retrieve exons grouped by genes.

```
exonsByGene <- exonsBy(GRCh37.txdb, "gene")

# GRangesList object, a list of GRanges
exonsByGene

## GRangesList object of length 30150:
## $ENSG00000000003
## GRanges object with 17 ranges and 2 metadata columns:
##   seqnames      ranges strand | exon_id      exon_name
##   <Rle>        <IRanges> <Rle> | <integer>  <character>
```

11.2. WORKING WITH GENOMIC COORDINATES AND REGIONS IN R193

```
## [1] X 99883667-99884983 - | 591006 ENSE00001459322
## [2] X 99885756-99885863 - | 591007 ENSE00000868868
## [3] X 99887482-99887565 - | 591008 ENSE00000401072
## [4] X 99887538-99887565 - | 591009 ENSE00001849132
## [5] X 99888402-99888536 - | 591010 ENSE00003554016
## ...
## [13] X 99890555-99890743 - | 591018 ENSE00003662440
## [14] X 99891188-99891686 - | 591019 ENSE00001886883
## [15] X 99891605-99891803 - | 591020 ENSE00001855382
## [16] X 99891790-99892101 - | 591021 ENSE00001863395
## [17] X 99894942-99894988 - | 591022 ENSE00001828996
## -----
## seqinfo: 265 sequences (1 circular) from GRCh37 genome
##
## ...
## <30149 more elements>
```

11.2.16 Merging All Exons

Merging exons (exons can overlap with each other) helps simplify analysis, such as quantifying the overall exonic content of genes. To get a single range representing all exons, we can reduce them.

```
allExons <- exons(GRCh37.txdb) %>%
  GenomicRanges::reduce()

allExons

## GRanges object with 279054 ranges and 0 metadata columns:
##           seqnames      ranges strand
##             <Rle>      <IRanges>  <Rle>
## [1]       1 29554-30039    +
## [2]       1 30267-30667    +
## [3]       1 30976-31109    +
## [4]       1 69091-70008    +
## [5]       1 160446-160690   +
## ...
## [279050] HSCHR7_1_CTG6 141373867-141374020   -
## [279051] HSCHR7_1_CTG6 141385280-141385438   -
## [279052] HSCHR7_1_CTG6 141386361-141386460   -
## [279053] HSCHR7_1_CTG6 141401359-141401418   -
## [279054] HSCHR7_1_CTG6 141401689-141401956   -
## -----
## seqinfo: 265 sequences (1 circular) from GRCh37 genome
```

11.2.17 Introns

Identifying introns is crucial for studying gene splicing and understanding the non-coding regions within genes. To find intronic regions, we can use the `intronsByTranscript` function.

```
introns <- intronsByTranscript(GRCh37.txdb) %>%
  unlist() %>%
  GenomicRanges::reduce()

introns

## GRanges object with 185379 ranges and 0 metadata columns:
##           seqnames      ranges strand
##             <Rle>      <IRanges>  <Rle>
## [1]          1 30040-30563    +
## [2]          1 30668-30975    +
## [3]          1 160691-161313   +
## [4]          1 317782-334128   +
## [5]          1 334298-439466   +
## ...
## [185375] HSCHR7_1_CTG6 141365119-141366086   -
## [185376] HSCHR7_1_CTG6 141366187-141373866   -
## [185377] HSCHR7_1_CTG6 141374021-141385279   -
## [185378] HSCHR7_1_CTG6 141385439-141386360   -
## [185379] HSCHR7_1_CTG6 141386461-141401688   -
## -----
## seqinfo: 265 sequences (1 circular) from GRCh37 genome
```

11.2.18 Getting All Genes

Having a complete list of genes is essential for various genomics analyses, including differential gene expression studies. Obtaining all genes is straightforward.

```
allGenes <- genes(GRCh37.txdb)

allGenes

## GRanges object with 30150 ranges and 1 metadata column:
##           seqnames      ranges strand |
##             <Rle>      <IRanges>  <Rle> |
## ENSG000000000003          X 99883667-99894988   - |
## ENSG000000000005          X 99839799-99854882   + |
## ENSG00000000419          20 49551404-49575092   - |
```

11.2. WORKING WITH GENOMIC COORDINATES AND REGIONS IN R195

```
##   ENSG000000000457          1 169818772-169863408 - |
##   ENSG000000000460          1 169631245-169823221 + |
##   ...
##   ENSG00000273488          ...     ...     ...     ... .
##   ENSG00000273490 HSCHR19LRC_LRC_J_CTG1 54693789-54697585 + |
##   ENSG00000273491           HG1308_PATCH 130600118-130603315 + |
##   ENSG00000273492          21    27543189-27589700 + |
##   ENSG00000273493          3     58315692-58315845 + |
##   gene_id
##   <character>
##   ENSG000000000003 ENSG000000000003
##   ENSG000000000005 ENSG000000000005
##   ENSG000000000419 ENSG000000000419
##   ENSG000000000457 ENSG000000000457
##   ENSG000000000460 ENSG000000000460
##   ...
##   ENSG00000273488 ENSG00000273488
##   ENSG00000273490 ENSG00000273490
##   ENSG00000273491 ENSG00000273491
##   ENSG00000273492 ENSG00000273492
##   ENSG00000273493 ENSG00000273493
##   -----
##   seqinfo: 265 sequences (1 circular) from GRCh37 genome
```

11.2.19 Promoters

Promoter regions are critical for understanding gene regulation and identifying potential binding sites for transcription factors. To find promoter regions, typically defined as the region from -1kb to +500bp around the transcription start site (TSS), we can use the promoters function.

```
promoterRegions <- promoters(genes(GRCh37.txdb),
                             upstream = 1000,
                             downstream = 500)

promoterRegions

## GRanges object with 30150 ranges and 1 metadata column:
##           seqnames      ranges strand
##           <Rle>      <IRanges> <Rle> |
##   ENSG000000000003        X 99894489-99895988 - |
##   ENSG000000000005        X 99838799-99840298 + |
##   ENSG000000000419       20 49574593-49576092 - |
##   ENSG000000000457        1 169862909-169864408 - |
```

```

##   ENSG000000000460          1 169630245-169631744 + |
##   ...                      ...   ...   ...
##   ENSG00000273488          3 100079031-100080530 + |
##   ENSG00000273490 HSCHR19LRC_LRC_J_CTG1 54692789-54694288 + |
##   ENSG00000273491     HG1308_PATCH 130599118-130600617 + |
##   ENSG00000273492          21 27542189-27543688 + |
##   ENSG00000273493          3 58314692-58316191 + |
##           gene_id
##           <character>
##   ENSG000000000003 ENSG000000000003
##   ENSG000000000005 ENSG000000000005
##   ENSG00000000419 ENSG00000000419
##   ENSG00000000457 ENSG00000000457
##   ENSG00000000460 ENSG00000000460
##   ...
##   ENSG00000273488 ENSG00000273488
##   ENSG00000273490 ENSG00000273490
##   ENSG00000273491 ENSG00000273491
##   ENSG00000273492 ENSG00000273492
##   ENSG00000273493 ENSG00000273493
##   -----
##   seqinfo: 265 sequences (1 circular) from GRCh37 genome

```

11.2.20 Full Genome

Having the entire genome as a single object is useful for genome-wide analyses and visualizations. To represent the entire genome as a GRanges object:

```

chrom_granges <- as(seqinfo(GRCh37.txdb), "GRanges")
chrom_granges

```

```

## GRanges object with 265 ranges and 0 metadata columns:
##           seqnames      ranges strand
##           <Rle>    <IRanges>  <Rle>
##           1          1 1-249250621 *
##           2          2 1-243199373 *
##           3          3 1-198022430 *
##           4          4 1-191154276 *
##           5          5 1-180915260 *
##           ...
##   HSCHR7_1_CTG6 HSCHR7_1_CTG6 1-159144671 *
##   HSCHR9_1_CTG1 HSCHR9_1_CTG1 1-141228243 *
##   HSCHR9_1_CTG35 HSCHR9_1_CTG35 1-141221627 *
##   HSCHR9_2_CTG35 HSCHR9_2_CTG35 1-141219511 *

```

```
##   HSCHR9_3_CTG35 HSCHR9_3_CTG35 1-141224529      *
##   -----
##   seqinfo: 265 sequences (1 circular) from GRCh37 genome
```

11.2.21 Full Transcriptome

Merging overlapping transcripts simplifies transcript-level analyses and helps identify the full extent of genes. To represent the entire transcriptome, we can merge overlapping features.

```
collapsed_tx <- GenomicRanges::reduce(transcripts(GRCh37.txdb))

# Set strand information to '*'
strand(collapsed_tx) <- "*"
```

11.2.22 Intergenic Regions

Intergenic regions often contain important regulatory elements, and identifying them can provide insights into gene regulation. To find regions that are not within any annotated genes, we can use the setdiff function.

```
intergenicRegions <- GenomicRanges::setdiff(chrom_granges, collapsed_tx)

intergenicRegions

## GRanges object with 24100 ranges and 0 metadata columns:
##           seqnames      ranges strand
##                 <Rle>      <IRanges>  <Rle>
## [1]          1        1-29553    *
## [2]          1      31110-34553    *
## [3]          1      36082-69090    *
## [4]          1      70009-89294    *
## [5]          1     133567-134900    *
## ...
## [24096] HSCHR7_1_CTG6 141493731-159144671    *
## [24097] HSCHR9_1_CTG1      1-141228243    *
## [24098] HSCHR9_1_CTG35      1-141221627    *
## [24099] HSCHR9_2_CTG35      1-141219511    *
## [24100] HSCHR9_3_CTG35      1-141224529    *
## -----
## seqinfo: 265 sequences (1 circular) from GRCh37 genome
```

11.2.23 Exploring Untranslated Regions (UTRs)

UTRs play crucial roles in post-transcriptional regulation, and analyzing them can provide insights into gene regulation mechanisms. If you're interested in untranslated regions (UTRs) of genes, you can use functions like `fiveUTRsByTranscript` and `threeUTRsByTranscript` provided by the `GenomicFeatures` package.

```
# To get 5' UTRs by transcript
fiveUTRs <- fiveUTRsByTranscript(GRCh37.txdb)

# To get 3' UTRs by transcript
threeUTRs <- threeUTRsByTranscript(GRCh37.txdb)
```

11.2.24 Conclusion

In this lesson, we've explored various genomic features and their manipulation using R packages such as `GenomicFeatures`, `AnnotationHub`, and `rtracklayer`. These tools are invaluable for genomics research, allowing you to analyze and interpret genomic data effectively. Whether you're working with ChIP-seq, RNA-seq, or genome annotation, understanding genomic features is essential to uncover the secrets of the genome.

11.3 Exploring CpG Islands and Shores in Genomic Data

In this lesson, we will delve into the fascinating world of genomics and learn how to manipulate and analyze genomic data using the powerful tools available in the Bioconductor package. We will specifically focus on CpG islands and their shores, exploring how to extract and analyze these critical genomic features.

11.3.1 Introduction to CpG Islands

CpG islands are regions of DNA that contain a high frequency of cytosine-guanine (CpG) dinucleotide pairs. These regions are essential for regulating gene expression and have critical roles in various biological processes, including DNA methylation and epigenetic modifications. Analyzing CpG islands can provide valuable insights into gene regulation and genome function.

In this lesson, we will use Bioconductor to fetch CpG island coordinates from the UCSC Genome Browser, extract CpG shores, and perform various genomic operations.

11.3.2 Fetching CpG Island Coordinates

CpG islands are critical genomic regions involved in gene regulation and epigenetic modifications. Accessing their coordinates is the first step in understanding their distribution across the genome and their potential functional roles.

Researchers often use CpG island coordinates to investigate gene promoters, identify potential regulatory elements, and study the epigenetic regulation of specific genes in various diseases, including cancer.

To begin, we will retrieve the CpG island coordinates from the UCSC Genome Browser using the `AnnotationHub` package. CpG islands are available for various species, and in this example, we are using *Homo sapiens* (human) data.

```
# Fetching CpG island coordinates from UCSC Genome Browser
library(AnnotationHub)
ah <- AnnotationHub()

AnnotationHub::query(ah, c("cpg", "UCSC"))

## AnnotationHub with 59 records
## # snapshotDate(): 2021-10-20
## # $dataprovider: UCSC
## # $species: Homo sapiens, Bos taurus, Pan troglodytes, Felis catus, Rattus n...
## # $rdataclass: GRanges
## # additional mcols(): taxonomyid, genome, description,
## #   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
## #   rdatapath, sourceurl, sourcetype
## # retrieve records with, e.g., 'object[["AH5086"]]''
##
##           title
## AH5086 | CpG Islands
## AH5096 | Evo Cpg
## AH5204 | CpG Islands
## AH5227 | Evo Cpg
## AH5344 | CpG Islands
## ...
## AH7109 | CpG Islands
## AH7116 | CpG Islands
## AH7135 | CpG Islands
## AH7168 | CpG Islands
## AH7203 | CpG Islands
```

use the first entry

```
cgi <- ah[["AH5086"]]
cgi

## GRanges object with 28691 ranges and 1 metadata column:
##           seqnames      ranges strand |      name
##           <Rle>      <IRanges> <Rle> | <character>
## [1]     chr1    28736-29810   * | CpG:_116
## [2]     chr1  135125-135563   * | CpG:_30
## [3]     chr1 327791-328229   * | CpG:_29
## [4]     chr1 437152-438164   * | CpG:_84
## [5]     chr1 449274-450544   * | CpG:_99
## ...
## [28687] chr9_g1000201_random 15651-15909   * | CpG:_30
## [28688] chr9_g1000201_random 26397-26873   * | CpG:_43
## [28689] chr11_g1000202_random 16284-16540   * | CpG:_23
## [28690] chr17_g1000204_random 54686-57368   * | CpG:_228
## [28691] chr17_g1000205_random 117501-117801   * | CpG:_23
## -----
## seqinfo: 93 sequences (1 circular) from hg19 genome
```

We now have the CpG island coordinates stored in the `cgi` GenomicRanges object.

11.3.3 Defining CpG Shores

CpG shores are regions located near CpG islands, and they play a crucial role in gene regulation. Defining these shores allows us to explore the regulatory landscape around CpG islands and identify regions of potential interest.

By analyzing CpG shores, researchers can gain insights into how epigenetic modifications in these regions affect gene expression. This knowledge is vital for understanding diseases that involve aberrant gene regulation.

CpG shores are regions located 2000 base pairs upstream and 2000 base pairs downstream of CpG islands. We can use Bioconductor to extract these shores.

```
# Extract the shore defined by 2000 bp upstream of CpG islands
shore1 <- trim(flank(cgi, width = 2000, start = TRUE))

# Extract the shore defined by 2000 bp downstream of CpG islands
shore2 <- trim(flank(cgi, width = 2000, start = FALSE))
```

`trim` will trim off the bases that exceed the chromosome ends since we extend 2000 bp upstream and downstream of the CpG sites. Some CpG sites can be very close to the ends of the chromosomes.

11.3.4 Combining and Analyzing CpG Shores

Combining the upstream and downstream shores and analyzing their overlap with CpG islands helps identify regions with unique genomic characteristics. This step allows researchers to pinpoint areas of interest for further investigation.

Researchers often use this analysis to identify differentially methylated regions (DMRs) associated with specific diseases or conditions. DMRs can serve as biomarkers or potential therapeutic targets.

Now, let's perform some genomic operations on these CpG shores. We'll combine the upstream and downstream shores and identify the features that are present in shores but not in CpG islands (i.e., shores not overlapping with islands).

```
# Combine the shores where they overlap
shore1_2 <- GenomicRanges::reduce(c(shore1, shore2))

# Extract the features (ranges) that are present in shores only and not in CpG islands
cpgi_shores <- GenomicRanges::setdiff(shore1_2, cgi)
cpgi_shores$name <- paste("shore", 1:length(cpgi_shores), sep = "_")

cpgi_shores

## GRanges object with 51914 ranges and 1 metadata column:
##           seqnames      ranges strand |      name
##           <Rle>      <IRanges> <Rle> | <character>
## [1]     chr1    26736-28735   * |    shore_1
## [2]     chr1    29811-31810   * |    shore_2
## [3]     chr1  133125-135124   * |    shore_3
## [4]     chr1  135564-137563   * |    shore_4
## [5]     chr1  325791-327790   * |    shore_5
## ...
## [51910] chrUn_g1000241  37274-39273   * | shore_51910
## [51911] chrUn_g1000242  10843-12842   * | shore_51911
## [51912] chrUn_g1000242  13100-15099   * | shore_51912
## [51913] chrUn_g1000243  28420-30419   * | shore_51913
## [51914] chrUn_g1000243  30716-32715   * | shore_51914
## -----
## seqinfo: 93 sequences (1 circular) from hg19 genome
```

Now, `cpgi_shores` contains the `GenomicRanges` object representing CpG shores, and each shore is labeled with a unique name.

11.3.5 Conclusion

In this lesson, we've explored the powerful capabilities of the Bioconductor package for working with genomic data. We've fetched CpG island coordinates, extracted CpG shores, and performed genomic operations to identify regions of interest. These techniques are fundamental for researchers and bioinformaticians working with genomics data to unravel the mysteries of the genome.

If you're interested in diving deeper into genomics analysis, consider exploring the tutorials provided by Bioconductor on their website. They offer a wealth of knowledge and resources to help you harness the full potential of genomic data analysis.

11.4 Real-World Applications: ChIP-seq

Understanding genomic features is crucial for various genomics tasks, including:

- ChIP-seq Analysis: You can use these genomic ranges to determine how many ChIP-seq peaks fall into promoters, exons, introns, or intergenic regions, helping you interpret the functional significance of your data.
- RNA-seq Analysis: Identifying which exons are covered by RNA-seq reads and counting reads in each exon allows you to quantify gene expression accurately.
- Functional Genomics: Investigating the genomic context of genes helps in understanding their regulatory elements, including promoters and enhancers.
- Genome Annotation: These tools are essential for creating comprehensive annotations of genomes, enabling researchers to understand gene structures and functions.

11.4.1 read in peak file

11.4.2 Identify promoters that overlap with the peaks

11.4.3 use the ChIPseeker package

11.5 Analyzing and Visualizing Genomic Data

In this lesson, we will explore several essential tools and techniques used in genomics research, including `GEOquery` for data retrieval, gene ID conversion using

`biomaRt` and `org.Hs.eg.db`, and visualization with `ComplexHeatmap`. These tools are commonly used in genomics to analyze and visualize gene expression data. We will briefly mention `DESeq2` for differential expression analysis.

11.5.1 DESeq2

You can explore docs here: <https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>

`DESeq2` is a powerful Bioconductor package used for differential expression analysis. It is particularly helpful when working with RNA-seq data. This tool helps identify genes that are differentially expressed under different conditions. The tutorial on Bioconductor is very comprehensive and I will leave the students to read by themselves. We will use it in our final project.

11.5.2 GEOquery

You can explore docs here: <https://bioconductor.org/packages/release/bioc/vignettes/GEOquery/inst/doc/GEOquery.html>

`GEOquery` is a valuable R package for downloading and importing gene expression data directly from public repositories such as the Gene Expression Omnibus (GEO).

```
# Loading the GEOquery library
library(GEOquery)

# Downloading and importing data from GEO
GSE197576 <- getGEO(GEO = "GSE197576", GSEMatrix = TRUE, destdir = "~/Downloads")

GSE197576

## $GSE197576_series_matrix.txt.gz
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 0 features, 12 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM5920759 GSM5920760 ... GSM5920770 (12 total)
##   varLabels: title geo_accession ... tissue:ch1 (43 total)
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
```

```

##   pubMedIds: 35487218
## Annotation: GPL18573

# Accessing expression data
exprs(GSE197576$GSE197576_series_matrix.txt.gz)

##      GSM5920759 GSM5920760 GSM5920761 GSM5920762 GSM5920763 GSM5920764
##      GSM5920765 GSM5920766 GSM5920767 GSM5920768 GSM5920769 GSM5920770

```

In this example, it returns an empty `ExpressionSet` object. We used `exprs` to return the matrix but it has no features. You may try a different GEO accession.

`GEOquery` simplifies the process of retrieving gene expression data from public repositories like GEO. Researchers use it to access valuable datasets for their studies.

11.5.3 Converting Gene IDs

In genomics research, integrating data from various sources often involves working with different gene identifier systems. Converting gene IDs is a crucial step to harmonize and standardize data for downstream analysis. Here, we discuss two commonly used methods, `biomaRt` and `org.Hs.eg.db`, and explain why these tasks are essential in a researcher's environment.

11.5.3.1 Using `biomaRt`

You can explore docs here: https://bioconductor.org/packages/release/bioc/vignettes/biomaRt/inst/doc/accessing_ensembl.html

Converting gene IDs using `biomaRt` is essential for mapping Ensembl gene IDs to more recognizable gene symbols and associated information.

```

library(biomaRt)
ensembl <- useMart("ensembl", dataset = "hsapiens_gene_ensembl")

gene_info <- getBM(attributes = c("ensembl_gene_id", "gene_biotype",
                                    "chromosome_name", "hgnc_symbol"),
                    filters = "ensembl_gene_id",
                    values = c("ENSG00000164307"),
                    mart = ensembl)
print(gene_info)

##   ensembl_gene_id   gene_biotype chromosome_name hgnc_symbol
## 1 ENSG00000164307 protein_coding                 5        ERAP1

```

11.5.3.2 Using org.Hs.eg.db

You can explore docs here: <https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html>

Mapping ENTREZIDs to official gene symbols using `org.Hs.eg.db` is crucial for researchers working with gene expression data.

Why Researchers Do This:

- Consistency in Annotation: ENTREZIDs are a widely accepted and consistent gene identifier system. Mapping other identifiers to ENTREZIDs ensures that gene information is uniform and can be compared across different studies.
- Integration with Other Databases: Many databases and tools use ENTREZIDs (e.g., the KEGG pathway database) as a standard for gene annotation. Converting identifiers to ENTREZIDs facilitates seamless integration with these resources.
- Gene Symbol Mapping: Once converted to ENTREZIDs, researchers can efficiently map these to official gene symbols, providing meaningful and interpretable gene names for further analysis and reporting.

```
# Loading the org.Hs.eg.db library
library(org.Hs.eg.db)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)

# Accessing gene information
hg19_genes <- genes(TxDb.Hsapiens.UCSC.hg19.knownGene)

# Mapping ENTREZIDs to official gene symbols
map <- AnnotationDbi::select(org.Hs.eg.db, keys = hg19_genes$gene_id,
                             columns = "SYMBOL", keytype = "ENTREZID")

head(map, n = 10)

##      ENTREZID      SYMBOL
## 1          1       A1BG
## 2          10      NAT2
## 3         100      ADA
## 4        1000     CDH2
## 5       10000     AKT3
## 6 100008586    GAGE12F
## 7 100009676 ZBTB11-AS1
## 8       10001     MED6
```

```
## 9      10002      NR2E3
## 10     10003      NAALAD2
```

In genomics research, this type of mapping is essential when you have data that uses ENTREZIDs, and you want to work with more interpretable gene symbols for analysis or visualization. It allows you to associate gene identifiers with their official names, making the data more understandable and facilitating downstream analyses and interpretation.

11.5.4 ComplexHeatmap

You can explore docs here: <https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html>

Visualizing gene expression data is a crucial step in genomics research because it helps researchers gain insights into how genes are expressed under different conditions or in various samples. **ComplexHeatmap** is a versatile R package that serves as an artistic palette for creating intricate and informative heatmaps. Here's why visualizing gene expression data using **ComplexHeatmap** is essential:

1. Identify Expression Patterns: Gene expression data often involve a large number of genes and samples. Heatmaps allow you to quickly identify patterns in gene expression, such as clusters of genes with similar expression profiles. This can reveal groups of genes that are co-regulated or have similar functions.
2. Visualize Differential Expression: When comparing gene expression between different conditions or treatments (e.g., healthy vs. diseased tissue), heatmaps can highlight genes that are significantly upregulated or downregulated. This visualization aids in pinpointing genes of interest for further investigation.
3. Sample Relationships: Heatmaps also help in understanding the relationships between samples. For example, you can identify outliers, detect batch effects, or confirm the consistency of replicates by examining how samples cluster based on their expression profiles.
4. Publication-Ready Figures: **ComplexHeatmap** generates high-quality heatmap images that are suitable for inclusion in research papers and presentations. It provides the ability to export heatmaps in various formats (e.g., PDF, PNG) for easy sharing and publication.
5. Interactive Exploration: In addition to static heatmaps, **InteractiveComplexHeatmap** from the same author supports interactive exploration. You can zoom in on specific sections of the heatmap, hover over cells to view gene names or expression values, and provide interactive tools for your audience to explore the data themselves.

Let's go through an example by simulating a count matrix from a gene expression experiment:

```
library(ComplexHeatmap)

set.seed(123)

# Parameters for the negative binomial distribution
size_parameter <- 3 # Size parameter (dispersion)
mean_parameter <- 10 # Mean parameter

# Number of genes (rows) and samples (columns)
num_genes <- 10
num_samples <- 20

# Simulate a count table with negative binomial distribution
expr_data <- matrix(rnbinom(num_genes * num_samples,
                           size = size_parameter,
                           mu = mean_parameter),
                     nrow = num_genes,
                     ncol = num_samples)
```

check the quantiles of the data. This is important because we need to know the ranges of the data so we can map the values to color.

```
quantile(expr_data, c(0, 0.2, 0.5, 0.8, 1))
```

```
##    0%   20%   50%   80% 100%
##    0     4     9    14    31
```

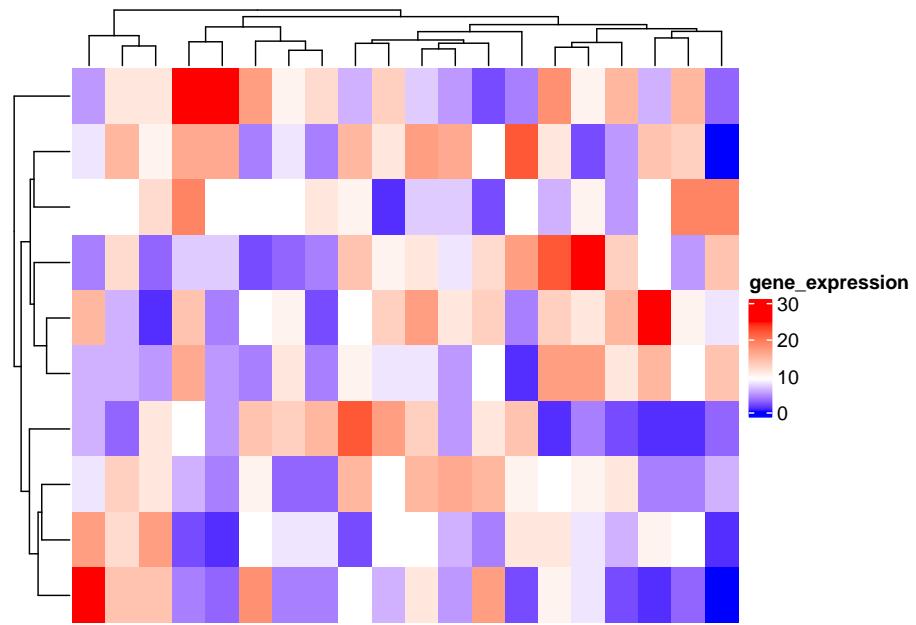
Map the color to the values

```
color_map<- circlize::colorRamp2(c(0, 9, 25), c("blue", "white", "red"))
```

Any value beyond 25 will be mapped to the same intensity of redness.

Make the heatmap:

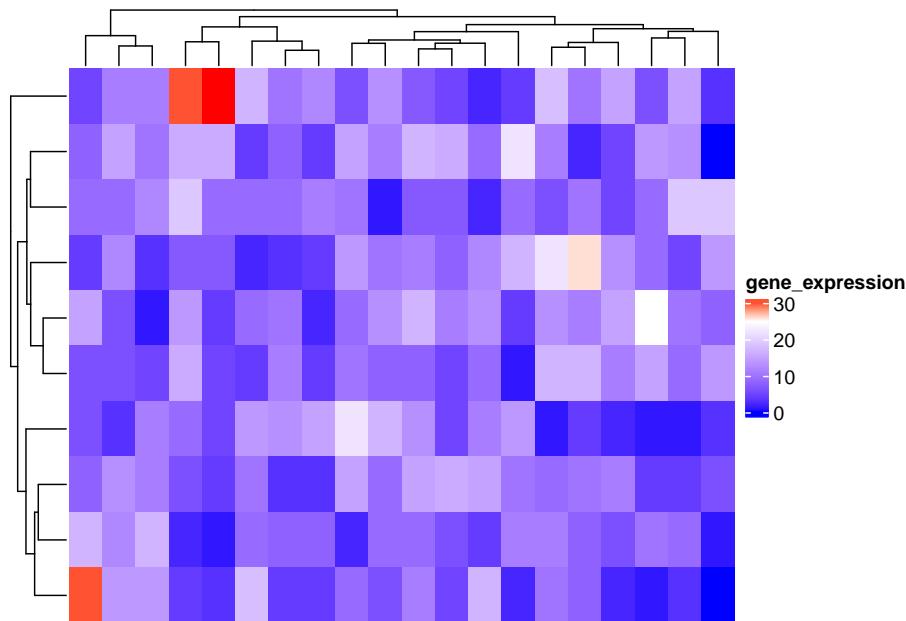
```
Heatmap(expr_data, col=color_map, name = "gene_expression")
```



By default, it will cluster both the rows and columns.

Let's see how it looks if we change the color mapping

```
color_map2<- circlize::colorRamp2(c(0, 25, 31), c("blue", "white", "red"))
Heatmap(expr_data, col=color_map2, name = "gene_expression")
```



Now, we see very fewer red cells, but the underlying data is the same! How you map the values to color makes a big difference on how the heatmap look. Note the legend reflects our changes in the color mapping.

We will use it again in the next section.

11.5.5 Conclusion

This lesson has provided an overview of essential tools and techniques used in genomics research. We have explored the following key topics:

- **DESeq2:** We learned about DESeq2 for differential expression analysis, which is crucial for identifying genes that are differentially expressed under different conditions, such as in disease versus healthy states.
- **GEOquery:** We discussed how to retrieve gene expression data from public repositories like GEO using the GEOquery package. Accessing publicly available datasets is a valuable resource for genomics research.
- **Gene ID Conversion:** We covered two methods, `biomaRt` and `org.Hs.eg.db`, for converting gene IDs. This step is essential for integrating data from various sources and ensuring consistent gene identification.
- **ComplexHeatmap:** We explored the `ComplexHeatmap` package, a powerful tool for visualizing gene expression data through heatmaps. Visualization aids in identifying patterns and trends in large genomic datasets.

These tools and techniques are indispensable for genomics researchers, enabling them to analyze, integrate, and visualize genomic data effectively. By mastering these skills, researchers can gain valuable insights into gene expression patterns, biological processes, and potential biomarkers for various conditions and diseases.

11.6 Real-World Example - TCGA Analysis

In this lesson, we will explore a real-world example of analyzing cancer genomics data from The Cancer Genome Atlas (TCGA) project. TCGA is one of the largest and most renowned cancer sequencing projects, providing access to a wealth of genomic data from various cancer types. We will use R and several bioinformatics packages to download raw RNA-seq counts for 33 different cancer types, convert them to TPM (transcripts per million), and visualize the data in a heatmap.

11.6.1 Introduction to TCGA

The Cancer Genome Atlas (TCGA) project is a groundbreaking initiative that has sequenced approximately 10,000 treatment-naive tumors across 33 different cancer types. It has generated a diverse range of data types, including whole-exome sequencing, whole-genome sequencing, copy-number variation analysis (SNP arrays), bulk RNA-seq, protein expression data (Reverse-Phase Protein Array), and DNA methylation profiles. TCGA has significantly contributed to our understanding of cancer biology and has opened up new avenues for cancer research.

11.6.2 Why Analyze TCGA Data?

Analyzing TCGA data can provide valuable insights into the molecular basis of cancer. Researchers can use this data to identify potential biomarkers, discover novel therapeutic targets, and gain a deeper understanding of the genetic alterations associated with specific cancer types. Moreover, TCGA data is freely accessible, making it a valuable resource for the scientific community.

11.6.3 Getting Started

We will use the `recount3` package to access TCGA data. `recount3` is an online resource that provides RNA-seq gene, exon, and exon-exon junction counts, along with coverage bigWig files for thousands of studies in both human and mouse. It represents the third generation of the ReCount project.

11.6.4 Step 1: Install and Load Required Packages

```
# Install the recount3 package if not already installed
# BiocManager::install("recount3")

# Load necessary libraries
library(recount3)
library(purrr)
library(dplyr)
library(ggplot2)
```

- **recount3:** This package is crucial for accessing TCGA data. It allows us to retrieve RNA-seq gene counts and other genomic information from the TCGA project.
- **purrr:** The purrr package is used for functional programming in R. We use it to apply functions to elements of a list, which is particularly useful for handling multiple datasets or projects.
- **dplyr:** dplyr is a powerful package for data manipulation. It helps us filter and process data efficiently, making it easier to work with large datasets like TCGA.
- **ggplot2:** For data visualization, we use the ggplot2 package. It allows us to create high-quality graphs and plots, which can be essential for presenting our findings.

By loading these packages, we ensure that we have access to the tools needed to analyze and visualize TCGA data effectively.

11.6.5 Step 2: Retrieve TCGA Project Information

Let's fetch information about available TCGA projects. TCGA encompasses a wide range of cancer types and studies. We filter the projects to focus only on those that originate from TCGA data sources.

```
# Get information about available TCGA projects
human_projects <- available_projects()

# Filter projects that are from TCGA data sources
tcga_info <- subset(
  human_projects,
  file_source == "tcga" & project_type == "data_sources"
)
```

```
head(tcga_info)

##      project organism file_source      project_home project_type n_samples
## 8710     ACC    human   tcga data_sources/tcga data_sources      79
## 8711    BLCA   human   tcga data_sources/tcga data_sources     433
## 8712    BRCA   human   tcga data_sources/tcga data_sources    1256
## 8713    CESC   human   tcga data_sources/tcga data_sources     309
## 8714    CHOL   human   tcga data_sources/tcga data_sources      45
## 8715    COAD   human   tcga data_sources/tcga data_sources    546
```

This step is essential to identify the relevant projects and data sources within TCGA. By narrowing our focus to TCGA data sources, we ensure that we are working with the specific datasets we need for our analysis.

11.6.6 Step 3: Create a RangedSummarizedExperiment Object

Now that we have identified the TCGA data sources of interest, we proceed to create a data structure called a `RangedSummarizedExperiment` object. This object allows us to efficiently organize and work with genomic data, including RNA-seq counts.

```
# Create a list of RangedSummarizedExperiment objects
proj_info <- purrr::map(seq(nrow(tcga_info)), ~tcga_info[., ])
rse_tcga <- purrr::map(proj_info, ~create_rse(.x))
```

Here, we create a list of `RangedSummarizedExperiment` objects, one for each project within TCGA. These objects will serve as containers for the gene expression data, making it easier to manipulate and analyze the information.

The first time you use `recount3`, it will ask:

```
/Users/tommytang/Library/Caches/org.R-project.R/R/recount3
does not exist, create directory? (yes/no): yes
```

`rse_tcga` is a list of `RangedSummarizedExperiment` objects, let's take a look at one of them. `RangedSummarizedExperiment` is a child of the `SummarizedExperiment`.

```
rse_tcga[[1]]
```

```

## class: RangedSummarizedExperiment
## dim: 63856 79
## metadata(8): time_created recount3_version ... annotation recount3_url
## assays(1): raw_counts
## rownames(63856): ENSG00000278704.1 ENSG00000277400.1 ...
##   ENSG00000182484.15_PAR_Y ENSG00000227159.8_PAR_Y
## rowData names(10): source type ... havana_gene tag
## colnames(79): 43e715bf-28d9-4b5e-b762-8cd1b69a430e
##   1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872 ...
##   a08b85ea-d1e7-4b77-8dec-36294305b9f7
##   aa2d53e5-d389-4332-9dd5-a736052e48f8
## colData names(937): rail_id external_id ... recount_seq_qc.errq
##   BigWigURL

```

11.6.7 Step 4: Explore TCGA Data

With our `RangedSummarizedExperiment` objects in place, we can now explore the TCGA data in more detail. Let's look at three aspects: gene counts, gene information, and metadata.

11.6.7.1 Gene Counts

We start by examining the raw RNA-seq counts, which represent how many times each gene was sequenced in each sample.

```

# Access raw gene counts from one RangedSummarizedExperiment object
raw_counts <- rse_tcgas[[1]]@assays@data$raw_counts[1:5, 1:5]

raw_counts

##                                     43e715bf-28d9-4b5e-b762-8cd1b69a430e
## ENSG00000278704.1                               0
## ENSG00000277400.1                               0
## ENSG00000274847.1                               0
## ENSG00000277428.1                               0
## ENSG00000276256.1                               0
##                                     1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872
## ENSG00000278704.1                               0
## ENSG00000277400.1                               0
## ENSG00000274847.1                               0
## ENSG00000277428.1                               0
## ENSG00000276256.1                               0
##                                     93b382e4-9c9a-43f5-bd3b-502cc260b886
## ENSG00000278704.1                               0

```

```

## ENSG00000277400.1          0
## ENSG00000274847.1          0
## ENSG00000277428.1          0
## ENSG00000276256.1          0
##           1f39dadd-3655-474e-ba4c-a5bd32c97a8b
## ENSG00000278704.1          0
## ENSG00000277400.1          0
## ENSG00000274847.1          0
## ENSG00000277428.1          0
## ENSG00000276256.1          0
##           8c8c09b9-ec83-45ec-bc4c-0ba92de60acb
## ENSG00000278704.1          0
## ENSG00000277400.1          0
## ENSG00000274847.1          0
## ENSG00000277428.1          0
## ENSG00000276256.1          0

```

This line extracts a subset of raw gene counts from the first dataset in the `RangedSummarizedExperiment` object `rse_tcga`. It takes the first 5 genes and the first 5 samples, providing a small portion of the data for analysis or exploration.

The `raw_counts` matrix provides a glimpse into the data, allowing us to see the sequencing counts for a subset of genes and samples.

11.6.7.2 Gene Information

Next, we retrieve information about the genes included in the dataset. This information includes details such as genomic location, gene type, and gene names.

```

# Access gene information
gene_info <- rse_tcga[[1]]@rowRanges

gene_info

```

	seqnames	ranges	strand	source
##	<Rle>	<IRanges>	<Rle>	<factor>
##	ENSG00000278704.1	GL000009.2	56140-58376	- ENSEMBL
##	ENSG00000277400.1	GL000194.1	53590-115018	- ENSEMBL
##	ENSG00000274847.1	GL000194.1	53594-115055	- ENSEMBL
##	ENSG00000277428.1	GL000195.1	37434-37534	- ENSEMBL
##	ENSG00000276256.1	GL000195.1	42939-49164	- ENSEMBL
##
##	ENSG00000124334.17_PAR_Y	chrY	57184101-57197337	+ HAVANA

```

##   ENSG00000185203.12_PAR_Y      chrY 57201143-57203357    - | HAVANA
##   ENSG00000270726.6_PAR_Y      chrY 57190738-57208756    + | HAVANA
##   ENSG00000182484.15_PAR_Y      chrY 57207346-57212230    + | HAVANA
##   ENSG00000227159.8_PAR_Y      chrY 57212184-57214397    - | HAVANA
##           type bp_length     phase          gene_id
##           <factor> <numeric> <integer>          <character>
##   ENSG00000278704.1    gene     2237     <NA> ENSG00000278704.1
##   ENSG00000277400.1    gene     2179     <NA> ENSG00000277400.1
##   ENSG00000274847.1    gene     1599     <NA> ENSG00000274847.1
##   ENSG00000277428.1    gene      101     <NA> ENSG00000277428.1
##   ENSG00000276256.1    gene     2195     <NA> ENSG00000276256.1
##           ...
##           ...     ...     ...
##   ENSG00000124334.17_PAR_Y    gene     2504     <NA> ENSG00000124334.17_PA..
##   ENSG00000185203.12_PAR_Y    gene     1054     <NA> ENSG00000185203.12_PA..
##   ENSG00000270726.6_PAR_Y    gene      773     <NA> ENSG00000270726.6_PA..
##   ENSG00000182484.15_PAR_Y    gene     4618     <NA> ENSG00000182484.15_PA..
##   ENSG00000227159.8_PAR_Y    gene     1306     <NA> ENSG00000227159.8_PA..
##           gene_type   gene_name    level
##           <character> <character> <character>
##   ENSG00000278704.1    protein_coding BX004987.1      3
##   ENSG00000277400.1    protein_coding AC145212.2      3
##   ENSG00000274847.1    protein_coding AC145212.1      3
##   ENSG00000277428.1    misc_RNA        Y_RNA      3
##   ENSG00000276256.1    protein_coding AC011043.1      3
##           ...
##           ...     ...     ...
##   ENSG00000124334.17_PAR_Y    protein_coding      IL9R      2
##   ENSG00000185203.12_PAR_Y    antisense        WASIR1      2
##   ENSG00000270726.6_PAR_Y    processed_transcript AJ271736.10      2
##   ENSG00000182484.15_PAR_Y    transcribed_unproces.. WASH6P      2
##   ENSG00000227159.8_PAR_Y    unprocessed_pseudogene DDX11L16      2
##           havana_gene      tag
##           <character> <character>
##   ENSG00000278704.1      <NA>      <NA>
##   ENSG00000277400.1      <NA>      <NA>
##   ENSG00000274847.1      <NA>      <NA>
##   ENSG00000277428.1      <NA>      <NA>
##   ENSG00000276256.1      <NA>      <NA>
##           ...
##           ...     ...     ...
##   ENSG00000124334.17_PAR_Y OTTHUMG00000022720.1      PAR
##   ENSG00000185203.12_PAR_Y OTTHUMG00000022676.3      PAR
##   ENSG00000270726.6_PAR_Y OTTHUMG00000184987.2      PAR
##   ENSG00000182484.15_PAR_Y OTTHUMG00000022677.5      PAR
##   ENSG00000227159.8_PAR_Y OTTHUMG00000022678.1      PAR
##   -----
##   seqinfo: 374 sequences from an unspecified genome; no seqlengths

```

The `gene_info` object provides essential context about the genes being studied, enabling us to interpret the gene expression data more effectively.

11.6.7.3 Metadata Information

Lastly, we access metadata information associated with the samples. Metadata contains additional details about each sample, such as unique identifiers, external IDs, and the TCGA study to which each sample belongs.

```
# Access metadata information
metadata_info<- rse_tcga[[1]]@colData@listData %>% as.data.frame() %>% `[,`(1:5, 1:5)

metadata_info

##   rail_id           external_id study
## 1 106797 43e715bf-28d9-4b5e-b762-8cd1b69a430e ACC
## 2 110230 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872 ACC
## 3 110773 93b382e4-9c9a-43f5-bd3b-502cc260b886 ACC
## 4 110869 1f39dadd-3655-474e-ba4c-a5bd32c97a8b ACC
## 5 116503 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb ACC
##          tcga_tcga_barcode      tcga_gdc_file_id
## 1 TCGA-OR-A5KU-01A-11R-A29S-07 43e715bf-28d9-4b5e-b762-8cd1b69a430e
## 2 TCGA-P6-A50G-01A-22R-A29S-07 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872
## 3 TCGA-OR-A5K5-01A-11R-A29S-07 93b382e4-9c9a-43f5-bd3b-502cc260b886
## 4 TCGA-OR-A5K4-01A-11R-A29S-07 1f39dadd-3655-474e-ba4c-a5bd32c97a8b
## 5 TCGA-OR-A5LP-01A-11R-A29S-07 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb
```

NOTE: `[` is a function itself for subsetting.

This line creates a dataframe (`metadata_info`) containing metadata information from the first dataset in the `RangedSummarizedExperiment` object `rse_tcga`. It selects the first 5 rows and first 5 columns of this metadata for examination or analysis. Metadata often includes details about the samples, helping researchers understand their characteristics and context within the dataset.

This metadata is crucial for tracking and organizing the samples and understanding their context within the TCGA project.

11.6.8 Step 5: Data Transformation - Converting Raw Counts to TPM

In this step, we will convert the raw RNA-seq counts into TPM (Transcripts Per Million) values. This transformation normalizes the data, making it comparable

across samples and genes. TPM accounts for both the length of genes and the total number of reads in each sample.

First, we define a function called `count2tpm` that takes a `RangedSummarizedExperiment` object (`rse`) as input and performs the TPM conversion. Here's what each part of the function does:

1. We extract the raw gene count matrix from the `rse` object.
2. We retrieve the effective gene length for each gene, which is needed for TPM calculation. In this example, we use the gene length information provided in the TCGA data.
3. We calculate the Reads Per Kilobase (RPK) for each gene by dividing the raw counts by the gene length.
4. We calculate the sum of RPK values for each sample (column) and divide it by 1,000,000 (`1e6`) to scale to a per-million basis.
5. Finally, we calculate TPM values by dividing the RPK values for each gene by the per-million scale factor.

```
genes_of_interest <- c("MSLN", "EGFR", "ERBB2", "CEACAM5", "NECTIN4", "EPCAM",
                      "MUC16", "MUC1", "CD276", "FOLH1", "DLL3", "VTCN1",
                      "PROM1", "PVR", "CLDN6", "MET", "FOLR1", "TNFRSF10B",
                      "TACSTD2", "CD24")

count2tpm <- function(rse) {
  count_matrix <- rse@assays@data$raw_counts
  gene_length <- rse@rowRanges$bp_length
  reads_per_rpk <- count_matrix / gene_length
  per_mil_scale <- colSums(reads_per_rpk) / 1e6
  tpm_matrix <- t(t(reads_per_rpk) / per_mil_scale)

  # Make sure they match the ENSG and gene order
  gene_ind <- rse@rowRanges$gene_name %in% genes_of_interest
  tpm_submatrix <- tpm_matrix[gene_ind,]
  rownames(tpm_submatrix) <- rse@rowRanges[gene_ind, ]$gene_name

  return(tpm_submatrix)
}
```

After defining the `count2tpm` function, we apply it to each of our `RangedSummarizedExperiment` objects stored in `rse_tcga`. This step converts the raw counts to TPM values and subsets the data to include only the genes of interest (specified in the `genes_of_interest` vector). The resulting `tpm_data` is a list of TPM matrices for each TCGA project.

11.6.9 Step 6: Combine Data and Metadata

In this step, we will combine the TPM data matrices from different TCGA projects and merge the associated metadata. This process is essential for creating a comprehensive dataset for further analysis and visualization.

11.6.9.1 Combining TPM Data Matrices

We have already obtained TPM values for each TCGA project and stored them in the `tpm_data` list, where each element represents a TPM matrix for one project. Now, we will combine these matrices into a single matrix, `tpm_data2`, which will contain TPM values for all samples across all projects.

```
# Convert raw count matrix per cancer type to TPM and subset to only the genes of interest
tpm_data <- map(rse_tcga, count2tpm)

# Combine the TPM data matrices into one matrix
tpm_data2 <- do.call(cbind, tpm_data)
```

In the code above, we use the `map()` function to apply the `count2tpm` function to each `RangedSummarizedExperiment` object in `rse_tcga`. The `do.call()` function then combines the resulting TPM matrices horizontally (column-wise), creating the `tpm_data2` matrix.

11.6.9.2 Combining Metadata

Next, we need to combine the metadata associated with each TCGA project into a single metadata table. This metadata contains information about the samples, such as sample type, study, and unique identifiers.

```
# Get the metadata columns for each project and convert them to data frames
metadata <- map(rse_tcga, ~x@colData@listData %>% as.data.frame())

# Combine the metadata data frames into one data frame
metadata2 <- do.call(rbind, metadata)
```

In this code, we again use the `map()` function to extract metadata columns from each `RangedSummarizedExperiment` object. We convert these columns into data frames and then use `bind_rows()` to vertically stack them into the `metadata2` data frame. `do.call` is a function from base R. `do.call(cbind)` is similar to `dplyr::bind_rows`, and `do.call(rbind)` is similar to `dplyr::bind_cols`.

11.6.9.3 Checking the Dimensions

Finally, let's check the dimensions of our combined data matrix and metadata data frame to ensure that the combination was successful.

```
# Check the dimensions of the combined TPM data matrix
dim(tpm_data2)

## [1] 20 11348

# Check the dimensions of the combined metadata data frame
dim(metadata2)

## [1] 11348 937
```

Running the `dim()` function on `tpm_data2` will give you the dimensions of the TPM data matrix, which should indicate the number of genes and samples in the combined dataset. Similarly, checking the dimensions of `metadata2` will confirm the number of metadata columns and rows, which correspond to the samples and their associated information.

With this combined dataset, you can proceed with various analyses, visualizations, and explorations of gene expression patterns and relationships with clinical metadata across different TCGA projects.

11.6.10 Step 7: Renaming Metadata Columns

Now, let's proceed with renaming some of the metadata columns for clarity and convenience. We will select specific columns from the metadata and create a new column called `sample_type` based on the `tcga.cgc_sample_sample_type` column's values. This new column will categorize samples into “cancer,” “metastatic,” or “normal” based on their sample type.

```
metadata2 <- metadata2 %>%
  dplyr::select(tcga.tcga_barcode, tcga.cgc_sample_sample_type, study) %>%
  mutate(sample_type = case_when(
    tcga.cgc_sample_sample_type == "Additional - New Primary" ~ "cancer",
    tcga.cgc_sample_sample_type == "Additional Metastatic" ~ "metastatic",
    tcga.cgc_sample_sample_type == "Metastatic" ~ "metastatic",
    tcga.cgc_sample_sample_type == "Primary Blood Derived Cancer - Peripheral Blood" ~ "cancer",
    tcga.cgc_sample_sample_type == "Primary Tumor" ~ "cancer",
    tcga.cgc_sample_sample_type == "Recurrent Tumor" ~ "cancer",
    tcga.cgc_sample_sample_type == "Solid Tissue Normal" ~ "normal"
  ))
```

In the code above, we use the `select()` function to choose specific columns from the metadata that we want to retain. We then create a new `sample_type` column using `mutate()` and `case_when()` based on the values of `tcga.cgc_sample_sample_type`.

11.6.11 Step 8: Merging into a single DataFrame

With the TPM data matrix and the updated metadata, we can combine them into a single dataframe named `final_df`. This dataframe will have TPM values for the selected genes along with associated metadata columns.

```
# Combine the TPM data matrix and metadata into a single dataframe
final_df <- cbind(t(tpm_data2), metadata2)
```

In this code, we use `cbind()` to merge the transposed TPM data matrix (`t(tpm_data2)`) with the metadata (`metadata2`) to create the `final_df` dataframe.

11.6.11.1 Displaying the Head of the Combined Dataframe

Finally, let's check the first few rows of the combined dataframe to ensure that the merging process was successful:

```
# Display the first few rows of the combined dataframe
head(final_df)
```

	TACSTD2	VTCN1	MUC1
## 43e715bf-28d9-4b5e-b762-8cd1b69a430e	0.7035937	0.00000000	0.67502205
## 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872	25.4360736	0.00000000	2.01525394
## 93b382e4-9c9a-43f5-bd3b-502cc260b886	1.5756197	0.00000000	0.90784666
## 1f39dadd-3655-474e-ba4c-a5bd32c97a8b	0.2702156	0.09099681	0.04293345
## 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb	0.4122814	0.00000000	0.11484380
## 85a86b91-4f24-4e77-ae2d-520f8e205efc	4.5469193	4.85973690	0.04208195
##	NECTIN4	FOLH1	FOLR1
## 43e715bf-28d9-4b5e-b762-8cd1b69a430e	0.08620727	7.213342	0.00000000
## 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872	0.07279804	23.552286	0.12154673
## 93b382e4-9c9a-43f5-bd3b-502cc260b886	0.69905270	2.853812	1.01000271
## 1f39dadd-3655-474e-ba4c-a5bd32c97a8b	0.01652257	1.157070	0.27942068
## 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb	0.03168398	2.408137	0.04922458
## 85a86b91-4f24-4e77-ae2d-520f8e205efc	0.06828305	1.010411	0.02248965
##	MSLN	CLDN6	ERBB2
## 43e715bf-28d9-4b5e-b762-8cd1b69a430e	0.06674445	0.09704962	1.879518
## 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872	0.95554610	0.25458796	7.777976

```

## 93b382e4-9c9a-43f5-bd3b-502cc260b886 0.04563568 0.25701910 2.905926
## 1f39dadd-3655-474e-ba4c-a5bd32c97a8b 0.03154912 0.24746913 4.914280
## 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb 0.26968788 0.12576720 1.494744
## 85a86b91-4f24-4e77-ae2d-520f8e205efc 0.01336404 0.01823883 13.474689
##                                     MUC16      DLL3 CEACAM5      PVR
## 43e715bf-28d9-4b5e-b762-8cd1b69a430e 0.0011479879 0.49589978      0 52.08113
## 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872 0.0008049670 2.52244014      0 40.87926
## 93b382e4-9c9a-43f5-bd3b-502cc260b886 0.0026190288 0.77074712      0 33.26727
## 1f39dadd-3655-474e-ba4c-a5bd32c97a8b 0.0051705741 0.10636402      0 28.26457
## 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb 0.0004894306 0.04483123      0 41.66776
## 85a86b91-4f24-4e77-ae2d-520f8e205efc 0.0000000000 0.01184285      0 30.18711
##                                     EPCAM      PROM1      CD24      EGFR
## 43e715bf-28d9-4b5e-b762-8cd1b69a430e 4.521984 0.025311008 0.55036003 1.286481
## 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872 9.530414 0.023576862 9.67272890 5.373307
## 93b382e4-9c9a-43f5-bd3b-502cc260b886 42.358567 0.000000000 0.06939934 4.600918
## 1f39dadd-3655-474e-ba4c-a5bd32c97a8b 16.316524 0.007783431 0.84522244 3.010374
## 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb 12.529742 0.019204339 0.21369023 16.476552
## 85a86b91-4f24-4e77-ae2d-520f8e205efc 2.430109 0.043719865 4.95506593 2.010338
##                                     MET TNFRSF10B
## 43e715bf-28d9-4b5e-b762-8cd1b69a430e 0.9320235 12.80547
## 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872 8.0610999 31.46289
## 93b382e4-9c9a-43f5-bd3b-502cc260b886 0.1295387 65.57967
## 1f39dadd-3655-474e-ba4c-a5bd32c97a8b 2.9728030 24.31636
## 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb 19.7360055 21.11014
## 85a86b91-4f24-4e77-ae2d-520f8e205efc 8.6087283 37.91574
##                                     tcga.tcgab_barcode
## 43e715bf-28d9-4b5e-b762-8cd1b69a430e TCGA-OR-A5KU-01A-11R-A29S-07
## 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872 TCGA-P6-A50G-01A-22R-A29S-07
## 93b382e4-9c9a-43f5-bd3b-502cc260b886 TCGA-OR-A5K5-01A-11R-A29S-07
## 1f39dadd-3655-474e-ba4c-a5bd32c97a8b TCGA-OR-A5K4-01A-11R-A29S-07
## 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb TCGA-OR-A5LP-01A-11R-A29S-07
## 85a86b91-4f24-4e77-ae2d-520f8e205efc TCGA-PK-A5H9-01A-11R-A29S-07
##                                     tcga.cgc_sample_sample_type study
## 43e715bf-28d9-4b5e-b762-8cd1b69a430e Primary Tumor ACC
## 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872 Primary Tumor ACC
## 93b382e4-9c9a-43f5-bd3b-502cc260b886 Primary Tumor ACC
## 1f39dadd-3655-474e-ba4c-a5bd32c97a8b Primary Tumor ACC
## 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb Primary Tumor ACC
## 85a86b91-4f24-4e77-ae2d-520f8e205efc Primary Tumor ACC
##                                     sample_type
## 43e715bf-28d9-4b5e-b762-8cd1b69a430e cancer
## 1a5db9fc-2abd-4e1b-b5ef-b1cf5e5f3872 cancer
## 93b382e4-9c9a-43f5-bd3b-502cc260b886 cancer
## 1f39dadd-3655-474e-ba4c-a5bd32c97a8b cancer
## 8c8c09b9-ec83-45ec-bc4c-0ba92de60acb cancer
## 85a86b91-4f24-4e77-ae2d-520f8e205efc cancer

```

With everything in a single dataframe, we are ready to do anything you want:) You will notice that this is the dataset we used in previous lectures and I just showed you how to get it from the source.

With everything now in a single dataframe, you can proceed with various data analysis tasks, such as identifying patterns, conducting differential expression analysis, or generating visualizations to gain insights from this comprehensive TCGA dataset.

11.6.12 Step 9: Create a Gene Expression Heatmap

To create the gene expression heatmap, we will follow these steps:

```
# import ComplexHeatmap
library(ComplexHeatmap)
```

11.6.12.1 Filter and Transform Data:

First, we'll filter the data to include only cancer samples, calculate the median expression for each gene within the same cancer type, and take the logarithm of the expression values to create a log2 transformation. This makes the data suitable for heatmap visualization.

```
tcga_df <- final_df %>%
  dplyr::filter(sample_type == "cancer") %>%
  group_by(sample_type, study) %>%
  summarise(across(1:20, ~log2(.x + 1))) %>%
  summarise(across(1:20, median)) %>%
  arrange(study) %>%
  dplyr::filter(!is.na(sample_type))
```

11.6.12.2 Create a Gene Expression Matrix:

Let's convert the summarized data into a matrix format. Rows represent cancer types, columns represent genes, and the matrix values are the median log2-transformed expression levels.

```
# Create a matrix 'tcga_mat' from 'tcga_df' excluding the first two columns and convert
tcga_mat <- tcga_df[, -c(1, 2)] %>% as.matrix()
rownames(tcga_mat) <- tcga_df %>% pull(study)

dim(tcga_mat)
```

```
## [1] 32 20
```

`dim(tcga_mat)` tells us that we have 32 rows (cancer types) and 20 columns (genes) in the resulting matrix. Let's print the matrix:

```
tcga_mat
```

```
##      TACSTD2      VTCN1      MUC1      NECTIN4      FOLH1      FOLR1      CD276
## ACC  0.4807693  0.000000000  0.2410192  0.04359167  1.6264037  0.02425151  5.297494
## BLCA 9.3725643  2.220041940  5.5991868  6.57397122  0.8918716  0.30467605  5.381322
## BRCA 8.0335879  5.770754248  7.5184905  5.51512285  1.4233695  1.39552103  5.784310
## CESC 9.7498568  1.914881813  6.5567883  6.43148129  0.7041431  0.87034258  4.965822
## CHOL 5.5273588  6.051437587  5.1893312  4.24539532  1.5456811  3.20878730  5.811681
## COAD 3.3928422  0.074410401  5.4975137  3.38347931  0.8062397  1.13574612  5.090426
## DLBC 0.6352793  0.002451433  1.1990101  1.01033523  0.2645891  0.03749739  3.266925
## ESCA 8.2368067  1.499102374  5.9425749  4.73125564  1.1132842  0.50368752  5.398227
## GBM  0.5207019  0.204680580  2.7142306  0.21242126  2.3137440  2.37227202  5.916635
## HNSC 9.0661567  0.694964205  4.0868590  6.08414302  1.1091992  0.38462756  5.830814
## KICH 2.7171000  0.145383937  5.5732785  0.38002326  0.9940703  1.89521752  4.194597
## KIRC 2.7574839  0.307661347  4.6082145  0.28522181  3.8988559  5.87646625  4.802535
## KIRP 6.5632546  2.287412718  4.9965286  0.17556715  0.4862326  5.84513870  4.926838
## LGG  0.1926966  0.341667316  1.7542645  0.07059579  2.5460983  1.14655021  4.305272
## LIHC 0.6768282  0.097613683  0.4391886  0.14510725  2.4297942  0.24022185  4.218155
## LUAD 7.8276755  1.355902742  8.4823052  5.09395755  1.1373591  6.45229154  5.201273
## LUSC 8.4101169  3.040923288  5.2575659  5.71623462  1.9452955  2.82656424  5.812421
## MESO 1.2488321  0.053038967  4.5448185  0.57555047  0.8397168  0.21716736  5.864723
## OV   7.3244772  5.414729025  7.7623819  3.37780261  1.3177417  8.16821960  5.019641
## PAAD 8.0663095  3.093104563  8.1649247  5.04149982  1.1797304  3.50409020  5.631054
## PCPG 1.0556914  0.020899233  0.2936292  0.13453283  1.4682181  0.81474148  4.896633
## PRAD 8.7594024  1.127776607  2.9181243  4.41268848  7.9254012  0.68870009  5.627149
## READ 3.4642158  0.056164097  5.6436544  3.24900294  0.7618605  1.61010910  5.087184
## SARC 1.2359343  0.139256548  2.6489240  0.50282829  1.5561093  0.25330451  6.600442
## SKCM 2.2675734  0.097902971  1.4645616  1.38994449  0.6182104  0.07875303  6.006916
## STAD 6.6014640  0.525928740  7.3745946  2.94416228  0.8882602  1.23634263  4.948302
## TGCT 2.0988218  0.338403710  1.0458394  2.27417760  0.7553944  2.78491077  5.465165
## THCA 8.2074834  0.569348321  5.1203104  4.15113608  1.3056050  5.00909622  4.848942
## THYM 3.4738470  0.017066392  1.1953473  1.10766635  0.2974423  0.83108801  3.581791
## UCEC 7.8256710  5.512091113  7.5435743  4.42551293  2.9885794  5.72164350  5.635153
## UCS  4.6713276  2.773715527  4.5590560  2.71970946  2.1128680  3.81105575  5.946331
## UVM  0.1193770  0.015265577  0.5954239  0.02765640  0.4841875  0.08134510  4.948453
##      MSLN      CLDN6      ERBB2      MUC16      DLL3      CEACAM5
## ACC  0.04481252  0.100146747  1.671858  0.0012205866  0.272796764  0.000000000
## BLCA 0.68716010  0.070442282  5.919806  0.0336637234  0.208256653  1.778875643
## BRCA 0.44071809  0.102271454  6.271432  0.1706249992  0.127821241  2.086550224
## CESC 5.62058497  0.093658539  5.440120  1.5432603634  0.829009483  5.852352010
```

```

## CHOL 0.60189628 0.327171872 5.643004 0.0085288100 0.039092934 0.621102898
## COAD 4.45458393 0.074751017 5.451245 0.0076928548 0.187729160 10.878821129
## DLBC 0.35785162 0.003087473 2.287948 0.0233179585 0.993488458 0.013941585
## ESCA 2.47373379 0.181754311 5.399585 0.2162955799 0.201342375 7.197142265
## GBM 0.92682384 0.149746286 3.975576 0.0013188121 4.821499549 0.0000000000
## HNSC 1.96099164 0.095296810 4.968078 0.2377308618 0.339048581 4.297622737
## KICH 0.05250724 0.043960617 4.687870 0.0007542286 0.010856365 0.0000000000
## KIRC 0.62610191 0.065242760 4.640322 0.0136965478 0.022333527 0.012439869
## KIRP 3.31497246 0.135890060 5.756473 0.0041568708 0.132753240 0.007485503
## LGG 0.72084397 0.159661042 3.215578 0.0011538098 7.342314792 0.0000000000
## LIHC 0.03555045 0.053806519 3.900610 0.0014097586 0.002397106 0.005960490
## LUAD 6.40328757 0.515735694 5.797269 1.1197883023 0.342515956 6.795916804
## LUSC 2.66588838 0.148816023 4.807536 0.2162983747 0.544306196 4.197498762
## MESO 9.10430072 0.423970579 4.832141 1.6452238522 0.233216970 0.0000000000
## OV 8.95238588 5.140635778 5.072491 4.6261895707 0.971504155 0.021736955
## PAAD 7.51409138 0.599000931 5.573353 1.3005505006 0.112664555 7.670442709
## PCPG 0.04768524 0.141276749 2.243807 0.0006859763 0.176422947 0.0000000000
## PRAD 0.69895768 0.047194319 5.811768 0.0312615665 0.304708642 0.605374582
## READ 4.71118780 0.071464799 5.539358 0.0059694499 0.273579598 11.213512117
## SARC 0.18614478 0.128843904 4.080999 0.0034064531 0.166898040 0.0000000000
## SKCM 0.10666438 0.063647636 4.131557 0.0052637362 2.811811044 0.132423575
## STAD 5.24125577 0.150099973 5.290161 0.0541710205 0.117172504 7.877833103
## TGCT 0.51413379 7.520373820 4.118593 0.1542327692 3.401171404 0.070429833
## THCA 0.33791649 0.397080016 6.005224 0.0813373551 0.183221365 0.019130524
## THYM 0.22949846 0.157873855 3.329512 0.0079995046 0.096684822 0.0000000000
## UCEC 5.00045864 0.679964295 5.431228 2.5515510076 0.244214559 0.775156469
## UCS 3.32976790 3.507766256 5.372831 0.5394212880 2.112957592 0.177189912
## UVM 0.04446844 0.0000000000 3.891753 0.0000000000 1.581155709 0.0000000000
## PVR EPCAM PROM1 CD24 EGFR MET TNFRSF10B
## ACC 5.236533 2.54354419 0.01273144 0.3505085 2.5059929 1.224786 4.334822
## BLCA 4.616088 6.36016443 0.20996061 8.1322461 4.0402929 4.300953 4.400105
## BRCA 3.834349 7.34815989 2.38468060 9.2653878 2.3867813 2.572547 4.026348
## CESC 4.730008 6.26554142 0.16395043 8.4236687 4.6460625 4.552453 4.979278
## CHOL 5.344069 7.97991762 4.37980601 8.4224757 3.9911158 4.949958 5.485029
## COAD 5.252634 9.60464201 4.47048763 9.3837212 3.2617795 5.377500 5.022903
## DLBC 2.163426 0.65838277 0.04557316 4.9255259 0.7242894 1.279921 4.059292
## ESCA 5.327993 7.57273004 2.16415031 8.6033079 5.3580321 5.291490 4.893869
## GBM 4.081546 0.38659163 2.68647622 5.1230942 6.7923898 2.005737 4.332658
## HNSC 4.922199 4.68395428 0.11238413 8.1208121 5.7435940 4.889075 4.895662
## KICH 5.132339 6.56645497 0.65202457 8.2691455 3.2130325 6.078532 4.564581
## KIRC 4.517746 4.65016367 2.21935887 10.0347534 5.6655666 6.309732 5.241214
## KIRP 4.655545 6.17909091 4.99464557 10.6188107 3.9479224 6.983211 5.494021
## LGG 3.433609 0.74956575 1.74465625 6.1547041 5.9290943 0.682397 3.480682
## LIHC 4.980117 0.78819179 0.06097059 5.6470623 3.6541165 4.775692 3.945549
## LUAD 4.743134 8.26018569 1.72616107 7.4193353 4.3624974 5.558406 4.989287
## LUSC 4.620441 7.12538734 0.97017148 7.5103897 5.2339820 5.074190 4.589941

```

```

## MESO 4.577427 1.10033694 0.05612307 2.2708816 4.6625905 5.780901 5.282852
## OV 3.750990 7.69797342 1.18943790 8.8827993 2.3811636 2.427727 3.382974
## PAAD 5.080733 8.06801056 4.61622807 8.5264790 3.7740009 5.247906 5.201323
## PCPG 4.892586 2.84560470 0.07544391 8.2692372 0.8566280 1.003257 2.480249
## PRAD 4.821357 7.61044397 0.75473658 7.6350131 4.1718705 2.386331 3.680519
## READ 5.301009 9.66106282 4.36896752 9.3307734 3.3160883 5.297970 4.827250
## SARC 4.520111 0.28448581 0.15045652 3.3333123 3.7024832 2.351284 4.271224
## SKCM 4.571221 0.26562350 0.58196240 1.8289097 0.9286357 3.559768 4.674391
## STAD 4.997742 8.32728127 3.93302612 8.9745717 3.8687001 4.725972 4.507678
## TGCT 5.091183 6.87150918 4.57983863 5.9294673 1.7305245 1.975280 3.881551
## THCA 3.901983 8.42405535 0.27570756 9.8733054 4.3838020 7.102301 4.882152
## THYM 2.833464 3.87134280 0.23738095 1.6799685 4.2379247 3.469279 4.077410
## UCEC 4.354987 8.44878219 4.28734222 9.3528813 2.6807882 3.738476 4.634779
## UCS 4.348322 6.58161996 3.48922500 7.8918205 3.3461966 3.826455 4.321049
## UVM 3.877644 0.08178735 0.13753182 0.1211111 0.4593655 6.347510 3.523352

```

11.6.12.3 Define a Cell Function for Grid Lines

```

cell_fun = function(j, i, x, y, w, h, fill) {
  grid.rect(x = x, y = y, width = w, height = h, gp = gpar(col = "black", fill = NA))
}

```

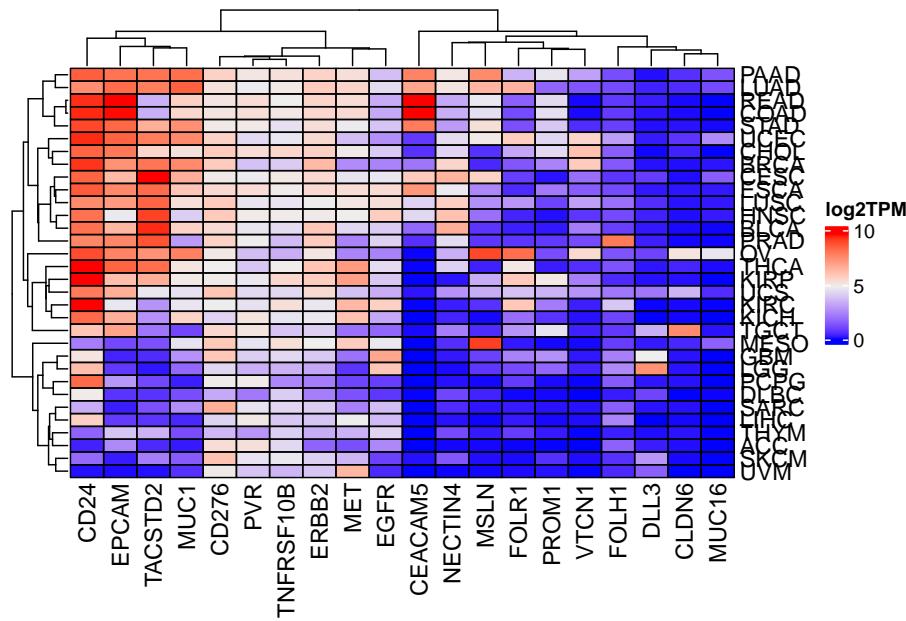
We define a custom cell function (`cell_fun`) to add black grid lines to the heatmap for better visualization.

11.6.12.4 Create the Heatmap

```

Heatmap(tcga_mat, cluster_columns = TRUE, cell_fun = cell_fun, name = "log2TPM")

```



Now, we create the heatmap using the `Heatmap` function from the `ComplexHeatmap` package. We specify `cluster_columns = TRUE` to cluster the columns (genes) for better visualization. The `cell_fun` parameter is set to our custom function for adding grid lines, and we name the data as “log2TPM.”

The resulting heatmap will show the median gene expression levels for each gene across different cancer types.

Grid lines will help in distinguishing individual cells within the heatmap so we added them. This visualization can provide insights into gene expression patterns in various cancer types, which may be useful for further analysis and interpretation.

11.6.13 Step 10: Sanity Check and Scaling of Gene Expression

11.6.13.1 Sanity Check

Before proceeding with scaling, we will conduct a sanity check of the gene expression heatmap to see if the results make biological sense. For example, we can observe if certain genes are highly expressed in specific cancer types, which could indicate potential biomarkers or interesting biological phenomena.

```

# Sanity check
sanity_check_genes <- c("MSLN", "FOLH1")

# Extract the rows (genes) corresponding to sanity check genes
sanity_check_data <- tcga_mat[, colnames(tcga_mat) %in% sanity_check_genes]

# Print the results
print(sanity_check_data)

##          FOLH1      MSLN
## ACC  1.6264037 0.04481252
## BLCA 0.8918716 0.68716010
## BRCA 1.4233695 0.44071809
## CESC 0.7041431 5.62058497
## CHOL 1.5456811 0.60189628
## COAD 0.8062397 4.45458393
## DLBC 0.2645891 0.35785162
## ESCA 1.1132842 2.47373379
## GBM  2.3137440 0.92682384
## HNSC 1.1091992 1.96099164
## KICH 0.9940703 0.05250724
## KIRC 3.8988559 0.62610191
## KIRP 0.4862326 3.31497246
## LGG  2.5460983 0.72084397
## LIHC 2.4297942 0.03555045
## LUAD 1.1373591 6.40328757
## LUSC 1.9452955 2.66588838
## MESO 0.8397168 9.10430072
## OV   1.3177417 8.95238588
## PAAD 1.1797304 7.51409138
## PCPG 1.4682181 0.04768524
## PRAD 7.9254012 0.69895768
## READ 0.7618605 4.71118780
## SARC 1.5561093 0.18614478
## SKCM 0.6182104 0.10666438
## STAD 0.8882602 5.24125577
## TGCT 0.7553944 0.51413379
## THCA 1.3056050 0.33791649
## THYM 0.2974423 0.22949846
## UCEC 2.9885794 5.00045864
## UCS  2.1128680 3.32976790
## UVM  0.4841875 0.04446844

```

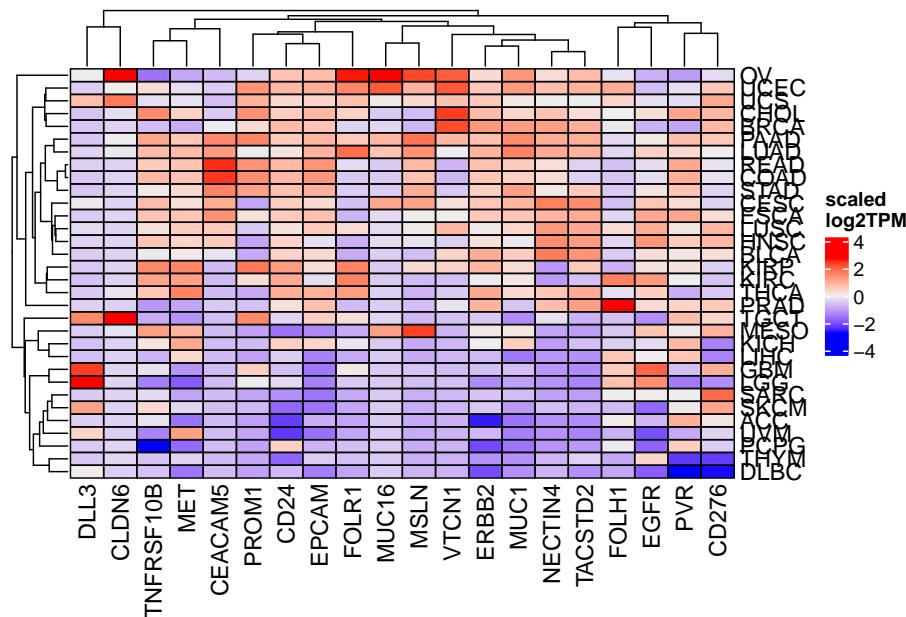
We see MSLN is high in MESO, FOLH1 is high in prostate cancer (PRAD). We are probably on the right track!

11.6.13.2 Scaling the data

To visualize gene expression in a more comparative manner, we can scale the expression values for each gene across the cancer types. Scaling standardizes the data, making it easier to identify relative expression levels.

```
# Scale the expression data
scaled_tcga_mat <- scale(tcga_mat)

# Create a scaled heatmap
Heatmap(scaled_tcga_mat,
        cluster_columns = TRUE,
        cell_fun = cell_fun,
        name = "scaled\\nlog2TPM")
```



Here, we use the `scale()` function to standardize the expression values across cancer types for each gene. We then create a new heatmap using the scaled data.

By comparing the original and scaled heatmaps, we can gain insights into how the expression of genes varies relative to each other across different cancer types. This scaling helps us focus on the relative expression patterns rather than the absolute values.

11.6.14 Conclusion

In this comprehensive example, we have walked through various essential steps for performing gene expression analysis using R, focusing on the analysis of The Cancer Genome Atlas (TCGA) data as an illustrative example. Gene expression analysis is a fundamental component of genomics research, enabling us to uncover insights into the molecular mechanisms underlying complex biological processes, such as cancer.

11.7 Section completed

Congratulations on completing this section!

You've learned crucial genomics data handling and analysis techniques, like using the `GenomicRanges` package, analyzing CpG islands, converting gene IDs, and creating visualizations with `ComplexHeatmap`. These skills are vital for bioinformatics, giving you the confidence to navigate genomic data complexities.

Don't hesitate to interact with your peers and instructors in the Q&A section and comments. Share your experiences, ask questions, and offer support. Your engagement enriches everyone's learning journey.

As you approach the final project, remember that what you've learned is more than just theory—it's practical knowledge you can apply. Use this opportunity to showcase your skills in a real-world scenario. Dive into the challenge with enthusiasm and curiosity.

Good luck! Let's make this final project a success together!

Chapter 12

Final Project: Analyzing RNAseq Data from GEO

12.1 Final Project Overview

In this final project, we will work with RNAseq data obtained from the **GEO** database, specifically the dataset with ITPR3 and RELB knockout in the SW480 cell line under varying oxygen conditions. Our primary objectives are to clean and prepare the metadata, identify differentially expressed genes, explore data distribution, and visualize gene expression patterns.

The steps we'll take:

1. To start, we will download the RNAseq count table and the associated metadata. With the help of the `dplyr` package in R, we will clean and organize the metadata. Our focus will be on selecting and subsetting the samples that are most relevant to our analysis - specifically, two wild-type samples under normoxia and two under hypoxia. For the initial phase of the project, we will disregard knockout samples to simplify our analysis.
2. Next, we will delve into exploring the RNAseq count matrix. Without considering knockout samples, we will calculate essential summary statistics to gain insights into gene expression levels' variability and distribution between the two conditions - normoxia and hypoxia.
3. After understanding the dataset's characteristics, we will proceed to identify differentially expressed genes. This step is crucial for uncovering genes that exhibit significant expression changes in response to varying oxygen levels. We will utilize well-established tool, `DESeq2`, to perform differential gene expression analysis.

4. Following the identification of differentially expressed genes, we will continue exploring the count matrix by calculating summary statistics specifically for these genes. This allows us to gain a deeper understanding of how their expression patterns differ between normoxia and hypoxia conditions.
5. To visualize data p-value distribution more effectively, we will create histograms. These histograms will offer a visual representation of how p values are distributed.
6. Additionally, we will use boxplots to compare gene expression levels between normoxia and hypoxia. Boxplots provide a concise summary of the data's central tendency, spread, and potential outliers, aiding in the identification of expression differences.
7. Principal Component Analysis (PCA) will be employed to obtain insights into the overall structure of the data and any potential clustering patterns. This dimensionality reduction technique will help us visualize how samples group based on their gene expression profiles.
8. Finally, we will create a heatmap using R. This heatmap will visualize the expression patterns of the identified differentially expressed genes across samples, providing a comprehensive view of how these genes respond to changes in oxygen levels in the SW480 cell line.

In summary, this project will take us through the full process of RNAseq data analysis, focusing on the hypoxia vs normoxia comparison in the SW480 cell line. We'll clean the data, pinpoint differentially expressed genes, explore their distributions, and visualize gene expression patterns.

Are you ready? Let's go!

12.2 How to pre-process RNAseq data

This is a bonus section on how to pre-process RNAseq data. In this course, we mainly focus on how to analyze RNAseq data for downstream analysis. We will start with a count matrix (next lesson) downloaded from GEO.

However, in real-world data analysis, sequencing data comes as a FASTQ file. FASTQ files are just normal text files with 4 lines for each read. Go to the link to understand the format.

Watch this video:

12.2.1 RNAseq pre-processing steps

1. The first step is to do Quality control of the FASTQ files using FASTQC.

2. Trim adaptors and low-quality bases using tools such as trimmomatic or fastp. Trimming of the reads is optional.
3. Align the reads to transcriptome using STAR. The single-cell RNAseq version is called STAR-solo from the same lab.
4. Quantify the number of reads fall into each gene using FeatureCounts.

I have written a Snakemake pipeline to pre-process RNAseq fastq file to get a count matrix at <https://github.com/crazyhottommy/pyflow-RNAseq>

The bash script to align the fastq files to transcriptome using STAR:

```
STAR --runMode alignReads \
    --runThreadN 5 \
    --genomeDir /path/to/the/STAR/index \
    --genomeLoad NoSharedMemory \
    --readFilesIn mysample_R1.fastq.gz mysample_R2.fastq.gz \
    --readFilesCommand zcat \
    --twoPassMode Basic \
    --runRNGseed 777 \
    --outFilterType Normal \
    --outFilterMultimapNmax 20 \
    --outFilterMismatchNmax 10 \
    --outFilterMultimapScoreRange 1 \
    --outFilterMatchNminOverLread 0.33 \
    --outFilterScoreMinOverLread 0.33 \
    --outReadsUnmapped None \
    --alignIntronMin 20 \
    --alignIntronMax 500000 \
    --alignMatesGapMax 1000000 \
    --alignSjOverhangMin 8 \
    --alignSjStitchMismatchNmax 5 -1 5 5 \
    --sjdbScore 2 \
    --alignSjDboverhangMin 1 \
    --sjdbOverhang 100 \
    --chimSegmentMin 20 \
    --chimJunctionOverhangMin 20 \
    --chimSegmentReadGapMax 3 \
    --quantMode GeneCounts \
    --outMultimapperOrder Random \
    --outSAMstrandField intronMotif \
    --outSAMattributes All \
    --outSAMunmapped Within KeepPairs \
    --outSAMtype BAM Unsorted \
    --limitBAMsortRAM 30000000000 \
    --outSAMmode Full \
```

```
--outSAMheaderHD @HD VN:1.4 \
--outFileNamePrefix mysample
```

Then quantifying using `FeatureCounts`:

```
featureCounts -T 5 -p -t exon -g gene_id -a gene.gtf -o mysample_featureCount.txt mysample
```

`mysample_featureCount.txt` will be a count table for one sample.

Alternative Alignment-free RNAseq quantification tools such as `salmon` and `kallisto` are also very popular. I recommend you to read the tutorial of `STAR`, `FeatureCounts`, `salmon` and `kallisto` to learn how to use those command line tools.

12.2.2 How to use salmon to preprocess GEO fastq to counts

Please refer to this blog post and this youtube video if you want to learn more:

12.3 Download and subset Count Matrix

In this lesson, you will learn how to explore a count matrix in R, a common task in data analysis. We'll cover downloading the data, reading it into R, examining the data's dimensions and column names, subsetting the data, converting it into a matrix, and adding row names.

12.3.1 Downloading the Count Matrix

To begin, we need to download the count matrix, which is a tab-separated values (TSV) file containing gene expression data. You can obtain it from the following FTP address: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE197nnn/GSE197576/suppl/GSE197576_raw_gene_counts_matrix.tsv.gz

You can use the `wget` command in Unix to download the processed count matrix file as follows:

```
wget https://ftp.ncbi.nlm.nih.gov/geo/series/GSE197nnn/GSE197576/suppl/GSE197576_raw_g
```

If you don't have access to a Unix-like command-line environment, you can download the file manually through your web browser. Simply open your web browser and go to the following URL: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE197nnn/GSE197576/suppl/GSE197576_raw_gene_counts_matrix.tsv.gz

12.3.2 Reading the Count Matrix in R

Now that you have the data, let's read it into R using the `readr` package. The first step is to load the required libraries and read the TSV file:

```
library(dplyr)
library(readr)

raw_counts <- read_tsv("~/Downloads/GSE197576_raw_gene_counts_matrix.tsv.gz")
```

12.3.3 Examining the Data

It's crucial to understand the data structure. We'll start by examining the dimensions of the data and the column names:

```
# Check the dimensions of the data
dim(raw_counts) # This will show the number of rows and columns
```

```
## [1] 43809    13

# List the column names
colnames(raw_counts) # This will display the names of all columns
```

```
## [1] "gene"                  "01_SW_sgCTRL_Norm"   "02_SW_sgCTRL_Norm"
## [4] "03_SW_sgITPR3_1_Norm"  "04_SW_sgITPR3_1_Norm"  "07_SW_sgRELB_3_Norm"
## [7] "08_SW_sgRELB_3_Norm"   "11_SW_sgCTRL_Hyp"     "12_SW_sgCTRL_Hyp"
## [10] "13_SW_sgITPR3_1_Hyp"  "14_SW_sgITPR3_1_Hyp"  "17_SW_sgRELB_3_Hyp"
## [13] "18_SW_sgRELB_3_Hyp"
```

The first six rows:

```
head(raw_counts)

## # A tibble: 6 x 13
##   gene      `01_SW_sgCTRL_Norm` `02_SW_sgCTRL_Norm` `03_SW_sgITPR3_1_Norm` 
##   <chr>        <dbl>          <dbl>            <dbl>                
## 1 DDX11L1       0              0                0
## 2 WASH7P        18             11               28
## 3 MIR6859-1     5              1                6
## 4 MIR1302-2HG   0              0                0
## 5 MIR1302-2     0              0                0
## 6 FAM138A       0              0                0
```

```
## # i 9 more variables: `04_SW_sgITPR3_1_Norm` <dbl>,
## #   `07_SW_sgRELB_3_Norm` <dbl>, `08_SW_sgRELB_3_Norm` <dbl>,
## #   `11_SW_sgCTRL_Hyp` <dbl>, `12_SW_sgCTRL_Hyp` <dbl>,
## #   `13_SW_sgITPR3_1_Hyp` <dbl>, `14_SW_sgITPR3_1_Hyp` <dbl>,
## #   `17_SW_sgRELB_3_Hyp` <dbl>, `18_SW_sgRELB_3_Hyp` <dbl>
```

Notice that the first column name is the gene name, and the other 12 columns are the sample names.

12.3.4 Subsetting the Data

Next, let's narrow down our data to the specific samples we need for comparison. To do this, we'll create a logical vector by matching column names that contain "sgCTRL" or "gene":

```
columns_to_select <- colnames(raw_counts) %>%
  stringr::str_detect("sgCTRL|gene")

columns_to_select

## [1] TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE
## [13] FALSE
```

The `stringr::str_detect` function searches for patterns in the column names. The resulting `columns_to_select` variable is a logical vector that helps us select the relevant columns.

Now, let's use this logical vector to subset the data frame:

```
counts_sub <- raw_counts[, columns_to_select]

head(counts_sub)

## # A tibble: 6 x 5
##   gene      `01_SW_sgCTRL_Norm` `02_SW_sgCTRL_Norm` `11_SW_sgCTRL_Hyp` 
##   <chr>        <dbl>          <dbl>            <dbl>      
## 1 DDX11L1       0              0              0
## 2 WASH7P        18             11             23
## 3 MIR6859-1     5              1              8
## 4 MIR1302-2HG   0              0              1
## 5 MIR1302-2     0              0              0
## 6 FAM138A       0              0              0
## # i 1 more variable: `12_SW_sgCTRL_Hyp` <dbl>
```

12.3.5 Converting to a Matrix

To perform various analyses, it's often more convenient to work with a matrix. We'll remove the first column (gene names) and convert the data frame to a matrix:

```
#subset the dataframe by removing the first column using negative index
# and then use as.matrix to convert it to a matrix
raw_counts_mat<- counts_sub[, -1] %>%
  as.matrix()

head(raw_counts_mat)
```

	01_SW_sgCTRL_Norm	02_SW_sgCTRL_Norm	11_SW_sgCTRL_Hyp	12_SW_sgCTRL_Hyp
## [1,]	0	0	0	0
## [2,]	18	11	23	45
## [3,]	5	1	8	12
## [4,]	0	0	1	0
## [5,]	0	0	0	0
## [6,]	0	0	0	0

Here, we use the `%>%` (pipe) operator to perform multiple operations sequentially. The `as.matrix()` function converts the data frame to a matrix.

12.3.6 Adding Row Names

The matrix lacks row names, which can be crucial for identifying genes. We can add the gene names as row names:

```
rownames(raw_counts_mat) <- raw_counts$gene

head(raw_counts_mat)
```

	01_SW_sgCTRL_Norm	02_SW_sgCTRL_Norm	11_SW_sgCTRL_Hyp	12_SW_sgCTRL_Hyp
## DDX11L1	0	0	0	0
## WASH7P	18	11	23	45
## MIR6859-1	5	1	8	12
## MIR1302-2HG	0	0	1	0
## MIR1302-2	0	0	0	0
## FAM138A	0	0	0	0
##	12_SW_sgCTRL_Hyp			
## DDX11L1	0			
## WASH7P	45			

```
## MIR6859-1          12
## MIR1302-2HG        0
## MIR1302-2          0
## FAM138A             0
```

Now, our matrix has gene names associated with each row.

12.4 Calculate the total exon length per gene

We want to find the differentially expressed genes between hypoxia and normoxia. The ideal workflow is to use the DESeq2 R package, which models the count data with the negative binomial distribution. For now, let's use a t-test to compare and we will use DESeq2 later.

However, because different samples have different sequencing depths, and different genes have different lengths, we must first normalize the counts to transcript per million (TPM).

12.4.1 Why Normalize Gene Counts to TPM?

When working with gene expression data, it's essential to account for variations in sequencing depths (the number of reads obtained for each sample) and gene lengths (some genes are longer than others). Normalization helps ensure that our gene expression values are comparable across different samples and genes.

Transcript Per Million (TPM) is a commonly used normalization method in RNA-seq analysis. It scales the gene counts to a common unit (per million) based on gene length and sequencing depth.

The general process:

1. For each gene, we divide its raw count values by its total exon length. This step essentially converts the raw counts into counts per unit exon length.
2. We sum up the values for each column (sample). This calculation gives us the total counts for each sample.
3. For each gene in each sample, we divide the value obtained in Step 1 by the total count for that sample calculated in Step 2. This step scales the counts relative to the sequencing depth of each sample.
4. To bring the values to a common scale and make them more interpretable, we multiply the result from Step 3 by 1,000,000 (1e6). This step converts the values to TPM, where the final unit is "Transcripts Per Million."

The normalization process ensures that the gene expression values are now comparable across different samples, making it easier to identify genes that are differentially expressed between conditions (e.g., hypoxia and normoxia).

12.4.2 Creating a function for normalization

Let's write a function. Before that, we need to know the gene length of all the genes in the data frame.

We will use the Bioconductor package `TxDb.Hsapiens.UCSC.hg19.knownGene` to get the gene length, or more exactly, the total exon lengths for each gene (most of the RNAseq reads are from the exons).

```
#if (!require("BiocManager", quietly = TRUE))
#   install.packages("BiocManager")

#BiocManager::install("TxDb.Hsapiens.UCSC.hg19.knownGene")

library(TxDb.Hsapiens.UCSC.hg19.knownGene)

TxDb.Hsapiens.UCSC.hg19.knownGene

## TxDb object:
## # Db type: TxDb
## # Supporting package: GenomicFeatures
## # Data source: UCSC
## # Genome: hg19
## # Organism: Homo sapiens
## # Taxonomy ID: 9606
## # UCSC Table: knownGene
## # Resource URL: http://genome.ucsc.edu/
## # Type of Gene ID: Entrez Gene ID
## # Full dataset: yes
## # miRBase build ID: GRCh37
## # transcript_nrow: 82960
## # exon_nrow: 289969
## # cds_nrow: 237533
## # Db created by: GenomicFeatures package from Bioconductor
## # Creation time: 2015-10-07 18:11:28 +0000 (Wed, 07 Oct 2015)
## # GenomicFeatures version at creation time: 1.21.30
## # RSQLite version at creation time: 1.0.0
## # DBSCHEMVERISON: 1.1
```

It is a `TxDb` object. We can use functions such as `genes` and `exons` to get the genes or exons.

240 CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

```
# make it a shorter name
txdb<- TxDb.Hsapiens.UCSC.hg19.knownGene

genes(txdb)

## GRanges object with 23056 ranges and 1 metadata column:
##      seqnames      ranges strand |   gene_id
##           <Rle>      <IRanges> <Rle> | <character>
##    1     chr19  58858172-58874214   - |       1
##   10    chr8   18248755-18258723   + |      10
##  100   chr20  43248163-43280376   - |     100
## 1000  chr18  25530930-25757445   - |   1000
## 10000 chr1  243651535-244006886   - | 10000
## ...
## 9991  chr9  114979995-115095944   - | 9991
## 9992  chr21 35736323-35743440   + | 9992
## 9993  chr22 19023795-19109967   - | 9993
## 9994  chr6  90539619-90584155   + | 9994
## 9997  chr22 50961997-50964905   - | 9997
## -----
## seqinfo: 93 sequences (1 circular) from hg19 genome
```

It returns a `GRanges` object with the chromosome name, start, end, strand and the `gene_id`. Note that `gene_id` here is the ENTREZ ID.

However, to be accurate, we want the exons, not the whole genes which contain introns. Let's get the exons:

```
exons<- exonsBy(txdb, by = "gene")
exons

## GRangesList object of length 23459:
## $`1`
## GRanges object with 15 ranges and 2 metadata columns:
##      seqnames      ranges strand | exon_id exon_name
##           <Rle>      <IRanges> <Rle> | <integer> <character>
## [1]  chr19  58858172-58858395   - | 250809  <NA>
## [2]  chr19  58858719-58859006   - | 250810  <NA>
## [3]  chr19  58859832-58860494   - | 250811  <NA>
## [4]  chr19  58860934-58862017   - | 250812  <NA>
## [5]  chr19  58861736-58862017   - | 250813  <NA>
## ...
## [11] chr19  58868951-58869015   - | 250821  <NA>
## [12] chr19  58869318-58869652   - | 250822  <NA>
## [13] chr19  58869855-58869951   - | 250823  <NA>
```

```

## [14] chr19 58870563-58870689      - | 250824 <NA>
## [15] chr19 58874043-58874214      - | 250825 <NA>
## -----
## seqinfo: 93 sequences (1 circular) from hg19 genome
##
## $`10`
## GRanges object with 2 ranges and 2 metadata columns:
##   seqnames      ranges strand | exon_id  exon_name
##   <Rle>        <IRanges> <Rle> | <integer> <character>
## [1] chr8 18248755-18248855      + | 113603 <NA>
## [2] chr8 18257508-18258723      + | 113604 <NA>
## -----
## seqinfo: 93 sequences (1 circular) from hg19 genome
##
## ...
## <23457 more elements>

```

This returns a GRangesList object and each element of the list is a GRanges containing all the exons for that gene.

Let's calculate the total exon lengths for each gene by the `width` function:

```

# width of exons per gene
width(exons)

## IntegerList of length 23459
## [[1]] 224 288 663 1084 282 297 273 270 36 96 65 335 97 127 172
## [[10]] 101 1216
## [[100]] 326 103 130 65 102 72 128 116 144 123 62 161
## [[1000]] 1394 165 140 234 234 143 254 186 138 173 145 156 147 227 106 112 519
## [[10000]] 218 68 5616 50 103 88 215 129 123 69 66 132 145 112 126 158 41
## [[100008586]] 92 121 126 127
## [[100009676]] 2784
## [[10001]] 704 28 116 478 801 109 130 83 92 160 52
## [[10002]] 308 127 104 222 176 201 46 106 918 705
## [[10003]] 191 112 187 102 126 187 94 1193 99 ... 64 68 92 91 265 82 93 1054
## ...
## <23449 more elements>

# sum the exons width up per gene
head(sum(width(exons)))

```

	1	10	100	1000	10000	100008586
##	4309	1317	1532	4473	7459	466

Note that the `width` and `sum` functions are vectorized. It will calculate across all genes.

Let's turn it into a tibble using the `enframe` function:

```
exon_len<- sum(width(exons)) %>%
  tibble::enframe(name = "ENTREZID", value = "exon_length")

head(exon_len)

## # A tibble: 6 x 2
##   ENTREZID  exon_length
##   <chr>          <int>
## 1 1                  4309
## 2 10                 1317
## 3 100                1532
## 4 1000               4473
## 5 10000              7459
## 6 100008586         466
```

Next, let's map the ENTREZID to the official gene symbol so we can match the rownames of the RNAseq count matrix. We will need the `org.Hs.eg.db` Bioconductor package (install it if you do not have it).

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

map<- AnnotationDbi::select(org.Hs.eg.db,
                            keys = exon_len$ENTREZID,
                            columns= "SYMBOL",
                            keytype = "ENTREZID")

head(map)

##     ENTREZID SYMBOL
## 1 1          A1BG
## 2 10         NAT2
## 3 100        ADA
## 4 1000       CDH2
## 5 10000      AKT3
## 6 100008586 GAGE12F
```

12.4.3 join the exon length table

Read this article to understand different join functions in `dplyr`.

```
map<- left_join(exon_len, map)
head(map)

## # A tibble: 6 x 3
##   ENTREZID  exon_length SYMBOL
##   <chr>      <int> <chr>
## 1 1          4309  A1BG
## 2 10         1317  NAT2
## 3 100        1532  ADA
## 4 1000       4473  CDH2
## 5 10000      7459  AKT3
## 6 100008586 466   GAGE12F
```

One of the key problems with genomics is that gene IDs are not always 1:1 mappable. Different versions of the genome (hg19 vs hg38 for humans) may have slightly different gene symbols.

```
table(rownames(raw_counts_mat) %in% map$SYMBOL)
```

```
## 
## FALSE  TRUE
## 20559 23250
```

12.4.4 what genes are not in the mapping table?

```
base::setdiff(rownames(raw_counts_mat), map$SYMBOL) %>%
  head(n = 20)
```

```
## [1] "MIR6859-1"    "MIR1302-2HG"   "MIR1302-2"    "FAM138A"      "LOC100996442"
## [6] "DDX11L17"     "WASH9P"       "MIR6859-2"    "LOC107985721" "LOC112268260"
## [11] "LOC100132287" "LOC105378947" "LOC101928626" "MIR12136"     "LINC01409"
## [16] "FAM87B"       "LOC107984850"  "LOC284600"    "LOC107985728" "LOC100288175"
```

Most of the differences are from non-coding RNA (LOC genes) or microRNAs. Many of those genes have a limited number of counts, we can ignore them for the moment.

```
not_in_map<- setdiff(rownames(raw_counts_mat), map$SYMBOL)

raw_counts_mat[not_in_map, ] %>%
  head(n = 15)
```

244 CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

	01_SW_sgCTRL_Norm	02_SW_sgCTRL_Norm	11_SW_sgCTRL_Hyp
## MIR6859-1	5	1	8
## MIR1302-2HG	0	0	1
## MIR1302-2	0	0	0
## FAM138A	0	0	0
## LOC100996442	9	3	17
## DDX11L17	0	0	0
## WASH9P	52	32	68
## MIR6859-2	0	0	0
## LOC107985721	0	0	0
## LOC112268260	0	0	0
## LOC100132287	0	0	0
## LOC105378947	0	0	0
## LOC101928626	0	0	0
## MIR12136	3	1	0
## LINC01409	7	8	19
	12_SW_sgCTRL_Hyp		
## MIR6859-1		12	
## MIR1302-2HG		0	
## MIR1302-2		0	
## FAM138A		0	
## LOC100996442		15	
## DDX11L17		0	
## WASH9P		76	
## MIR6859-2		0	
## LOC107985721		0	
## LOC112268260		0	
## LOC100132287		0	
## LOC105378947		0	
## LOC101928626		0	
## MIR12136		0	
## LINC01409		24	

subset only the common genes for the map file and the count matrix. Make sure the order of the genes is the same for both data.

```
common_genes<- intersect(rownames(raw_counts_mat), map$SYMBOL)

## select only the common genes and re-order them by common_genes
map<- map %>%
  dplyr::slice(match(common_genes, SYMBOL))

# subset the common genes and re-order them by common_genes
raw_counts_mat<- raw_counts_mat[common_genes, ]

head(map)
```

```
## # A tibble: 6 x 3
##   ENTREZID exon_length SYMBOL
##   <chr>      <int> <chr>
## 1 100287102     2838 DDX11L1
## 2 653635        8050 WASH7P
## 3 79501          918 OR4F5
## 4 729737        5474 LOC729737
## 5 729759        1878 OR4F29
## 6 81399          939 OR4F16
```

The order of the genes is the same for `map` and `raw_counts_mat`.

```
head(raw_counts_mat)
```

	01_SW_sgCTRL_Norm	02_SW_sgCTRL_Norm	11_SW_sgCTRL_Hyp	12_SW_sgCTRL_Hyp
## DDX11L1	0	0	0	0
## WASH7P	18	11	23	45
## OR4F5	0	0	0	0
## LOC729737	3	3	16	16
## OR4F29	0	0	0	0
## OR4F16	1	0	0	3

12.5 Normalizing Raw Counts to Transcripts per Million (TPM)

TPM normalization is essential for comparing gene expression levels across different samples and genes. We will write a function in the R programming language to perform this conversion and explain the steps involved.

12.5.1 The `count2tpm` Function:

We will create a function called `count2tpm` in R to perform TPM normalization. This function takes two arguments: a count matrix and a vector of exon lengths. Let's break down the code step by step.

```
count2tpm <- function(count_matrix, exon_length) {
  # Calculate reads per base pair per gene
  reads_per_bp_gene <- count_matrix / exon_length
```

```

# Calculate the total reads per base pair for each sample
reads_per_bp_sample <- colSums(reads_per_bp_gene)

# Normalize to the library size and calculate TPM
tpm_matrix <- t(t(reads_per_bp_gene) / reads_per_bp_sample) * 1000000
return(tpm_matrix)
}

```

1. We start by defining the `count2tpm` function, which takes two arguments: `count_matrix` (raw gene expression counts) and `exon_length` (a vector of gene exon lengths).
2. We calculate the number of reads per base pair for each gene by dividing the count matrix by the exon length vector. This step helps us account for gene length differences.
3. We sum the reads per base pair values for each sample (column-wise) to calculate the total reads per base pair for each sample. This is crucial for library size normalization.
4. To normalize the data to library size, we divide the transposed `reads_per_bp_gene` matrix by the `reads_per_bp_sample` vector. The transposition allows us to perform element-wise division efficiently. Finally, we multiply the result by 1,000,000 to obtain TPM values.

12.5.2 Applying the Function

Now, let's apply the `count2tpm` function to our raw count matrix and exon length vector. Here's how you can do it:

```

tpm <- count2tpm(raw_counts_mat, map$exon_length)

head(tpm)

```

	01_SW_sgCTRL_Norm	02_SW_sgCTRL_Norm	11_SW_sgCTRL_Hyp	12_SW_sgCTRL_Hyp
## DDX11L1	0.0000000	0.0000000	0.0000000	0.0000000
## WASH7P	0.31352121	0.24557646	0.4016983	0.7781915
## OR4F5	0.0000000	0.0000000	0.0000000	0.0000000
## LOC729737	0.07684343	0.09849323	0.4109445	0.4068975
## OR4F29	0.0000000	0.0000000	0.0000000	0.0000000
## OR4F16	0.14932231	0.0000000	0.0000000	0.4447598

These values represent the TPM-normalized gene expression levels for each gene in different samples.

12.5.3 Conclusions

In this lesson, we have learned how to normalize raw gene expression counts to Transcripts per Million (TPM) using the `count2tpm` function in R. This normalization is crucial for comparing gene expression levels accurately across samples and genes, taking into account library size and gene length.

12.6 Analyzing Gene Expression Data Using t-Tests

Gene expression data provides valuable insights into how genes are activated or deactivated under different conditions, such as in response to diseases or environmental changes.

A t-test is a statistical test that helps us determine whether there is a significant difference between the means of two groups. In the context of gene expression analysis, we can use t-tests to identify genes that are differentially expressed between two experimental conditions. For example, we might want to know which genes are upregulated or downregulated in response to hypoxia (low oxygen levels) compared to normoxia (normal oxygen levels).

12.6.1 Hypothesis Testing

Before we dive into the code, let's understand the key components of hypothesis testing:

1. **Null Hypothesis (H_0):** This is the default assumption that there is no significant difference between the groups. In gene expression analysis, it means that the gene is not differentially expressed.
2. **Alternative Hypothesis (H_a):** This is the hypothesis we want to test. It suggests that there is a significant difference between the groups. In gene expression analysis, it implies that the gene is differentially expressed.
3. **p-value:** The p-value represents the probability of observing the data, assuming that the null hypothesis is true. A small p-value (typically less than 0.05) suggests that we can reject the null hypothesis and accept the alternative hypothesis.

12.6.2 Analyzing Specific Genes

Now, let's use t-tests to analyze the expression of specific genes and understand how they respond to hypoxia.

248CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

We'll start by examining the gene WASH7P. We perform a t-test to compare its expression levels between normoxia and hypoxia samples.

```
t.test(tpm["WASH7P", c(1,2)], tpm["WASH7P", c(3,4)])  
  
##  
## Welch Two Sample t-test  
##  
## data: tpm["WASH7P", c(1, 2)] and tpm["WASH7P", c(3, 4)]  
## t = -1.6227, df = 1.0651, p-value = 0.3404  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2.416652 1.795860  
## sample estimates:  
## mean of x mean of y  
## 0.2795488 0.5899449
```

In this case, the p-value is 0.3404, suggesting that there is no significant difference in the expression of the WASH7P gene between normoxia and hypoxia.

Next, we examine the VEGFA gene, which is known to be a key regulator of angiogenesis in response to hypoxia.

```
t.test(tpm["VEGFA", c(1,2)], tpm["VEGFA", c(3,4)])  
  
##  
## Welch Two Sample t-test  
##  
## data: tpm["VEGFA", c(1, 2)] and tpm["VEGFA", c(3, 4)]  
## t = -31.953, df = 1.8939, p-value = 0.00132  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -56.58778 -42.50235  
## sample estimates:  
## mean of x mean of y  
## 13.58050 63.12557
```

Here, the p-value is 0.00132, indicating a significant difference in VEGFA expression between normoxia and hypoxia. This suggests that VEGFA is likely upregulated under hypoxic conditions.

Now, let's analyze the SLC2A1 (GLUT1) gene, which plays a role in glucose transport during anaerobic glycolysis.

```
t.t.test(tpm["SLC2A1", c(1,2)], tpm["SLC2A1", c(3,4)])
```

```
##  
## Welch Two Sample t-test  
##  
## data: tpm["SLC2A1", c(1, 2)] and tpm["SLC2A1", c(3, 4)]  
## t = -31.938, df = 1.1196, p-value = 0.01354  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2460.877 -1295.615  
## sample estimates:  
## mean of x mean of y  
## 1360.405 3238.651
```

The p-value is 0.01354, indicating a significant difference in SLC2A1 expression between the two conditions. This suggests that SLC2A1 may be upregulated under hypoxic conditions as well.

12.6.3 Analyzing All Genes

To analyze all genes in our dataset, we can create a custom function called `mytest` that performs t-tests for each gene pair and extracts the p-values.

```
mytest <- function(x) t.t.test(x[c(1,2)], x[c(3,4)], var.equal = TRUE)$p.value  
pvals <- apply(tpm, 1, mytest)
```

```
head(pvals)
```

```
##      DDX11L1      WASH7P      OR4F5    LOC729737      OR4F29      OR4F16  
##      NaN 0.246130682      NaN 0.001173022      NaN 0.593224964
```

Here, we apply the `mytest` function to each row (gene) in our gene expression dataset (`tpm`) to calculate p-values.

Finally, we count how many genes have p-values smaller than 0.01 to identify differentially expressed genes:

```
sum(pvals < 0.01, na.rm = TRUE)
```

```
## [1] 3378
```

`pvals < 0.01` returns a logical vector of TRUE and FALSE. TRUE is 1 and FALSE is 0 under the hood in R. If you sum them up `sum(pvals < 0.01, na.rm = TRUE)` will tell you how many TRUES are in the vector.

There are 3378 genes with p-values smaller than 0.01!

12.6.4 Conclusion

In this lesson, we learned how to use t-tests to analyze gene expression data and identify differentially expressed genes. We examined specific genes and performed t-tests, understanding the significance of p-values and hypothesis testing. Additionally, we applied t-tests to all genes in the dataset to identify potential candidates for further investigation.

12.7 Analyzing Gene Expression Data with ggplot2

In this lesson, we will explore how to analyze gene expression data using the powerful ggplot2 library. We will focus on visualizing p-value distributions, comparing differentially expressed genes, and creating boxplots to gain insights into gene expression changes under different conditions.

12.7.1 Import Libraries and Load Data

First, let's import the necessary libraries and load your gene expression data. In this lesson, we assume you have a dataset containing gene names and p-values representing their significance.

```
# Import required libraries
library(ggplot2)
library(dplyr)
library(tidyr)

# Load your p-values data (replace 'pvals' with your actual dataset)
pval_df <- pvals %>%
  tibble::enframe(name = "gene", value = "pvalue")

# Display the first few rows of the p-value data
head(pval_df)

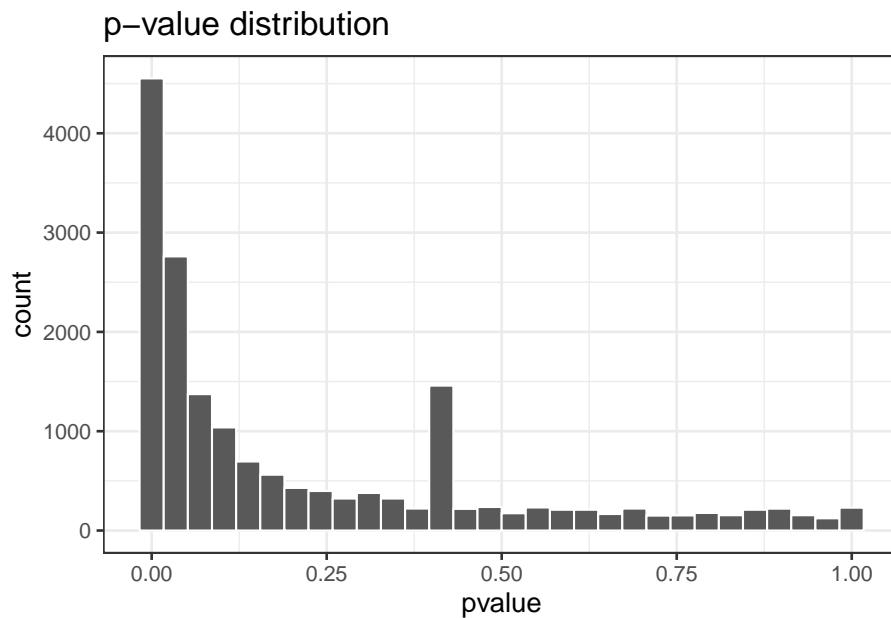
## # A tibble: 6 x 2
##   gene      pvalue
##   <chr>     <dbl>
## 1 DDX11L1    NaN
## 2 WASH7P     0.246
## 3 OR4F5      NaN
## 4 LOC729737  0.00117
## 5 OR4F29     NaN
## 6 OR4F16     0.593
```

This code converts your gene names and corresponding p-values into a data frame, making it easier to work with in ggplot2.

12.7.2 Visualizing P-Value Distribution

Next, let's create a histogram to visualize the distribution of p-values:

```
# Create a histogram of p-values
ggplot(pval_df, aes(x = pvalue)) +
  geom_histogram(color = "white") +
  theme_bw(base_size = 14) +
  ggtitle("p-value distribution")
```



This code uses ggplot2 to create a histogram, providing insights into the distribution of p-values across your genes. Understanding p-value distribution can help assess the significance of your results.

12.7.3 Identifying Differentially Expressed Genes

Now, we'll focus on comparing gene expression between hypoxia and normoxia conditions, specifically looking at up-regulated genes. We'll start by calculating the average expression levels for both conditions and identifying the up-regulated genes.

252CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

```
# Calculate average expression for normoxia and hypoxia conditions
avg_normoxia <- rowMeans(tpm[, c(1, 2)])
avg_hypoxia <- rowMeans(tpm[, c(3, 4)])

# Identify up-regulated genes
up_genes <- (avg_hypoxia - avg_normoxia) > 0

# Get the names of up-regulated genes
up_gene_names <- rownames(tpm)[up_genes]

head(up_gene_names)
```

```
## [1] "WASH7P"          "LOC729737"        "OR4F16"           "LOC100288069"    "LINC02593"
## [6] "SAMD11"
```

Here, we calculate the average expression for normoxia and hypoxia conditions and then identify up-regulated genes by comparing the averages. `up_gene_names` contains the names of these genes.

12.7.4 Selecting Differentially Expressed Genes

Not all up-regulated genes may be significantly different. Let's find the intersection of up-regulated genes with those that have a p-value less than 0.01 (significant up-regulation):

```
# Select differentially expressed genes (intersection of up-regulated genes and significant)
differential_genes <- pvals[pvals < 0.01 & !is.na(pvals)] %>%
  names()

# Find the intersection
differential_up_genes <- intersect(differential_genes, up_gene_names)

length(differential_up_genes)

## [1] 1990
```

`differential_up_genes` now contains the names of genes that are both up-regulated and significantly different under hypoxia.

12.7.5 boxplot to visualize gene expression changes between the two conditions.

12.7.5.1 preparing the data

Now, let's prepare our data for creating a boxplot to visualize gene expression changes between the two conditions. We will convert the gene expression data into a long format suitable for `ggplot2`:

```
# Convert gene expression data to long format
tpm[differential_genes, ] %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var = "gene") %>%
  tidyverse::pivot_longer(-1, names_to = "sample", values_to = "tpm")

## # A tibble: 13,512 x 3
##   gene     sample      tpm
##   <chr>    <chr>     <dbl>
## 1 LOC729737 01_SW_sgCTRL_Norm 0.0768
## 2 LOC729737 02_SW_sgCTRL_Norm 0.0985
## 3 LOC729737 11_SW_sgCTRL_Hyp  0.411
## 4 LOC729737 12_SW_sgCTRL_Hyp  0.407
## 5 NOC2L     01_SW_sgCTRL_Norm 128.
## 6 NOC2L     02_SW_sgCTRL_Norm 124.
## 7 NOC2L     11_SW_sgCTRL_Hyp  73.4
## 8 NOC2L     12_SW_sgCTRL_Hyp  70.9
## 9 PERM1     01_SW_sgCTRL_Norm 0.176
## 10 PERM1    02_SW_sgCTRL_Norm 0.113
## # i 13,502 more rows
```

This code converts the gene expression data into a long format with columns for gene names, sample names, and expression values.

Add another column to denote the condition by separating the sample column to two columns: sample and condition

```
tpm_df<- tpm[differential_genes, ] %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var="gene") %>%
  tidyverse::pivot_longer(-1, names_to = "sample", values_to = "tpm") %>%
  tidyverse::separate(sample, into = c("sample", "condition"), sep = "_sgCTRL_")

head(tpm_df)

## # A tibble: 6 x 4
```

```

##   gene      sample condition      tpm
##   <chr>    <chr>  <chr>       <dbl>
## 1 LOC729737 01_SW  Norm       0.0768
## 2 LOC729737 02_SW  Norm       0.0985
## 3 LOC729737 11_SW  Hyp        0.411
## 4 LOC729737 12_SW  Hyp        0.407
## 5 NOC2L     01_SW  Norm      128.
## 6 NOC2L     02_SW  Norm      124.

```

12.7.5.2 Customizing Boxplot Order

By default, ggplot2 orders boxplots alphabetically. To change the order, convert the condition column to a factor with the desired order:

```

# Define the order for boxplot
tpm_df$condition <- factor(tpm_df$condition, levels = c("Norm", "Hyp"))

```

This code ensures that the boxplot orders the conditions as “Norm” followed by “Hyp.”

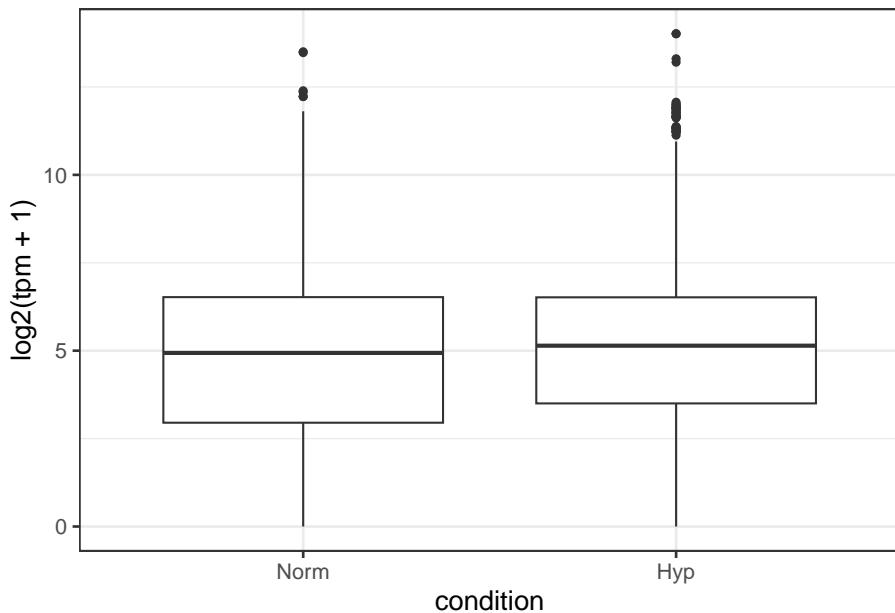
12.7.6 Creating the Boxplot

Now, we can create the boxplot to visualize gene expression changes between the two conditions:

```

# Create the boxplot
ggplot(tpm_df, aes(x = condition, y = log2(tpm + 1))) +
  geom_boxplot() +
  theme_bw(base_size = 14)

```

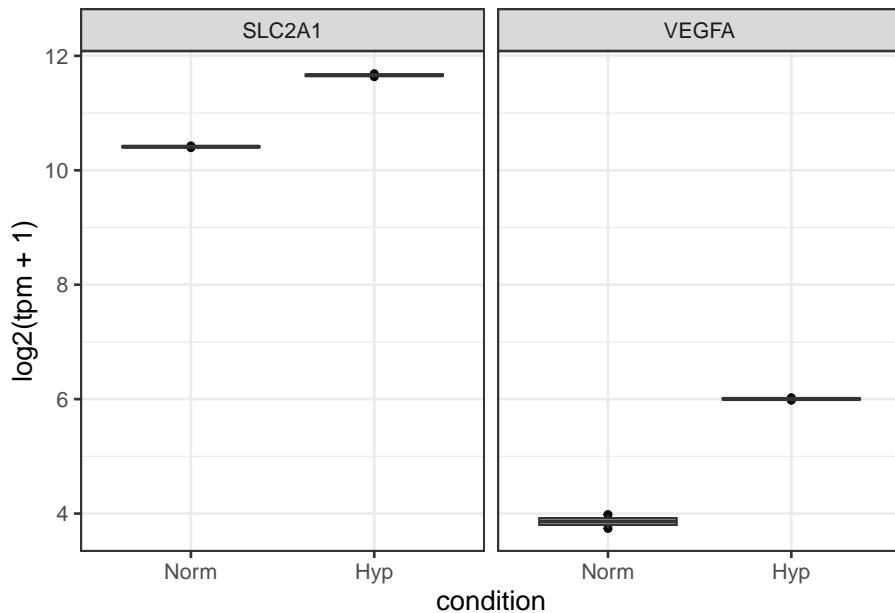


This code uses `ggplot2` to create a boxplot, showing the distribution of gene expression values between the “Norm” and “Hyp” conditions. The $\log_2(\text{tpm} + 1)$ transformation is often used to visualize RNA-Seq data.

12.7.7 Visualizing Raw Expression Values

Sometimes, it’s essential to examine the raw expression values to identify outliers. Here, we use a boxplot to visualize the raw values of selected genes:

```
# Select specific genes (e.g., "VEGFA" and "SLC2A1") for visualization
ggplot(tpm_df %>%
  filter(gene %in% c("VEGFA", "SLC2A1")),
  aes(x = condition, y = log2(tpm + 1))) +
  geom_point() +
  geom_boxplot() +
  facet_wrap(~ gene) +
  theme_bw(base_size = 14)
```



This code creates scatter plots for selected genes and overlays boxplots to help visualize the distribution of the gene expression levels.

12.7.8 Conclusion

In this lesson, we covered the entire process of analyzing gene expression data using ggplot2, from loading data to visualizing differential expression. Understanding the steps involved and customizing plots can provide valuable insights into your gene expression analysis.

12.8 Correcting for Multiple Comparisons in Statistical Analysis

This is a critical step when conducting hypothesis tests on a large number of data points, such as in genomics research. We will cover the need for correction, different methods to control errors, and demonstrate how to implement one of the widely-used methods, the False Discovery Rate (FDR) correction.

12.8.1 Why Correct for Multiple Comparisons?

Imagine you are a scientist studying the gene expression levels of thousands of genes in response to a treatment. You perform statistical tests to identify which

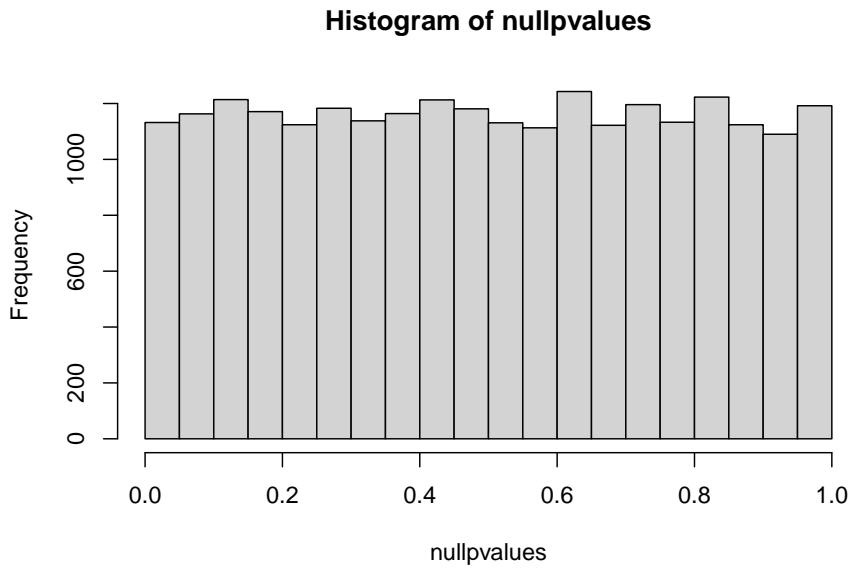
genes are significantly differentially expressed. If you run these tests without correction, you are likely to encounter a problem known as the “multiple comparisons problem.” In essence, the more tests you perform, the higher the chance of obtaining false positives (i.e., incorrectly identifying genes as significant).

To address this issue, we need correction methods that control the probability of making at least one false discovery while testing multiple hypotheses. In this project, we will focus on the **False Discovery Rate** (FDR) correction method, specifically using the **Benjamini & Hochberg** (BH) procedure.

12.8.2 The Multiple Comparisons Example

Let’s illustrate the concept with a practical example. Suppose you have gene expression data from 23,250 genes and you want to identify those that are differentially expressed between two conditions (e.g., control and treatment). You perform a statistical test for each gene and obtain p-values.

```
# Sample code to generate random p-values for demonstration purposes
m <- 23250 # Number of genes
n <- 100    # Number of comparisons
randomData <- matrix(rnorm(n * m), m, n)
nullpvalues <- apply(randomData, 1, mytest) # Simulated p-values
hist(nullpvalues)
```



258CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

If you were to plot the histogram of these p-values, you might expect them to follow a uniform distribution (a flat line) under the null hypothesis (no differential expression). However, due to the nature of p-values as random variables, you will still observe some p-values below the commonly used significance level of 0.05, even when no genes are differentially expressed.

Compare this histogram with the histogram for the real data. what do you see? Even if we randomly generated the data, you still see some p values are smaller than 0.05!! We randomly generated data, there should be No genes that deferentially expressed. However, we see a flat line across different p values.

p values are random variables. Mathematically, one can demonstrate that under the null hypothesis (and some assumptions are met, in this case, the test statistic T follows standard normal distribution), p-values follow a uniform (0,1) distribution, which means that $P(p < p_1) = p_1$.

This means that the probability we see a p value smaller than p_1 is equal to p_1 . That being said, with a 100 t-tests, under the null (no difference between control and treatment), we will see 1 test with a p value smaller than 0.01. And we will see 2 tests with a p value smaller than 0.02 etc.

We have 23250 genes in the matrix, and we did 23250 comparisons at one time. This explains why we see $23250 * 0.05 = 1162$ p-values are smaller than 0.05 in our randomly generated numbers. That's exactly what we see in the null distribution of the p-values.

In fact, checking the p-value distribution by histogram is a very important step during data analysis. You may want to read a blog post by David Robinson: How to interpret a p-value histogram.

12.8.3 Correcting for Multiple Comparisons: False Discovery Rate (FDR)

How do we control the false positives for multiple comparisons? One way is to use the Bonferroni correction to correct the familywise error rate (FWER): define a particular comparison as statistically significant only when the P value is less than alpha(often 0.05) divided by the number (m) of comparisons ($p < \alpha/m$).

Say we computed 100 t-tests, and got 100 p values, we only consider the genes with a p value smaller than $0.05/100$ as significant. This approach is very conservative and is used in Genome-wide association studies (GWAS). Since we often compare millions of genetic variations between (tens of thousands) cases and controls, this threshold will be very small!

Alternatively, we can use False Discovery Rate (FDR) to report the gene list. $FDR = \# \text{false positives} / \# \text{called significant}$. This approach does not use the

term statistically significant but instead use the term discovery. Let's control FDR for a gene list with $FDR = 0.05$. It means that of all the discoveries, 5% of them is expected to be false positives.

Benjamini & Hochberg (BH method) in 1995 proposed a way to control FDR: Let k be the largest i such that $p(i) \leq (i/m) * \alpha$, (m is the number of comparisons) then reject H_i for $i = 1, 2, \dots, k$

This process controls the FDR at level α . The method sets a different threshold p value for each comparison. Say we computed 100 t-tests, and got 100 p values, and we want to control the $FDR = 0.05$. We then rank the p values from small to big. if $p(1) \leq 1/100 * 0.05$, we then reject null hypothesis and accept the alternative. if $p(2) \leq 2/100 * 0.05$, we then reject the null and accept the alternative.

```
#remove the NAs
pvals<- pvals[!is.na(pvals)]

## order the pvals computed above and plot it.
alpha<- 0.05

#m is the number of comparisons
m<- length(pvals)

# let's arrange the p-value from small to big and get only the first 5000
top_5000_pvalue<- pval_df %>%
  dplyr::arrange(pvalue) %>%
  mutate(rank = row_number()) %>%
  dplyr::slice(1:5000)

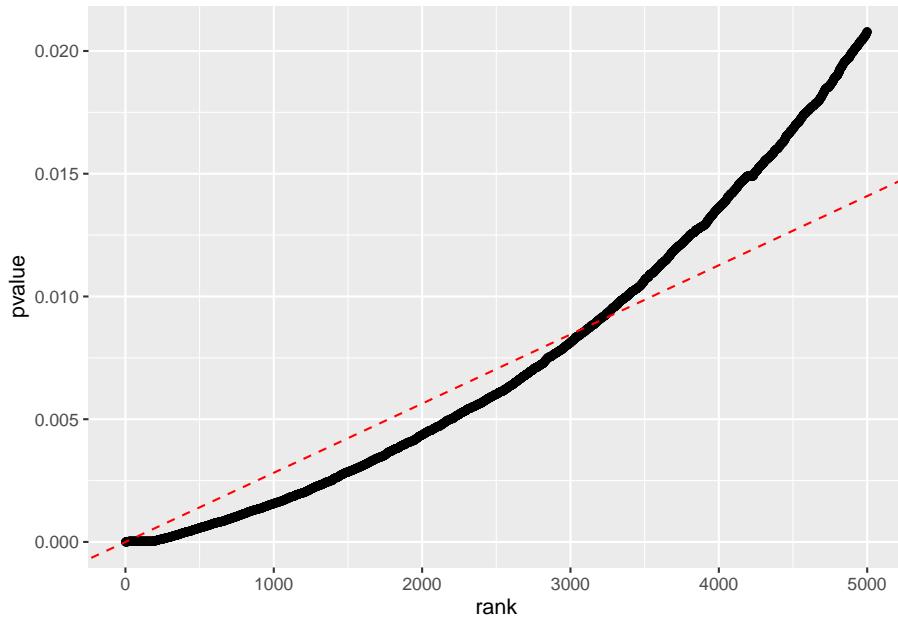
head(top_5000_pvalue)

## # A tibble: 6 x 3
##   gene      pvalue  rank
##   <chr>     <dbl> <int>
## 1 PPFIA4  0.000000907    1
## 2 GINS2   0.00000173     2
## 3 ZFP36L2 0.00000197     3
## 4 ECHS1   0.00000297     4
## 5 MOB3A   0.00000392     5
## 6 FLYWCH2 0.00000661     6
```

let's plot

260CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

```
ggplot(top_5000_pvalue, aes(x= rank, y = pvalue))+
  geom_point() +
  geom_abline(slope = alpha/m, intercept = 0, color = "red", linetype = 2)
```



p values that are below the red dotted line are controlled at FDR of 0.05.

We will use p.adjust function and the method “fdr” or “BH” to correct the p value, what the p.adjust function does is to recalculate the p-values.

With the FDR definition, p value is only significant if $p(i) \leq (i/m) * \alpha$. We can rewrite it to $p(i) * m/i \leq \alpha$. The p.adjust function returns $p(i) * m/i$ the adjusted p-value. We can then only accept the returned the p values if $p.adjust(pvals) \leq \alpha$.

```
top_5000_pvalue %>%
  mutate(padj = pvalue * m/rank) %>%
  head()
```

```
## # A tibble: 6 x 4
##   gene      pvalue  rank    padj
##   <chr>     <dbl> <int>   <dbl>
## 1 PPFIA4  0.000000907     1  0.0161
## 2 GINS2   0.00000173      2  0.0154
## 3 ZFP36L2 0.00000197      3  0.0116
## 4 ECHS1   0.00000297      4  0.0132
```

12.8. CORRECTING FOR MULTIPLE COMPARISONS IN STATISTICAL ANALYSIS 261

```
## 5 MOB3A 0.00000392      5 0.0139
## 6 FLYWCH2 0.00000661     6 0.0196
```

How many of those p-values are below the dotted red line?

```
top_5000_pvalue %>%
  mutate(padj = pvalue * m/rank) %>%
  filter(padj <= alpha) %>%
  filter(rank == which.max(rank))
```

```
## # A tibble: 1 x 4
##   gene      pvalue  rank    padj
##   <chr>     <dbl> <int>   <dbl>
## 1 RNASEH2A 0.00893 3173 0.0500
```

There are total 3173 p-values that are significant after FDR correction.

We can verify it using the p.adjust function in R:

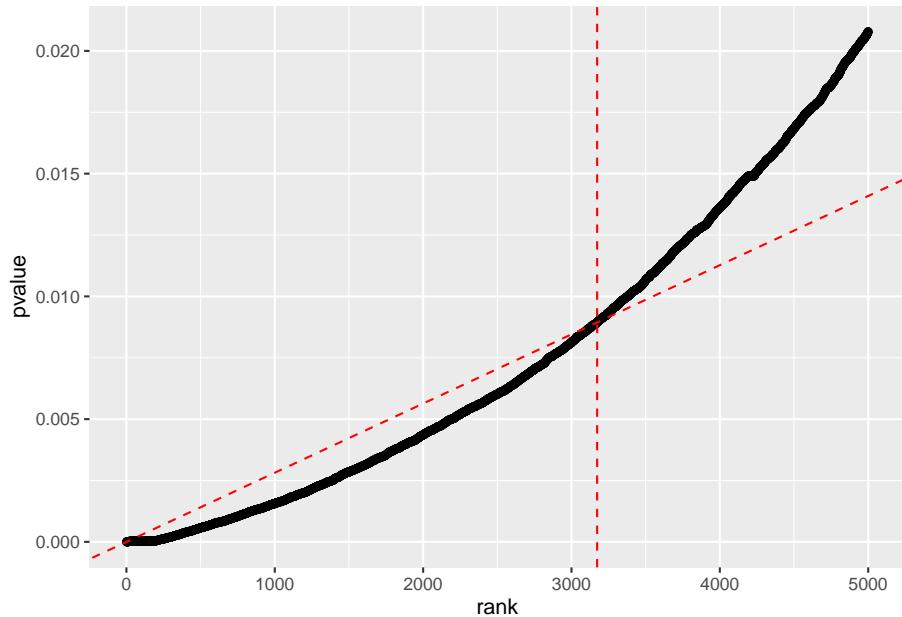
```
adjusted_pvalues <- p.adjust(pvals, method="fdr")
sum(adjusted_pvalues < 0.05)
```

```
## [1] 3173
```

This is the same as what we calculated manually!

We can plot a vertical line on the p-value ranking plot:

```
ggplot(top_5000_pvalue, aes(x= rank, y = pvalue)) +
  geom_point() +
  geom_abline(slope = alpha/m, intercept = 0, color = "red", linetype = 2) +
  geom_vline(xintercept = 3173, linetype = 2, color = "red")
```



12.8.4 Conclusion

Correcting for multiple comparisons is essential when conducting statistical tests on a large number of hypotheses. The False Discovery Rate (FDR) correction, such as the Benjamini & Hochberg method, allows us to control the rate of false discoveries while identifying significant results. This approach is valuable in various scientific disciplines to ensure the reliability of statistical findings.

12.9 Analyzing Differential Gene Expression with DESeq2

In this step, we will explore the DESeq2 workflow, a widely used bioinformatics package in R for identifying differentially expressed genes (DEGs) from RNA sequencing (RNA-seq) data. DESeq2 allows us to compare gene expression levels between different conditions or treatments and find genes that are significantly upregulated or downregulated.

DESeq2 is an R package within the Bioconductor project that performs differential expression analysis on RNA-seq data. It uses a negative binomial distribution to model read counts and estimates

12.9. ANALYZING DIFFERENTIAL GENE EXPRESSION WITH DESEQ2

the variance-mean dependence in the data to identify DEGs accurately. DESeq2 is particularly useful when dealing with count data from RNA-seq experiments. You can read docs here.

This Youtube video uses the same dataset and you may want to watch it if you prefer video.

12.9.1 Create a Sample Sheet

Before using DESeq2, it's essential to prepare a sample sheet that describes the experimental conditions of your samples. The sample sheet associates each sample with its respective experimental condition or treatment. This step is crucial for DESeq2 to understand the experimental design, as it enables the comparison of gene expression between conditions.

In our project, we will use a sample sheet with two conditions: "normoxia" and "hypoxia."

```
library(DESeq2)
coldata <- data.frame(condition = c("normoxia", "normoxia", "hypoxia", "hypoxia"))
rownames(coldata) <- colnames(raw_counts_mat)

coldata

##           condition
## 01_SW_sgCTRL_Norm  normoxia
## 02_SW_sgCTRL_Norm  normoxia
## 11_SW_sgCTRL_Hyp   hypoxia
## 12_SW_sgCTRL_Hyp   hypoxia
```

12.9.2 Create a DESeq2 Object

Next, we create a DESeq2 object using the count data and the sample sheet. This object will store the count data and associated sample information and will be used for differential expression analysis by DESeq2.

The design formula specifies the experimental design, taking the condition as the main factor.

```
dds <- DESeqDataSetFromMatrix(countData = raw_counts_mat,
                               colData = coldata,
                               design = ~ condition)

dds <- DESeq(dds)
```

12.9.3 Get Differential Results

To identify differentially expressed genes, we extract results from the DESeq2 analysis, specifying the contrast between two conditions (e.g., “hypoxia” vs. “normoxia”).

```
res <- results(dds, contrast = c("condition", "hypoxia", "normoxia"))

res

## log2 fold change (MLE): condition hypoxia vs normoxia
## Wald test p-value: condition hypoxia vs normoxia
## DataFrame with 23250 rows and 6 columns
##           baseMean log2FoldChange      lfcSE      stat     pvalue     padj
##           <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## DDX11L1       0.00000          NA        NA        NA        NA        NA
## WASH7P        23.25497       0.993523   0.588472   1.68831  0.0913516  0.1622577
## OR4F5         0.00000          NA        NA        NA        NA        NA
## LOC729737     8.93671       2.154987   0.975237   2.20970  0.0271257  0.0561441
## OR4F29        0.00000          NA        NA        NA        NA        NA
## ...           ...          ...        ...        ...        ...        ...
## APOBEC3A_B    0.000000         NA        NA        NA        NA        NA
## CCL3L1        0.228878       1.18653   4.99659   0.237469  0.812293   NA
## CCL4L1        0.000000         NA        NA        NA        NA        NA
## C4B_2         0.000000         NA        NA        NA        NA        NA
## HLA-DRB4      0.000000         NA        NA        NA        NA        NA
```

12.9.4 Explore the DESeq2 results and create visualizations.

A volcano plot is a commonly used visualization in RNA-seq analysis. It helps identify genes that are both statistically significant and biologically relevant.

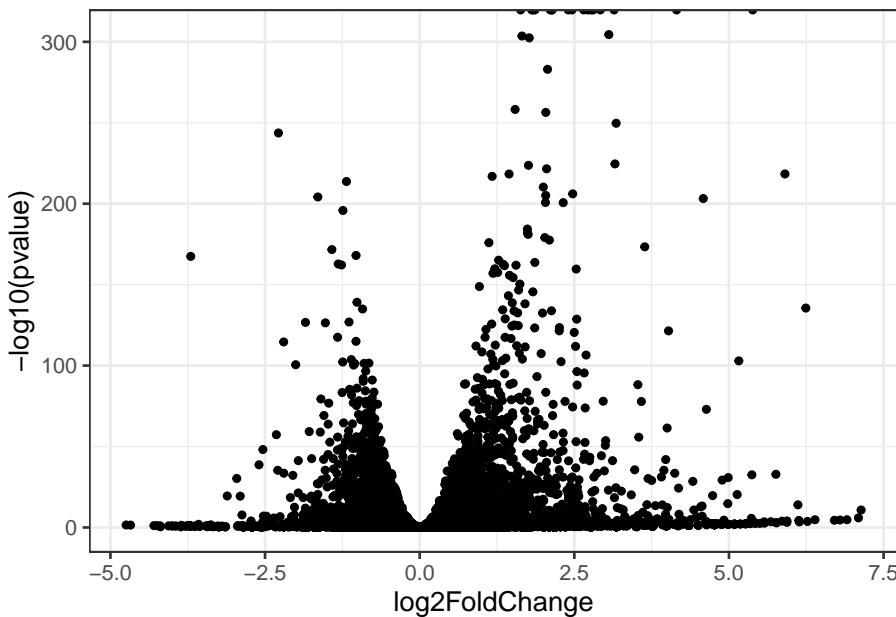
A volcano plot is a powerful visualization to simultaneously assess the significance and magnitude of gene expression changes. It helps identify genes that are both statistically significant and biologically relevant.

It's a scatter plot with the log2 fold change on the x-axis and the negative logarithm (base 10) of the p-value on the y-axis.

```
library(ggplot2)
```

12.9. ANALYZING DIFFERENTIAL GENE EXPRESSION WITH DESEQ2265

```
ggplot(data = as.data.frame(res)) +  
  geom_point(aes(x = log2FoldChange, y = -log10(pvalue))) +  
  theme_bw(base_size = 14)
```



In this plot, each point represents a gene. The x-axis shows how much a gene's expression changes (log2 fold change), while the y-axis indicates the significance of the change (-log10 p-value). Genes with a significant change will be far to the left or right on the plot, and those with very low p-values will be at the top.

12.9.5 Examining the Top Differentially Expressed Genes

Before creating visualizations, it's essential to understand which genes are the most differentially expressed. Examining the top genes by significance and fold change can provide insights into the dataset's characteristics.

```
top_differentially_expressed_genes <- res %>%  
  as.data.frame() %>%  
  arrange(padj, desc(log2FoldChange)) %>%  
  head(n = 30)
```

```
top_differentially_expressed_genes
```

##	baseMean	log2FoldChange	lfcSE	stat	pvalue
----	----------	----------------	-------	------	--------

266CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

## IGFBP3	22694.513	5.384140	0.05254478	102.46766	0.000000e+00
## ARRDC3	2211.253	4.153253	0.09299795	44.65962	0.000000e+00
## DDIT4	2912.023	3.143397	0.07280767	43.17398	0.000000e+00
## GPRC5A	2443.496	2.921410	0.07586026	38.51041	0.000000e+00
## STC2	9772.604	2.803328	0.05319908	52.69504	0.000000e+00
## ERRFI1	4315.414	2.773205	0.06711717	41.31886	0.000000e+00
## BNIP3L	8437.252	2.722813	0.04976713	54.71107	0.000000e+00
## LAMB3	3161.408	2.650830	0.06558866	40.41597	0.000000e+00
## LOXL2	3325.460	2.462177	0.06261047	39.32532	0.000000e+00
## BHLHE40	6352.428	2.402080	0.05114599	46.96517	0.000000e+00
## LBH	5294.026	2.139980	0.05538427	38.63877	0.000000e+00
## SERPINE1	25413.676	2.138890	0.05195800	41.16576	0.000000e+00
## NDRG1	27629.756	2.111619	0.04201763	50.25554	0.000000e+00
## FOSL2	7700.341	1.862466	0.04843765	38.45080	0.000000e+00
## LPCAT2	8017.173	1.829938	0.04637707	39.45781	0.000000e+00
## SLC2A3	61623.341	1.812379	0.03588353	50.50726	0.000000e+00
## PGK1	27104.315	1.631855	0.03996567	40.83142	0.000000e+00
## ADM	1965.214	3.058278	0.08189923	37.34196	3.424156e-305
## ITGA3	11970.535	1.653973	0.04435752	37.28732	2.634577e-304
## MYOF	7676.155	1.771628	0.04760186	37.21762	3.540676e-303
## ITGA5	3934.470	2.067309	0.05743290	35.99520	9.945951e-284
## PPME1	8988.338	1.543067	0.04488883	34.37530	5.899480e-259
## CALCOCO1	3722.413	2.038787	0.05952092	34.25328	3.897346e-257
## EGLN3	1866.247	3.176719	0.09398826	33.79910	2.033171e-250
## CHI3L1	2520.051	-2.283507	0.06839506	-33.38702	2.115471e-244
## CAV1	1328.035	3.157403	0.09853096	32.04478	2.595697e-225
## PPL	5609.751	1.758699	0.05499130	31.98141	1.977676e-224
## TIMP3	3279.084	2.051831	0.06447047	31.82590	2.837566e-222
## SIGLEC6	1189.280	5.905399	0.18690484	31.59575	4.222688e-219
## ITGB4	9662.811	1.445206	0.04574502	31.59263	4.661613e-219
		padj			
## IGFBP3		0.000000e+00			
## ARRDC3		0.000000e+00			
## DDIT4		0.000000e+00			
## GPRC5A		0.000000e+00			
## STC2		0.000000e+00			
## ERRFI1		0.000000e+00			
## BNIP3L		0.000000e+00			
## LAMB3		0.000000e+00			
## LOXL2		0.000000e+00			
## BHLHE40		0.000000e+00			
## LBH		0.000000e+00			
## SERPINE1		0.000000e+00			
## NDRG1		0.000000e+00			
## FOSL2		0.000000e+00			
## LPCAT2		0.000000e+00			

12.9. ANALYZING DIFFERENTIAL GENE EXPRESSION WITH DESEQ2267

```
## SLC2A3      0.000000e+00
## PGK1       0.000000e+00
## ADM        2.860883e-302
## ITGA3      2.085337e-301
## MYOF       2.662411e-300
## ITGA5      7.122722e-281
## PPME1      4.032831e-256
## CALCOCO1   2.548356e-254
## EGLN3       1.274036e-247
## CHI3L1     1.272583e-241
## CAV1        1.501411e-222
## PPL         1.101565e-221
## TIMP3       1.524077e-219
## SIGLEC6    2.189828e-216
## ITGB4      2.336867e-216
```

The code above arranges genes in descending order of adjusted p-value (padj) and then by log2 fold change. It retrieves the top 30 genes with the most significant differences in expression between the conditions. You may also notice that the top several genes are with p-value of 0, that's why you see the dots capped in the volcano plot.

You can use this list to focus further analysis or exploration on the genes with the most substantial changes.

12.9.6 Labeling Genes in the Volcano Plot

To identify and label genes of specific interest on the volcano plot, we filter genes based on criteria such as fold change and p-value.

To label genes of interest on the volcano plot, we first define criteria to identify them. In this example, we filter genes with an absolute log2 fold change greater than or equal to 2.5 and a p-value less than or equal to 0.001. We also exclude genes with names containing “LOC.”

```
genes_to_label <- res %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var = "gene") %>%
  filter(!stringr::str_detect(gene, "LOC")),
  abs(log2FoldChange) >= 2.5,
  padj <= 0.001)

head(genes_to_label)
```

##	gene	baseMean	log2FoldChange	lfcSE	stat	pvalue
1	SLC2A3	0.000000e+00	2.860883e-302			
2	PGK1	0.000000e+00				
3	ADM		2.085337e-301			
4	ITGA3		2.662411e-300			
5	MYOF		7.122722e-281			

```

## 1 ERRFI1 4315.41443      2.773205 0.06711717 41.318865 0.000000e+00
## 2 NPPB   25.25513       4.474210 0.86388658 5.179164 2.228827e-07
## 3 CTRC   13.68629       4.629341 1.21487165 3.810560 1.386526e-04
## 4 PADI2  46.96747       2.852354 0.46635589 6.116260 9.579672e-10
## 5 GJB4   88.58994       2.797718 0.33766499 8.285485 1.176284e-16
## 6 EDN2   366.47734      3.583731 0.19100291 18.762705 1.524399e-78
##
##      padj
## 1 0.000000e+00
## 2 1.066136e-06
## 3 4.506584e-04
## 4 5.721552e-09
## 5 1.227629e-15
## 6 1.302582e-76

```

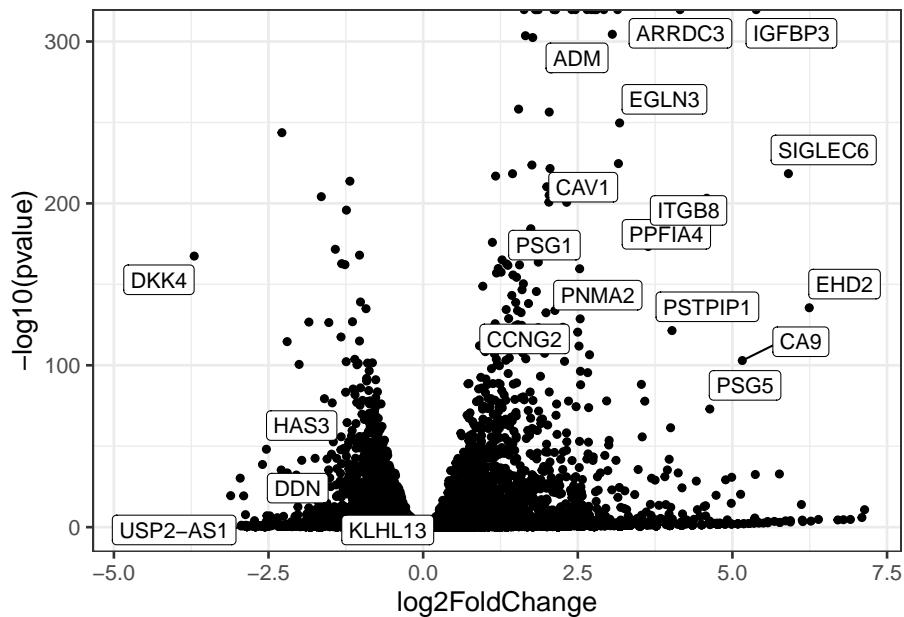
Now, we label these selected genes on the volcano plot using the `ggrepel` package to prevent label overlap.

```

library(ggrepel)

ggplot(data = as.data.frame(res), aes(x = log2FoldChange, y = -log10(pvalue))) +
  geom_point() +
  ggrepel::geom_label_repel(data = genes_to_label, aes(label = gene)) +
  theme_bw(base_size = 14)

```



12.9. ANALYZING DIFFERENTIAL GENE EXPRESSION WITH DESEQ2269

12.9.7 Coloring Points in the Volcano Plot

To further enhance the plot, we can color the points based on significance. Coloring points in the volcano plot based on significance provides additional information. It helps distinguish genes that are statistically significant from those that are not.

In this example, we color genes as “sig” (significant) or “not sig” (not significant) based on the criteria used for labeling.

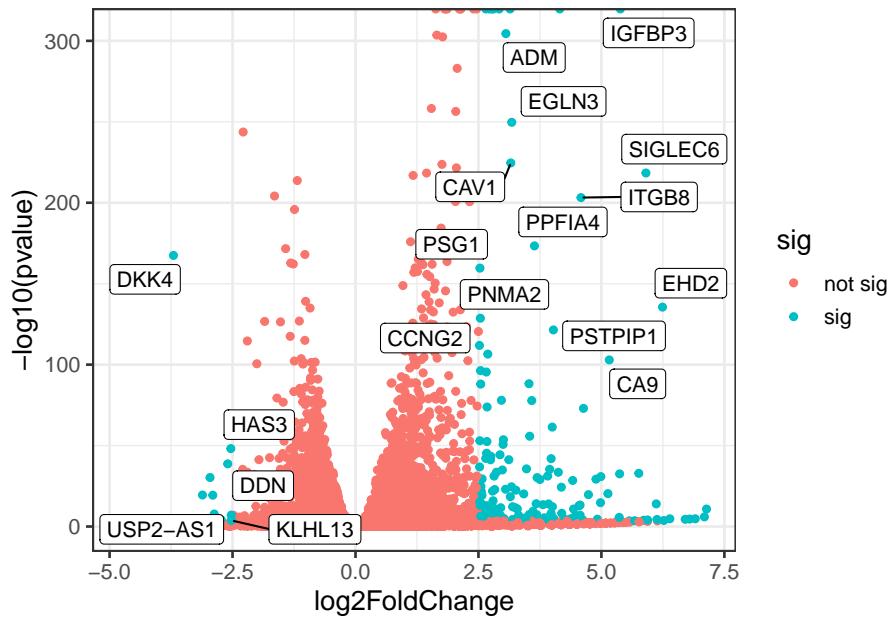
```
res2 <- res %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var = "gene") %>%
  mutate(sig = case_when(
    !stringr::str_detect(gene, "LOC") &
      abs(log2FoldChange) >= 2.5 &
      padj <= 0.001 ~ "sig",
    TRUE ~ "not sig"
  ))
head(res2)

##      gene  baseMean log2FoldChange      lfcSE      stat     pvalue     padj
## 1 DDX11L1 0.0000000          NA          NA          NA          NA          NA
## 2 WASH7P 23.2549734 0.9935235 0.5884719 1.6883109 0.09135156 0.16225771
## 3 OR4F5 0.0000000          NA          NA          NA          NA          NA
## 4 LOC729737 8.9367086 2.1549867 0.9752375 2.2097046 0.02712567 0.05614409
## 5 OR4F29 0.0000000          NA          NA          NA          NA          NA
## 6 OR4F16 0.9290025 1.3683069 2.9857917 0.4582727 0.64675651          NA
##
##      sig
## 1 not sig
## 2 not sig
## 3 not sig
## 4 not sig
## 5 not sig
## 6 not sig
```

Now, we update the volcano plot with colored points, horizontal and vertical lines, and labeled genes.

```
ggplot(res2, aes(x = log2FoldChange, y = -log10(pvalue))) +
  geom_point(aes(color = sig)) +
  ggrepel::geom_label_repel(data = genes_to_label, aes(label = gene))+
  theme_bw(base_size = 14)
```

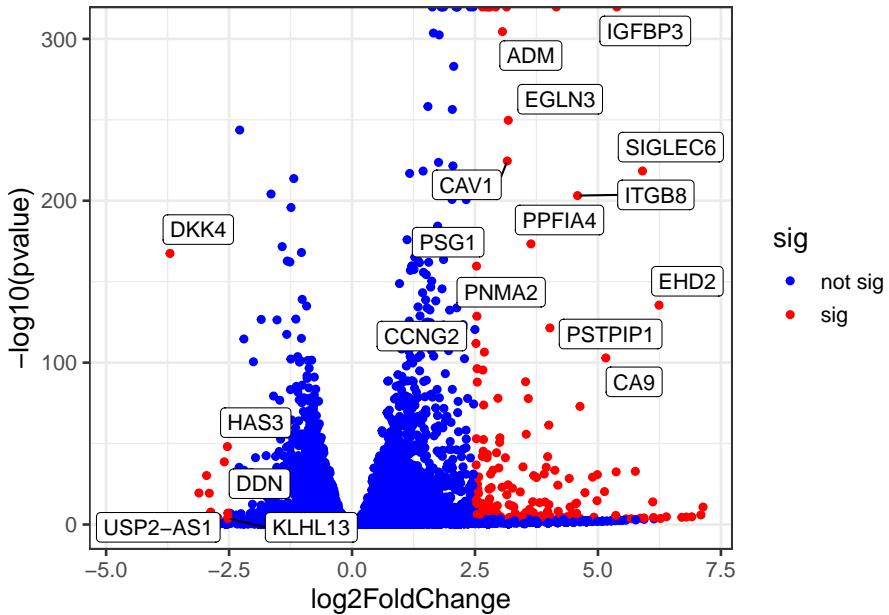
270CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO



Now, let's fix the color of the points and let's use red for the significant ones

```
ggplot(res2, aes(x = log2FoldChange, y = -log10(pvalue))) +
  geom_point(aes(color = sig)) +
  scale_color_manual(values = c("blue", "red")) +
  ggrepel::geom_label_repel(data = genes_to_label, aes(label = gene)) +
  theme_bw(base_size = 14)
```

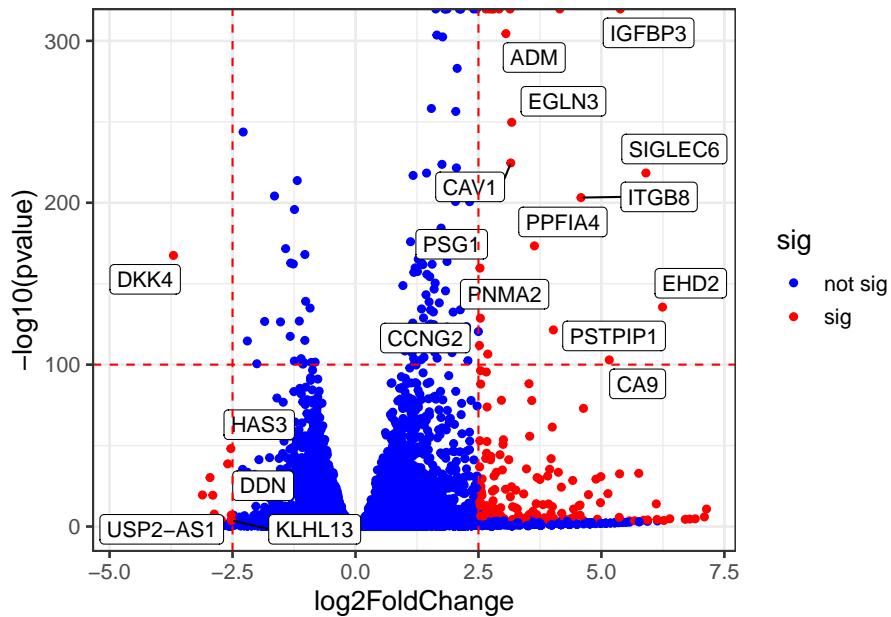
12.9. ANALYZING DIFFERENTIAL GENE EXPRESSION WITH DESEQ2



Finally, let's add horizontal and vertical lines.

1. Vertical Lines: Vertical lines are often added at specific fold change values, such as -2.5 and 2.5 in the provided code. These lines represent the threshold for fold change, indicating the point beyond which genes are considered significantly upregulated (right of the rightmost line) or significantly downregulated (left of the leftmost line).
2. Horizontal Line: The horizontal line, we choose at $-\log_{10}(p\text{value}) = 100$ or another chosen significance level that fits your own data, emphasizing the threshold for statistical significance. Genes falling above this line are considered highly significant based on their p-values.

```
ggplot(res2, aes(x = log2FoldChange, y = -log10(pvalue))) +
  geom_point(aes(color = sig)) +
  scale_color_manual(values = c("blue", "red")) +
  ggrepel::geom_label_repel(data = genes_to_label, aes(label = gene)) +
  geom_hline(yintercept = 100, linetype = 2, color = "red") +
  geom_vline(xintercept = c(-2.5, 2.5), linetype = 2, color = "red") +
  theme_bw(base_size = 14)
```



This final plot provides a clear visualization of DEGs, with significant genes highlighted in red and labeled for easy identification.

12.9.8 Conclusion

In this lesson, we've explored the DESeq2 workflow for differential gene expression analysis in RNA-seq data. We've learned how to create a sample sheet, perform differential analysis, and visualize the results using volcano plots. Additionally, we've discussed criteria for labeling and coloring genes on the plot, making it a powerful tool for identifying biologically relevant DEGs in real-world datasets.

12.10 Principal Component Analysis (PCA) using DESeq2

In this lesson, we will explore Principal Component Analysis (PCA), a dimensionality reduction technique commonly used in genomics to analyze high-dimensional gene expression data. We will use R and the DESeq2 package to perform PCA analysis and understand the steps involved.

12.10.1 What is Principal Component Analysis (PCA)?

Principal Component Analysis, or PCA, is a mathematical technique used to reduce the dimensionality of high-dimensional data while retaining the most important information. In genomics, PCA is often used to visualize and explore variations in gene expression data across different samples or conditions.

12.10.2 Why use PCA?

1. Dimension Reduction: Genomic data often contain thousands of genes, making it challenging to visualize or analyze. PCA helps reduce this complexity by summarizing the data into a few principal components.
2. Visualization: PCA allows us to visualize the similarities and differences between samples or conditions in a lower-dimensional space, making it easier to detect patterns or clusters.
3. Identification of Outliers: PCA can help identify outlier samples that deviate significantly from the majority of samples. PCA usually is my first step in exploratory data analysis to spot the wiredness of the data.

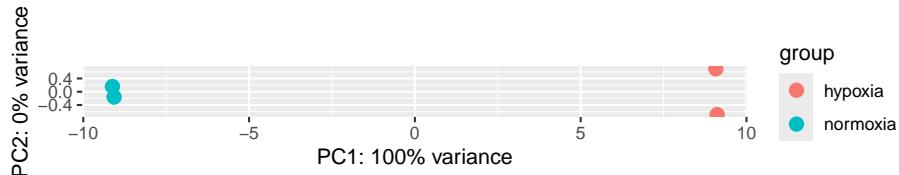
12.10.3 Using DESeq2 for PCA Analysis

DESeq2 includes a function called `plotPCA` for performing PCA analysis. However, in some cases, you may want to perform PCA manually for more customization. Let's walk through the steps.

DESeq2 has a `plotPCA` function to use:

```
#vsd is the normalized data
vsd <- vst(dds, blind=FALSE)

plotPCA(vsd, intgroup=c("condition"))
```



- `dds` is assumed to be a DESeqDataSet object containing raw gene expression data.
- `vst` stands for Variance Stabilizing Transformation, which is used to normalize the gene expression data. This transformation stabilizes the variance across samples, making the data suitable for PCA analysis.
- The `blind=FALSE` argument indicates that the design of the experiment is not blinded.
- Each point on the PCA plot represents a sample (e.g., a biological sample from an experiment).
- The position of each point on the plot is determined by its scores along the principal components (PC1 and PC2).
- The `intgroup` argument allows you to color the points based on a specific grouping variable (in this case, “condition”), which helps visualize how different conditions relate to each other in terms of gene expression patterns.

The resulting PCA plot provides insights into the underlying structure of the data. It clearly shows that the samples are separated in PC1 by the condition of hypoxia vs normoxia. PCA can be valuable for quality control, identifying experimental effects, or exploring the relationships between conditions in genomics research.

12.10.4 Plotting PCA by ourselves

Before performing Principal Component Analysis (PCA), it's essential to prepare our gene expression data properly. In genomics, data often require normalization to account for variations introduced during the experimental process. We are using DESeq2's Variance Stabilizing Transformation (VST) to normalize the data.

When conducting RNA-Seq experiments, variations can arise due to differences in sequencing depths, library sizes, and other technical factors. The VST helps to stabilize the variance across samples and makes the data more suitable for downstream analysis. By applying VST to our DESeqDataSet object named dds, you ensure that the data is in a suitable format for PCA, where the goal is to capture biological variations rather than technical noise.

12.10.4.1 Calculate Principal Components

Principal components are mathematical constructs that represent the major sources of variation in your data. Calculating principal components is a critical step in PCA.

In this step, we'll use the `prcomp` function to compute the principal components from the normalized gene expression data. The function takes the transpose of the normalized counts as input because PCA is typically performed on columns (samples) rather than rows (genes).

By calculating these principal components, we are summarizing the data in a way that retains the most significant sources of variation across all genes and samples. This is essential because genes that vary together might provide insights into shared biological processes or conditions.

```
# Get the normalized counts
normalized_counts <- assay(vsd) %>% as.matrix()

# Calculate the principal components
pca_prcomp <- prcomp(t(normalized_counts), center = TRUE, scale. = FALSE)

names(pca_prcomp)

## [1] "sdev"      "rotation"   "center"     "scale"      "x"
```

In PCA, each principal component is a linear combination of the original variables (in this case, genes). The coefficients of this linear combination are called loadings. Loadings represent the contribution of each original variable (gene) to the principal component. Let's retrieve them:

```
# the $x contains the PC loadings we want
pca_prcomp$x
```

	PC1	PC2	PC3	PC4
## 01_SW_sgCTRL_Norm	-13.25583	1.302834	-2.021424	7.071325e-14
## 02_SW_sgCTRL_Norm	-13.20037	-1.312503	2.022705	7.433980e-14
## 11_SW_sgCTRL_Hyp	13.26493	-2.483543	-1.063259	7.233500e-14
## 12_SW_sgCTRL_Hyp	13.19128	2.493211	1.061978	7.091842e-14

12.10.5 Create a DataFrame for Visualization

To visualize the results of PCA, we need to organize the principal component scores (PC1 and PC2) along with sample labels in a DataFrame.

This organization allows us to create a PCA plot where each point represents a sample, and we can color-code the points based on the sample labels. Also, this visualization helps us understand how samples relate to each other in a lower-dimensional space, where the primary sources of variation are captured by PC1 and PC2.

```
# Create a DataFrame for visualization
PC1_and_PC2 <- data.frame(PC1 = pca_prcomp$x[, 1], PC2 = pca_prcomp$x[, 2], type = row
```

12.10.6 Plot the PCA

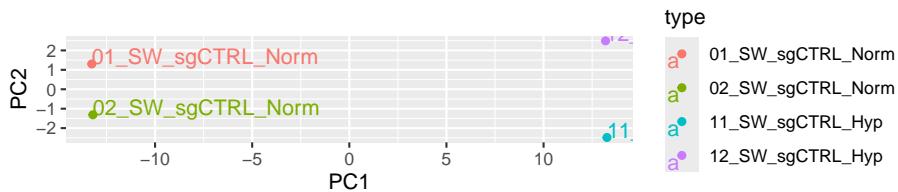
The final step is to create a PCA plot using the `ggplot2` package. This plot visually represents the relationships between samples in a reduced-dimensional space.

Visualization is a crucial aspect of PCA. By plotting PC1 against PC2, we visualize how samples cluster or spread apart based on their gene expression profiles. The `ggplot2` package provides a powerful and flexible way to create such plots. Each point in the plot corresponds to a sample, and the color of the points indicates the sample's label or condition. This visual representation allows us to identify patterns, clusters, or outliers in our data, aiding in the interpretation of the biological significance of the variation observed.

```
# Load ggplot2 library
library(ggplot2)

# Create a PCA plot
ggplot(PC1_and_PC2, aes(x = PC1, y = PC2, col = type)) +
  geom_point() +
```

```
geom_text(aes(label = type), hjust = 0, vjust = 0) +
coord_fixed()
```



12.10.7 Comparison with DESeq2's plotPCA

You may notice that the PCA plot we created manually is not identical to the one generated by DESeq2's `plotPCA` function. This difference arises because DESeq2's function uses a default set of genes (the top 500 most variable genes) for the analysis.

The choice of genes can significantly impact the results of PCA. DESeq2's default behavior is to focus on the genes with the highest variation across samples, as these are often the most informative for distinguishing between conditions. However, in some cases, you may want to use all genes or a specific subset based on your research question. Customizing the gene selection can provide different perspectives on the data and help you address specific hypotheses or explore different aspects of gene expression variation.

12.10.8 Conclusion

In this lesson, we've delved into the steps of performing Principal Component Analysis (PCA) in genomics using the DESeq2 package. PCA is a valuable technique for exploring complex gene expression data and understanding each

step in the process is essential for meaningful analysis and interpretation. By following these steps and considering the context and choices made during the analysis, you can gain deeper insights into your genomic data and draw biologically relevant conclusions.

12.11 Creating a Perfect Heatmap

Heatmaps are particularly useful in genomics and other scientific fields where large datasets need to be analyzed. We'll cover the importance of selecting significant genes, scaling data, and annotating heatmaps to enhance their interpretability.

Watch this youtube video to better your understanding.

12.11.1 Selecting Significant Genes

Our journey begins by identifying the genes that exhibit significant changes in expression. We define significant genes as those with an adjusted p-value (padj) less than or equal to 0.01 and an absolute log2 fold change (log2FoldChange) greater than or equal to 2. These criteria help us focus on genes that are most likely to be biologically relevant.

```
library(ComplexHeatmap)

significant_genes <- res %>%
  as.data.frame() %>%
  filter(padj <= 0.01, abs(log2FoldChange) >= 2) %>%
  rownames()

head(significant_genes, 10)

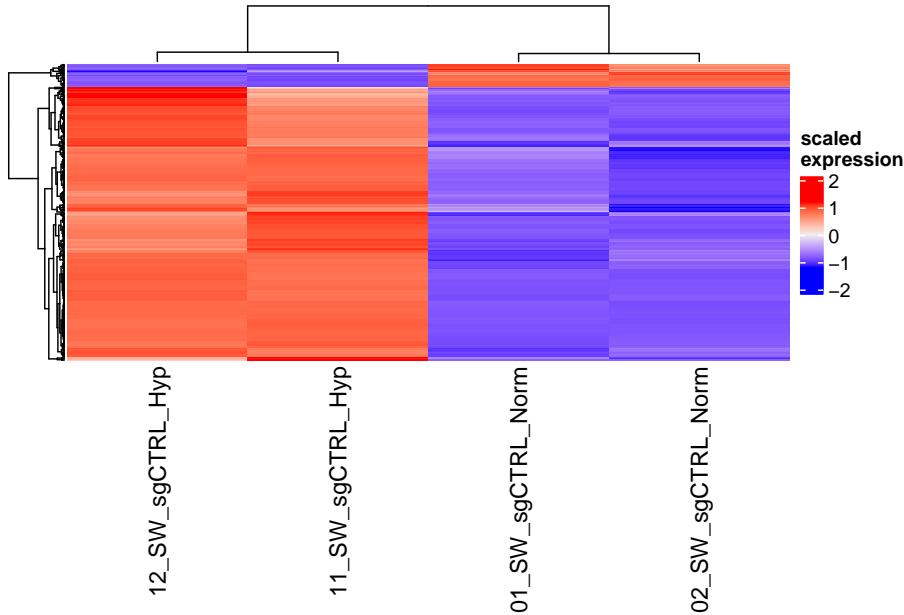
## [1] "HES2"    "ERRFI1"   "NPPB"     "CTRC"     "ESPNP"    "PADI2"    "CDA"      "CSMD2"
## [9] "GJB4"    "COL8A2"
```

The `significant_genes` variable now contains the names of genes meeting these criteria.

12.11.2 Creating the Heatmap

Next, we will create a heatmap using the selected significant genes. To ensure the heatmap is informative, we scale the data using the `scale` function. Scaling transforms the values so that they have a mean of 0 and a standard deviation of 1, making it easier to visualize relative differences.

```
significant_mat <- normalized_counts[significant_genes, ]  
  
Heatmap(  
  t(scale(t(significant_mat))),  
  show_row_names = FALSE,  
  name = "scaled\\nexpression"  
)
```



By scaling the data, we obtain a heatmap with a legend ranging from -2 to 2, representing z-scores after scaling. This scaling helps us see the patterns in gene expression more clearly.

12.11.3 Why Scaling is Important?

Scaling is essential because it helps in comparing genes with different expression ranges. Without scaling, genes with larger absolute expression values may dominate the visualization, making it challenging to discern subtle patterns. Scaling levels the playing field, allowing us to focus on the relative changes in gene expression.

12.11.4 Adding Annotations

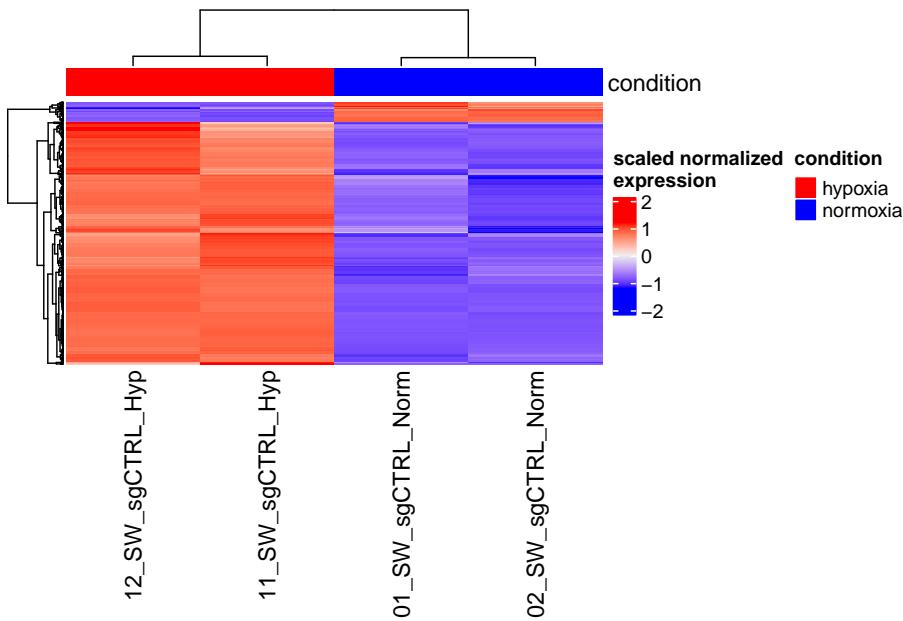
Annotations provide valuable context to our heatmap. In this example, we have annotations for different experimental conditions (e.g., “normoxia” and “hypoxia”). These annotations help us understand the conditions under which the gene expression data was collected.

```
coldata
```

```
##                  condition
## 01_SW_sgCTRL_Norm  normoxia
## 02_SW_sgCTRL_Norm  normoxia
## 11_SW_sgCTRL_Hyp   hypoxia
## 12_SW_sgCTRL_Hyp   hypoxia

col_anno <- HeatmapAnnotation(
  df = coldata,
  col = list(condition = c("hypoxia" = "red", "normoxia" = "blue"))
)

Heatmap(
  t(scale(t(significant_mat))),
  top_annotation = col_anno,
  show_row_names = FALSE,
  name = "scaled normalized\ncexpression"
)
```



Now, our heatmap includes color annotations at the top, with “hypoxia” conditions shown in red and “normoxia” conditions in blue.

Note, we used a named vector to denote the color for each condition:

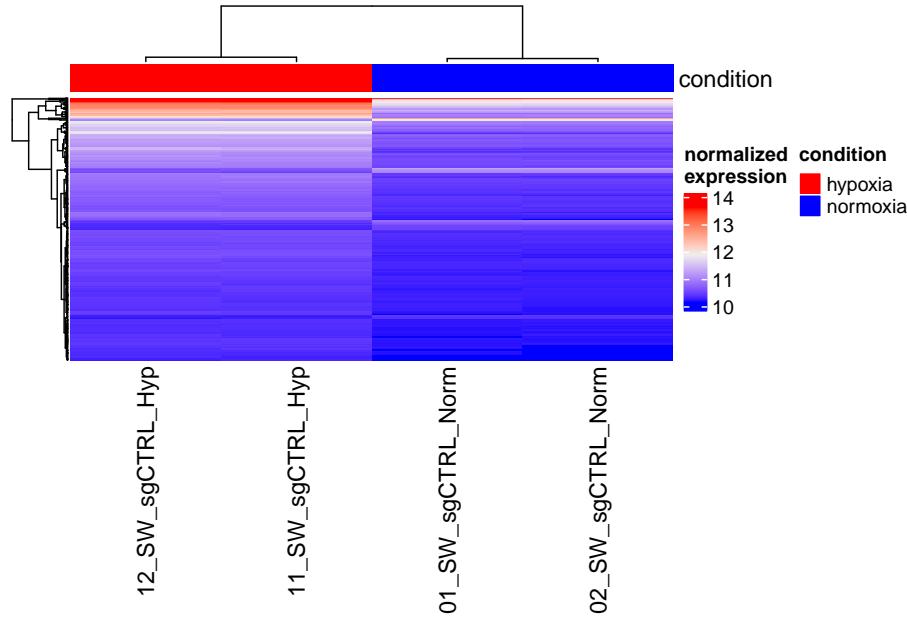
```
c("hypoxia" = "red", "normoxia" = "blue")
```

```
## hypoxia normoxia
##     "red"    "blue"
```

12.11.5 The Impact of Scaling

To highlight the difference that scaling makes, we can compare our scaled heatmap with one that doesn’t use scaling. The legend of the unscaled heatmap displays the normalized expression values directly.

```
Heatmap(
  significant_mat,
  top_annotation = col_anno,
  show_row_names = FALSE,
  name = "normalized\ncexpression"
)
```



The key takeaway here is that without scaling, we may not be able to discern patterns as easily, especially when dealing with genes that have vastly different expression levels.

12.11.6 Conclusion

In this lesson, we've explored the process of creating a perfect heatmap by selecting significant genes, scaling the data, and adding annotations. It is “perfect” because we pre-selected those genes and we are expected to see the pattern shown in the heatmap.

Scaling is crucial for visualizing relative changes, and annotations provide context to help us interpret the heatmap effectively. Heatmaps are powerful tools for identifying trends and patterns within complex datasets, making them invaluable in fields like genomics and beyond. I highly recommend you to read the tutorial <https://jokergoo.github.io/ComplexHeatmap-reference/book/>. It is very comprehensive and we only covered a small number of features.

12.12 Pathway Analysis Using Over-Representation and Gene Set Enrichment Analysis

In this lesson, we will delve into pathway analysis, a crucial step in the field of bioinformatics. Pathway analysis helps us understand the biological func-

12.12. PATHWAY ANALYSIS USING OVER-REPRESENTATION AND GENE SET ENRICHMENT ANALYSIS2

tions and processes associated with a set of genes or proteins. We will explore two widely used methods: Over-Representation Analysis (ORA) and Gene Set Enrichment Analysis (GSEA).

Watch this YouTube video if you prefer:

12.12.1 Understanding Pathway Analysis

Pathway analysis is used to identify biological pathways or Gene Ontology (GO) terms that are significantly enriched in a set of genes. It helps us interpret the functional significance of a group of genes and can reveal insights into the underlying biology of a particular condition or experiment.

For further understanding go here: <https://yulab-smu.top/biomedical-knowledge-mining-book/enrichment-overview.html>

12.12.2 Over-Representation Analysis (ORA)

Over-Representation Analysis (ORA) is a method that compares the list of genes of interest (e.g., DEGs) to a background set (all genes in the experiment) to identify pathways that are over-represented among the genes of interest.

Let's break down the steps for ORA:

1. Conversion of Gene Symbols: Genes are often represented by symbols. We convert these symbols to Entrez IDs, which are unique identifiers for genes, using the clusterProfiler package.
2. Selecting Background Genes: We filter out genes with zero expression (`baseMean = 0`) to create a list of background genes that were detected in the experiment.
3. Enrichment Analysis: Using the Gene Ontology (GO) database, we perform enrichment analysis for specific categories, such as Biological Processes (BP), Cellular Components (CC), or Molecular Functions (MF). For example, we focus on BP in this code.
4. Statistical Analysis: We apply statistical tests and corrections (e.g., Benjamini-Hochberg adjustment) to identify significantly enriched pathways.
5. Visualization: We visualize the results using bar plots and dot plots to highlight the most enriched pathways.

In many biological analyses, researchers work with gene symbols, which are human-readable names for genes. However, for more standardized and precise

analysis, it is often necessary to convert these gene symbols to Entrez IDs. Entrez IDs are unique numeric identifiers assigned to genes in the National Center for Biotechnology Information (NCBI) Entrez database, providing a consistent way to refer to genes across various databases and tools.

```
library(clusterProfiler)

#convert gene symbol to Entrez ID

significant_genes_map<- clusterProfiler::bitr(geneID = significant_genes,
                                              fromType="SYMBOL", toType="ENTREZID",
                                              OrgDb="org.Hs.eg.db")

head(significant_genes_map)

##      SYMBOL ENTREZID
## 1    HES2     54626
## 2  ERRFI1     54206
## 3   NPPB      4879
## 4    CTRC     11330
## 5  ESPNP     284729
## 6   PADI2     11240
```

In many RNA sequencing (RNA-seq) experiments, researchers aim to identify genes that are differentially expressed or have some other relevant feature compared to a control or reference group. However, it's crucial to consider all the genes that were detected in the experiment, not just the differentially expressed ones. These detected genes are often referred to as "background genes."

```
## background genes are genes that are detected in the RNAseq experiment
background_genes<- res %>%
  as.data.frame() %>%
  filter(baseMean != 0) %>%
  tibble::rownames_to_column(var = "gene") %>%
  pull(gene)

res_df<- res %>%
  as.data.frame() %>%
  filter(baseMean != 0) %>%
  tibble::rownames_to_column(var = "gene")

background_genes_map<- bitr(geneID = background_genes,
                             fromType="SYMBOL",
                             toType="ENTREZID",
```

12.12. PATHWAY ANALYSIS USING OVER-REPRESENTATION AND GENE SET ENRICHMENT ANALYSIS2

```
OrgDb="org.Hs.eg.db")  
  
head(background_genes_map)  
  
##      SYMBOL ENTREZID  
## 1     WASH7P    653635  
## 2     LOC729737  729737  
## 3     OR4F16    81399  
## 4 LOC100288069 100288069  
## 5     LINC00115  79854  
## 6     LINC01128  643837
```

The resulting `background_genes_map` object contains a mapping between gene symbols and their corresponding Entrez IDs for the background genes. This mapping is essential for subsequent gene ontology enrichment analysis to ensure that the analysis considers standardized gene identifiers.

12.12.3 Some concepts to remember

Gene Ontology(GO) defines concepts/classes used to describe gene function and relationships between these concepts. It classifies functions along three aspects:

- MF: Molecular Function molecular activities of gene products
- CC: Cellular Component where gene products are active
- BP: Biological Process pathways and larger processes made up of the activities of multiple gene products

GO terms are organized in a directed acyclic graph, where edges between terms represent parent-child relationship.

Now let's perform a Gene Ontology (GO) enrichment analysis using the `enrichGO` function from the `clusterProfiler` package. GO enrichment analysis is a widely used technique in bioinformatics to determine whether a set of genes is overrepresented in specific functional categories, such as biological processes (BP), cellular components (CC), or molecular functions (MF).

```
ego <- enrichGO(gene      = significant_genes_map$ENTREZID,  
                  universe   = background_genes_map$ENTREZID,  
                  OrgDb     = org.Hs.eg.db,  
                  ont       = "BP",  
                  pAdjustMethod = "BH",  
                  pvalueCutoff  = 0.01,
```

286CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

```

qvalueCutoff = 0.05,
readable     = TRUE)
head(ego)

##           ID          Description GeneRatio
## GO:0001525 GO:0001525      angiogenesis 44/316
## GO:0045765 GO:0045765 regulation of angiogenesis 26/316
## GO:1901342 GO:1901342 regulation of vasculature development 26/316
## GO:0045766 GO:0045766 positive regulation of angiogenesis 19/316
## GO:1904018 GO:1904018 positive regulation of vasculature development 19/316
## GO:0070482 GO:0070482 response to oxygen levels 27/316
##          BgRatio      pvalue    p.adjust   qvalue
## GO:0001525 456/14508 6.573666e-17 2.350743e-13 1.907055e-13
## GO:0045765 243/14508 2.070707e-11 3.570942e-08 2.896950e-08
## GO:1901342 247/14508 2.995757e-11 3.570942e-08 2.896950e-08
## GO:0045766 141/14508 2.212513e-10 1.582389e-07 1.283723e-07
## GO:1904018 141/14508 2.212513e-10 1.582389e-07 1.283723e-07
## GO:0070482 296/14508 3.289495e-10 1.960539e-07 1.590500e-07
##
## GO:0001525 NPPB/COL8A2/CCN1/F3/S1PR1/SEMA4A/SH2D2A/CHI3L1/CD34/TGFB2/COL4A3/ACKR3/CD
## GO:0045765
## GO:1901342
## GO:0045766
## GO:1904018
## GO:0070482
##          Count
## GO:0001525    44
## GO:0045765    26
## GO:1901342    26
## GO:0045766    19
## GO:1904018    19
## GO:0070482    27

```

The code you see is used to find significantly enriched biological processes (BP) among a list of significant genes (`significant_genes_map$ENTREZID`) compared to a background set of genes (`background_genes_map$ENTREZID`). Here's what each parameter in the `enrichGO` function does:

- gene: This is the list of significant genes that you want to analyze.
- universe: The universe of genes against which you want to test for enrichment (in this case, the background genes).
- OrgDb: Specifies the organism-specific gene annotation database (in this example, "org.Hs.eg.db" for Homo sapiens).

12.12. PATHWAY ANALYSIS USING OVER-REPRESENTATION AND GENE SET ENRICHMENT ANALYSIS2

- `ont`: Specifies the ontology to use (e.g., “BP” for Biological Process).
- `pAdjustMethod`: The method used to adjust p-values for multiple testing (e.g., “BH” for Benjamini & Hochberg correction).
- `pvalueCutoff`: The significance threshold for p-values.
- `qvalueCutoff`: The significance threshold for q-values, which are adjusted p-values.
- `readable`: A logical value indicating whether to include readable gene names.

Here, you can see the top enriched GO terms related to biological processes, along with their respective statistics.

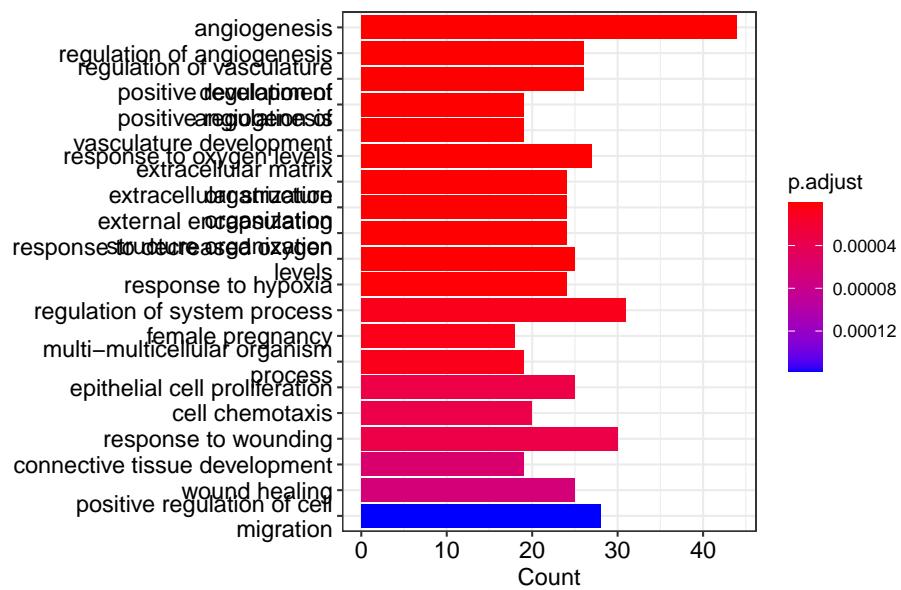
It is reassuring you see angiogenesis (hypoxia induces angiogenesis) and response to oxygen levels are the top GO terms that are enriched.

This analysis helps researchers understand which biological processes are over-represented in their list of significant genes, providing valuable insights into the underlying biology of the studied genes.

12.12.4 Visualizations

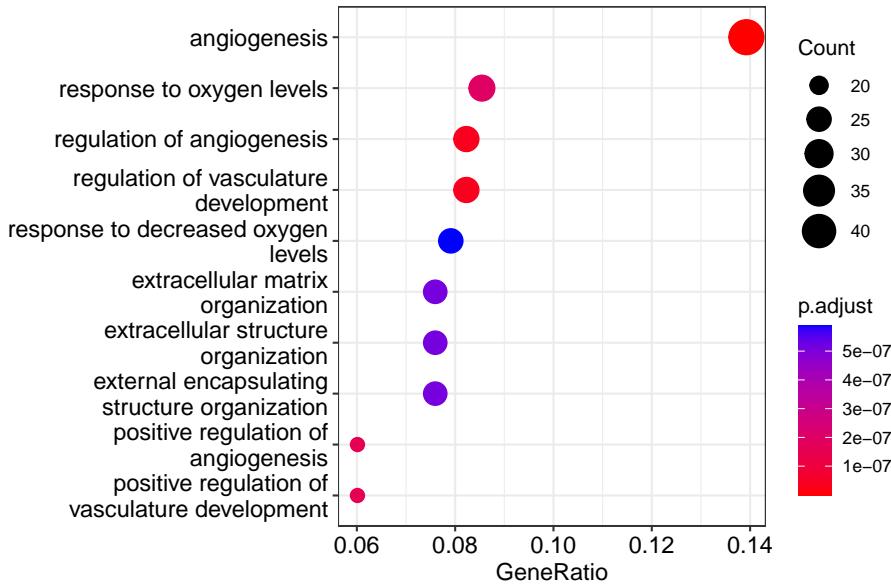
Now, let's plot a barplot to visualize the enriched GO terms. The `ego` object contains the enrichment results generated previously. Each bar in the plot represents a GO term, and the height of the bar indicates the significance or enrichment level of that term. The `showCategory` parameter specifies how many top categories you want to display in the plot. In this case, we are displaying the top 20 enriched GO terms related to biological processes.

```
barplot(ego, showCategory=20)
```



Now, let's a dotplot where similar to the barplot, each dot in the plot represents a GO term. The size and color of the dots convey information about the significance and enrichment level of each term. Larger and more colorful dots represent more significant terms. The ego object is used to generate this plot as well.

```
dotplot(ego)
```



These visualizations are important because they provide a quick and intuitive way to understand which biological processes are most relevant or significantly enriched in your dataset. Researchers can use these plots to identify key pathways or functions associated with their genes of interest.

For example, in a genomics study, these plots could help identify the biological processes that are most affected by differentially expressed genes, shedding light on the underlying molecular mechanisms of a particular condition or disease.

12.12.5 Introduction to MSigDB

You can check the docs [here](#).

MSigDB (Molecular Signatures Database) is a widely used resource for gene set enrichment analysis. It provides curated collections of gene sets representing various biological processes, pathways, and functional categories. These gene sets are organized into different categories, making it easier to explore and analyze.

The following are some of the key categories in MSigDB:

- H (Hallmark gene sets): These are a set of 50 gene sets representing well-defined biological states or processes.
- C1 (Positional gene sets): Gene sets based on chromosomal location.

290CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

- C2 (Curated gene sets): Manually curated gene sets from various sources, including pathway databases.
- C3 (Motif gene sets): Gene sets related to transcription factor binding motifs.
- C4 (Computational gene sets): Gene sets generated through computational methods.
- C5 (GO gene sets): Gene sets based on Gene Ontology terms.
- C6 (Oncogenic signatures): Gene sets associated with cancer-related processes.
- C7 (Immunologic signatures): Gene sets related to immune system processes.

Let's start by loading the `msigdbr` package and accessing the MSigDB database. We'll specifically focus on the "Hallmark gene sets" (category H) for Homo sapiens (human).

```
# Load the msigdbr package
library(msigdbr)

# Retrieve Hallmark gene sets for Homo sapiens
m_df <- msigdbr(species = "Homo sapiens")

head(m_df)

## # A tibble: 6 x 15
##   gs_cat gs_subcat     gs_name      gene_symbol entrez_gene ensembl_gene
##   <chr>  <chr>       <chr>        <chr>       <int>      <chr>
## 1 C3    MIR:Mir_Legacy AAACCAC_MIR140 ABCC4          10257 ENSG00000125257
## 2 C3    MIR:Mir_Legacy AAACCAC_MIR140 ABRAXAS2        23172 ENSG00000165660
## 3 C3    MIR:Mir_Legacy AAACCAC_MIR140 ACTN4           81   ENSG00000130402
## 4 C3    MIR:Mir_Legacy AAACCAC_MIR140 ACTN4           81   ENSG00000282844
## 5 C3    MIR:Mir_Legacy AAACCAC_MIR140 ACVR1          90   ENSG00000115170
## 6 C3    MIR:Mir_Legacy AAACCAC_MIR140 ADAM9          8754 ENSG00000168615
## # i 9 more variables: human_gene_symbol <chr>, human_entrez_gene <int>,
## #   human_ensembl_gene <chr>, gs_id <chr>, gs_pmid <chr>, gs_geoid <chr>,
## #   gs_exact_source <chr>, gs_url <chr>, gs_description <chr>
```

The `m_df` object now contains information about the Hallmark gene sets.

Let's use the `msigdbr` function with the `species` parameter set to "Homo sapiens" and the `category` parameter set to "H" to retrieve Hallmark gene sets. We then use the `dplyr::select` function to extract only the columns containing the gene set names (`gs_name`) and associated entrez gene IDs (`entrez_gene`).

12.12. PATHWAY ANALYSIS USING OVER-REPRESENTATION AND GENE SET ENRICHMENT ANALYSIS2

```
m_t2g <- msigdbr(species = "Homo sapiens", category = "H") %>%
  dplyr::select(gs_name, entrez_gene)
```

To get an overview of the number of genes in each gene set, we can create a table:

```
# Count the number of genes in each gene set
table(m_t2g$gs_name)
```

```
##                                     HALLMARK_ADIPOGENESIS
##                                     210
##                                     HALLMARK_ALLOGRAFT_REJECTION
##                                     335
##                                     HALLMARK_ANDROGEN_RESPONSE
##                                     102
##                                     HALLMARK_ANGIOGENESIS
##                                     36
##                                     HALLMARK_APICAL_JUNCTION
##                                     231
##                                     HALLMARK_APICAL_SURFACE
##                                     46
##                                     HALLMARK_APOPTOSIS
##                                     183
##                                     HALLMARK_BILE_ACID_METABOLISM
##                                     114
##                                     HALLMARK_CHOLESTEROL_HOMEOSTASIS
##                                     77
##                                     HALLMARK_COAGULATION
##                                     162
##                                     HALLMARK_COMPLEMENT
##                                     237
##                                     HALLMARK_DNA_REPAIR
##                                     170
##                                     HALLMARK_E2F_TARGETS
##                                     218
##                                     HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
##                                     204
##                                     HALLMARK_ESTROGEN_RESPONSE_EARLY
##                                     216
##                                     HALLMARK_ESTROGEN_RESPONSE_LATE
##                                     218
##                                     HALLMARK_FATTY_ACID_METABOLISM
##                                     165
```

292CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

##	HALLMARK_G2M_CHECKPOINT	
##		204
##	HALLMARK_GLYCOLYSIS	
##		215
##	HALLMARK_HEDGEHOG_SIGNALING	
##		36
##	HALLMARK_HEME_METABOLISM	
##		214
##	HALLMARK_HYPOXIA	
##		215
##	HALLMARK_IL2_STAT5_SIGNALING	
##		216
##	HALLMARK_IL6_JAK_STAT3_SIGNALING	
##		103
##	HALLMARK_INFLAMMATORY_RESPONSE	
##		222
##	HALLMARK_INTERFERON_ALPHA_RESPONSE	
##		140
##	HALLMARK_INTERFERON_GAMMA_RESPONSE	
##		286
##	HALLMARK_KRAS_SIGNALING_DN	
##		220
##	HALLMARK_KRAS_SIGNALING_UP	
##		220
##	HALLMARK_MITOTIC_SPINDLE	
##		215
##	HALLMARK_MTORC1_SIGNALING	
##		211
##	HALLMARK_MYC_TARGETS_V1	
##		236
##	HALLMARK_MYC_TARGETS_V2	
##		60
##	HALLMARK_MYOGENESIS	
##		212
##	HALLMARK_NOTCH_SIGNALING	
##		34
##	HALLMARK_OXIDATIVE_PHOSPHORYLATION	
##		220
##	HALLMARK_P53_PATHWAY	
##		215
##	HALLMARK_PANCREAS_BETA_CELLS	
##		44
##	HALLMARK_PEROXISOME	
##		110
##	HALLMARK_PI3K_AKT_MTOR_SIGNALING	
##		118

12.12. PATHWAY ANALYSIS USING OVER-REPRESENTATION AND GENE SET ENRICHMENT ANALYSIS2

```
##          HALLMARK_PROTEIN_SECRETION
##                               98
##  HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY
##                               58
##          HALLMARK_SPERMATOGENESIS
##                               144
##          HALLMARK_TGF_BETA_SIGNALING
##                               59
##          HALLMARK_TNFA_SIGNALING_VIA_NFKB
##                               228
##          HALLMARK_UNFOLDED_PROTEIN_RESPONSE
##                               115
##          HALLMARK_UV_RESPONSE_DN
##                               152
##          HALLMARK_UV_RESPONSE_UP
##                               191
##          HALLMARK_WNT_BETA_CATENIN_SIGNALING
##                               50
##          HALLMARK_XENOBIOTIC_METABOLISM
##                               224
```

```
head(m_t2g)
```

```
## # A tibble: 6 x 2
##   gs_name      entrez_gene
##   <chr>        <int>
## 1 HALLMARKADIPOGENESIS     19
## 2 HALLMARKADIPOGENESIS    11194
## 3 HALLMARKADIPOGENESIS    10449
## 4 HALLMARKADIPOGENESIS     33
## 5 HALLMARKADIPOGENESIS     34
## 6 HALLMARKADIPOGENESIS     35
```

Now that we have retrieved the Hallmark gene sets, let's perform gene set enrichment analysis using these sets.

We'll use a set of significant genes from our analysis (represented by `significant_genes_map$ENTREZID`) and the Hallmark gene sets (represented by `m_t2g`). This analysis will help us identify which Hallmark gene sets are enriched in our significant genes.

```
em <- enricher(significant_genes_map$ENTREZID, TERM2GENE = m_t2g,
                universe = background_genes_map$ENTREZID)
```

Here, we are using the `enricher` function from the `msigdbr` package to perform the enrichment analysis. It takes as input the list of significant genes, the

294CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

mapping between gene sets and genes (TERM2GENE), and the background gene set (typically all detected genes in the dataset).

Now, let's examine the results of the enrichment analysis:

```
head(em)
```

```
## HALLMARK_HYPOXIA HALLMARK_HYPOXIA
## HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
## HALLMARK_TNFA_SIGNALING_VIA_NFKB HALLMARK_TNFA_SIGNALING_VIA_NFKB HALLMARK_TNFA_SIGNALING_VIA_NFKB
## HALLMARK_KRAS_SIGNALING_UP HALLMARK_KRAS_SIGNALING_UP HALLMARK_KRAS_SIGNALING_UP
## HALLMARK_KRAS_SIGNALING_DN HALLMARK_KRAS_SIGNALING_DN HALLMARK_KRAS_SIGNALING_DN
## HALLMARK_INFLAMMATORY_RESPONSE HALLMARK_INFLAMMATORY_RESPONSE HALLMARK_INFLAMMATORY_RESPONSE
## HALLMARK_HYPOXIA HALLMARK_HYPOXIA
## HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
## HALLMARK_TNFA_SIGNALING_VIA_NFKB HALLMARK_TNFA_SIGNALING_VIA_NFKB HALLMARK_TNFA_SIGNALING_VIA_NFKB
## HALLMARK_KRAS_SIGNALING_UP HALLMARK_KRAS_SIGNALING_UP HALLMARK_KRAS_SIGNALING_UP
## HALLMARK_KRAS_SIGNALING_DN HALLMARK_KRAS_SIGNALING_DN HALLMARK_KRAS_SIGNALING_DN
## HALLMARK_INFLAMMATORY_RESPONSE HALLMARK_INFLAMMATORY_RESPONSE HALLMARK_INFLAMMATORY_RESPONSE
## GeneRatio BgRatio pvalue
## HALLMARK_HYPOXIA 29/133 191/3984 1.519657e-12
## HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION 24/133 177/3984 1.897877e-09
## HALLMARK_TNFA_SIGNALING_VIA_NFKB 19/133 188/3984 1.083454e-05
## HALLMARK_KRAS_SIGNALING_UP 18/133 176/3984 1.605325e-05
## HALLMARK_KRAS_SIGNALING_DN 14/133 149/3984 3.709578e-04
## HALLMARK_INFLAMMATORY_RESPONSE 15/133 174/3984 5.793946e-04
## p.adjust qvalue
## HALLMARK_HYPOXIA 6.534527e-11 4.958882e-11
## HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION 4.080436e-08 3.096536e-08
## HALLMARK_TNFA_SIGNALING_VIA_NFKB 1.552951e-04 1.178494e-04
## HALLMARK_KRAS_SIGNALING_UP 1.725724e-04 1.309607e-04
## HALLMARK_KRAS_SIGNALING_DN 3.190237e-03 2.420988e-03
## HALLMARK_INFLAMMATORY_RESPONSE 4.152328e-03 3.151093e-03
##
## HALLMARK_HYPOXIA 54206/1907/3491/2152/8497/3623/57007/8552
## HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION 1296/3491/4148/38
## HALLMARK_TNFA_SIGNALING_VIA_NFKB
## HALLMARK_KRAS_SIGNALING_UP
## HALLMARK_KRAS_SIGNALING_DN
## HALLMARK_INFLAMMATORY_RESPONSE
## Count
## HALLMARK_HYPOXIA 29
## HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION 24
## HALLMARK_TNFA_SIGNALING_VIA_NFKB 19
```

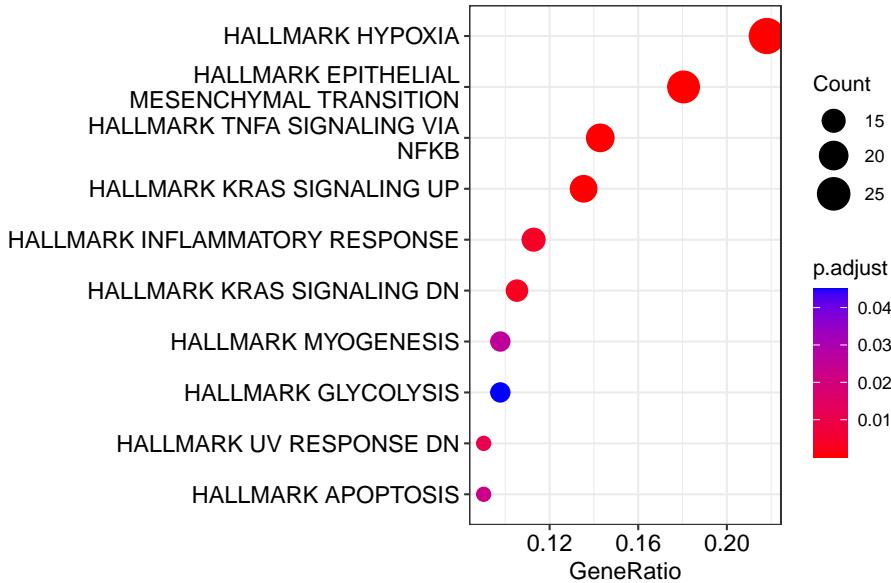
12.12. PATHWAY ANALYSIS USING OVER-REPRESENTATION AND GENE SET ENRICHMENT ANALYSIS

```
## HALLMARK_KRAS_SIGNALING_UP           18
## HALLMARK_KRAS_SIGNALING_DN            14
## HALLMARK_INFLAMMATORY_RESPONSE       15
```

The `em` object contains information about the enriched gene sets, including their names, descriptions, p-values, and adjusted p-values (q-values). The q-value represents the corrected p-value and is often used to control for false discoveries in enrichment analysis.

To visualize the enrichment results, we can create a dotplot that shows the enriched gene sets and their significance.

```
dotplot(em)
```



The dotplot will display the enriched gene sets as dots, with the size and color of the dots representing the significance of enrichment. This visualization helps identify the most significantly enriched gene sets.

Take a look, among the predefined gene sets, “HALLMARK HYPOXIA” is the pathway most strongly associated with the genes we are studying which makes perfect sense!

12.12.6 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is an alternative approach that considers the entire list of genes ranked by their association with a phenotype or condition.

Instead of just looking for over-represented pathways, GSEA assesses whether a predefined gene set (e.g., a pathway) shows a significant association with the phenotype across the entire ranked gene list.

Let's break down the steps for GSEA:

1. Accessing Gene Sets: We use the `msigdbr` package to access gene sets from resources like MSigDB. These gene sets represent predefined pathways or functional categories.
2. Conversion of Gene Symbols: Similar to ORA, we convert gene symbols to Entrez IDs for consistency.
3. Enrichment Analysis: We perform gene set enrichment analysis using the ranked gene list and predefined gene sets.
4. Statistical Analysis: Statistical tests determine whether specific gene sets are enriched at the top or bottom of the ranked list.
5. Visualization: We visualize the results using dot plots, which highlight gene sets enriched at the top or bottom of the ranked list.

12.12.7 GSEA in action

Before we dive into GSEA, we need to prepare our gene expression data. We have a dataset represented as `res_df`, which contains information about genes, including their fold change and p-values. GSEA requires the data to be pre-ranked based on a metric that combines fold change and significance, such as signed fold change multiplied by the negative logarithm of the p-value.

```
res_df <- res_df %>%
  mutate(signed_rank_stats = sign(log2FoldChange) * -log10(pvalue)) %>%
  left_join(background_genes_map, by = c("gene" = "SYMBOL")) %>%
  arrange(desc(signed_rank_stats))
```

One issue that may arise is the presence of infinite values in the data, which can cause errors during GSEA. To address this, we replace infinite values with large numbers so they will be ranked high:

```
res_df <- res_df %>%
  mutate(negative_log10pvalue = -log10(pvalue)) %>%
  mutate(negative_log10pvalue = ifelse(is.infinite(negative_log10pvalue), 1000, negative_log10pvalue)) %>%
  mutate(signed_rank_stats = sign(log2FoldChange) * negative_log10pvalue)
```

GSEA helps us determine whether a predefined set of genes, representing a biological pathway or function, is significantly enriched in our dataset. This

12.12. PATHWAY ANALYSIS USING OVER-REPRESENTATION AND GENE SET ENRICHMENT ANALYSIS2

analysis identifies whether genes within a specific set tend to be ranked higher or lower based on their association with a biological condition.

In the code below, we perform GSEA using our pre-ranked gene list and a reference gene set, represented by `m_t2g`:

```
gene_list <- res_df$signed_rank_stats
names(gene_list) <- res_df$ENTREZID

em2 <- GSEA(gene_list, TERM2GENE = m_t2g)

head(em2)
```

	ID	Description	setSize	
## HALLMARK_E2F_TARGETS	HALLMARK_E2F_TARGETS			
## HALLMARK_G2M_CHECKPOINT	HALLMARK_G2M_CHECKPOINT			
## HALLMARK_HYPOXIA	HALLMARK_HYPOXIA			
## HALLMARK_OXIDATIVE_PHOSPHORYLATION	HALLMARK_OXIDATIVE_PHOSPHORYLATION			
## HALLMARK_MYC_TARGETS_V1	HALLMARK_MYC_TARGETS_V1			
## HALLMARK_TNFA_SIGNALING_VIA_NFKB	HALLMARK_TNFA_SIGNALING_VIA_NFKB			
##		enrichmentScore	NES	pvalue
## HALLMARK_E2F_TARGETS	HALLMARK_E2F_TARGETS	-0.8226116	-2.080049	1.000000e-10
## HALLMARK_G2M_CHECKPOINT	HALLMARK_G2M_CHECKPOINT	-0.8006879	-2.022015	1.000000e-10
## HALLMARK_HYPOXIA	HALLMARK_HYPOXIA	0.9471260	1.967995	1.000000e-10
## HALLMARK_OXIDATIVE_PHOSPHORYLATION	HALLMARK_OXIDATIVE_PHOSPHORYLATION	-0.7802383	-1.973323	1.000000e-10
## HALLMARK_MYC_TARGETS_V1	HALLMARK_MYC_TARGETS_V1	-0.7234918	-1.828622	2.641130e-07
## HALLMARK_TNFA_SIGNALING_VIA_NFKB	HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.8959159	1.861281	3.946765e-07
##		p.adjust	qvalues	rank
## HALLMARK_E2F_TARGETS	1.250000e-09	7.105263e-10	1736	
## HALLMARK_G2M_CHECKPOINT	1.250000e-09	7.105263e-10	2206	
## HALLMARK_HYPOXIA	1.250000e-09	7.105263e-10	639	
## HALLMARK_OXIDATIVE_PHOSPHORYLATION	1.250000e-09	7.105263e-10	2634	
## HALLMARK_MYC_TARGETS_V1	2.641130e-06	1.501274e-06	1203	
## HALLMARK_TNFA_SIGNALING_VIA_NFKB	3.288971e-06	1.869520e-06	1067	
##		leading_edge		
## HALLMARK_E2F_TARGETS	tags=60%, list=10%, signal=55%			
## HALLMARK_G2M_CHECKPOINT	tags=60%, list=12%, signal=53%			
## HALLMARK_HYPOXIA	tags=44%, list=4%, signal=43%			
## HALLMARK_OXIDATIVE_PHOSPHORYLATION	tags=61%, list=15%, signal=53%			

```

## HALLMARK_MYC_TARGETS_V1           tags=75%, list=7%, signal=70%
## HALLMARK_TNFA_SIGNALING_VIA_NFKB   tags=36%, list=6%, signal=34%
##
## HALLMARK_E2F_TARGETS
## HALLMARK_G2M_CHECKPOINT
## HALLMARK_HYPOXIA
## HALLMARK_OXIDATIVE_PHOSPHORYLATION
## HALLMARK_MYC_TARGETS_V1           10399/6432/5687/5683/5682/6188/10549/11335/3251/
## HALLMARK_TNFA_SIGNALING_VIA_NFKB

```

This output includes several columns, such as the gene set ID, description, size, enrichment score, normalized enrichment score (NES), p-value, and more. Let's briefly explain some of these terms:

- Enrichment Score (ES): Measures the degree to which a gene set is overrepresented at the top or bottom of the ranked list of genes.
- Normalized Enrichment Score (NES): The ES adjusted for the size of the gene set and dataset.
- P-value: Indicates the statistical significance of the enrichment.
- FDR-adjusted p-value (q-value): Corrected p-value to account for multiple comparisons.

If you want to visualize it in a table format, you can run:

```
em2@result %>% View()
```

Once you run it, a table or data frame containing the results of the GSEA analysis will be displayed, allowing you to examine the details of the enriched gene sets, their enrichment scores, p-values, and other relevant information.

12.12.8 Visualizing GSEA Results

After performing GSEA and obtaining results, it's crucial to visualize these results effectively to gain insights and communicate findings. We'll use the gseaplot function to create visual representations of enriched gene sets.

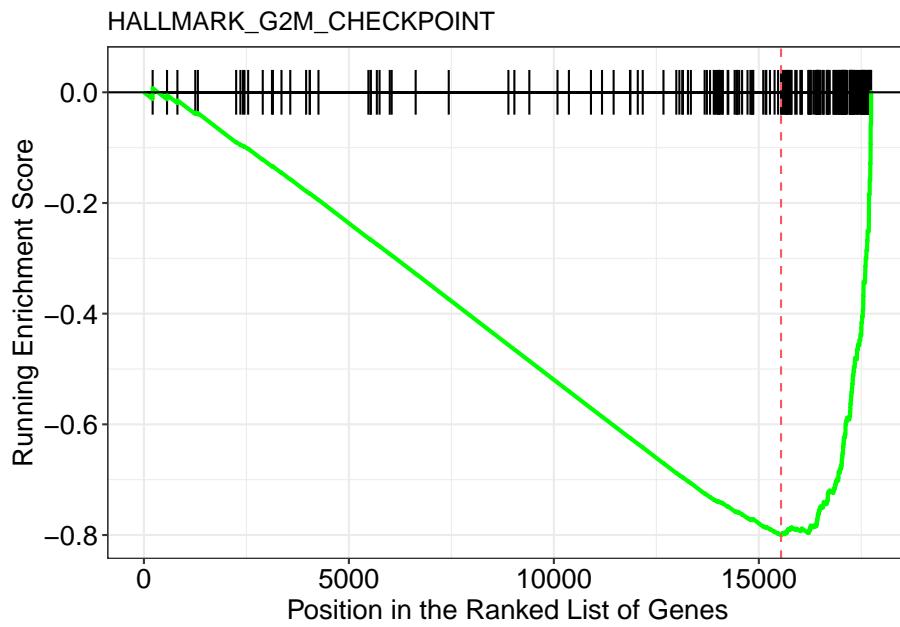
```

# save visualization in p1

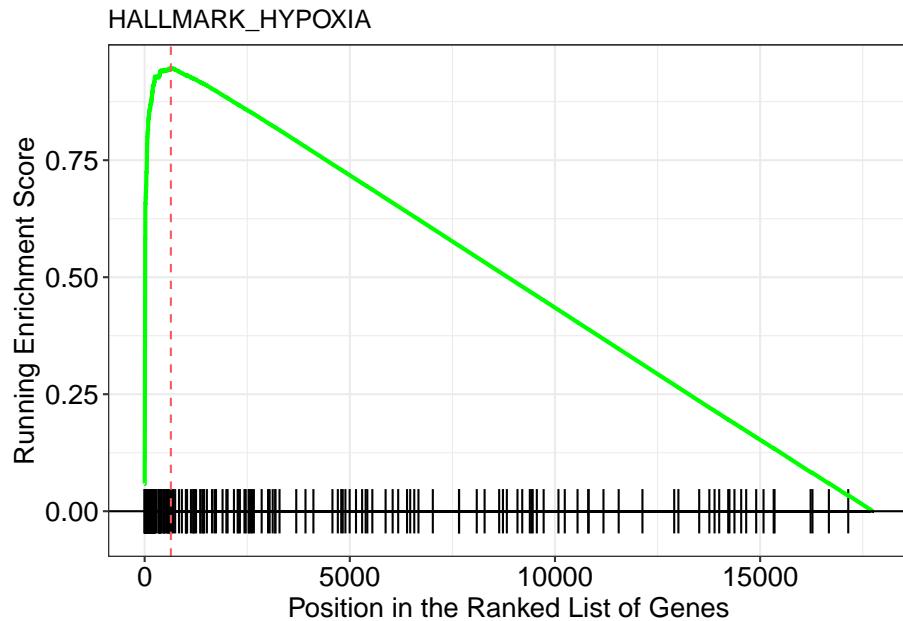
p1<- gseaplot(em2, geneSetID = "HALLMARK_G2M_CHECKPOINT",
               by = "runningScore", title = "HALLMARK_G2M_CHECKPOINT")
p1

```

12.12. PATHWAY ANALYSIS USING OVER-REPRESENTATION AND GENE SET ENRICHMENT ANALYSIS2



```
# save visualization in p2
p2 <- gseaplot(em2, geneSetID = "HALLMARK_HYPOXIA",
                 by = "runningScore", title = "HALLMARK_HYPOXIA")
p2
```



The X-axis is all your genes in the experiment (~ 20,000 in this case) pre-ranked by our metric. Each black bar is the gene in this gene set(pathway). You have an idea where are the genes located in the pre-ranked list.

Enrichment Score (ES) on the y-axis is calculated by some metric that ES is positive if the gene set is located at the top of the pre-ranked gene list. ES is negative if the gene set is located at the bottom of the pre-ranked gene list.

We see the hypoxia gene sets are at the front of the ranked list and the G2M gene sets are at the end of the ranked list.

This final step adds a vital dimension to our RNA-seq data analysis, enabling us to effectively convey the biological relevance of the enriched gene sets we've identified.

12.12.9 Conclusion

In conclusion, pathway analysis is a powerful tool in bioinformatics, providing crucial insights into the biological functions and processes associated with genes and proteins. This lesson covered two main types of pathway analysis: Over-Representation Analysis (ORA) and Gene Set Enrichment Analysis (GSEA).

Over-Representation Analysis (ORA) is best suited for scenarios where researchers have a specific list of genes of interest, such as differentially expressed genes (DEGs). ORA allows for the identification of pathways or functions that are significantly over-represented in this list, offering insights into the biological processes most affected by the condition or treatment under study.

Gene Set Enrichment Analysis (GSEA), on the other hand, is ideal for examining ranked lists of genes based on their association with a phenotype. Unlike ORA, GSEA does not rely on predefined cutoffs for significance and can provide a more nuanced view of how entire sets of genes related to specific biological processes or pathways are collectively associated with the phenotype.

Both ORA and GSEA are essential methods in bioinformatics, each serving different purposes:

- ORA provides a focused view on specific sets of genes, making it ideal for studies where the list of genes of interest is well-defined.
- GSEA offers a broader perspective, considering the **entire** spectrum of gene expression to reveal more subtle shifts in gene sets related to particular biological themes.

The choice between ORA and GSEA should be guided by the specific objectives of the research and the nature of the available data. By employing these methods, you can gain deeper insights into the complex interactions and functions of genes, leading to a better understanding of biological systems and the mechanisms underlying various conditions and diseases.

12.13 Congratulations for successfully completing this final project!

Your dedication and hard work have led you through a comprehensive journey in RNAseq data analysis, culminating in a thorough exploration of the hypoxia vs. normoxia comparison in the SW480 cell line.

Together, we've navigated the complexities of genomics data handling and analysis techniques, from cleaning and preparing metadata to identifying differentially expressed genes and visualizing gene expression patterns. Your commitment to mastering these skills is truly commendable.

As you move forward in your bioinformatics journey, remember that learning is a continuous process. The skills you've acquired here will serve as a solid foundation for tackling more advanced challenges in genomics and data analysis.

Don't forget to utilize the Q&A section and comments for any lingering questions or clarifications. Your engagement with your peers and instructors enriches the learning experience for everyone involved.

12.13.1 Final words

Congratulations to all the students who have completed this enriching course on programming in biology with a focus on the R programming language! You've

302CHAPTER 12. FINAL PROJECT: ANALYZING RNASEQ DATA FROM GEO

embarked on a fascinating journey into the world of computational analysis within the realm of biology, and you've successfully gained a valuable set of skills.

Throughout this course, we've covered a wide range of topics, starting with the fundamental concepts of programming. We've explored the core programming skills that are essential for any field, delving into variables, data types, and expressions.

The heart of the course introduced you to the powerful R programming language. You've learned basic data structures in R, and how to use R for data manipulation, analysis, and visualization – skills that are indispensable for working with complex biological data.

As you progressed, you delved into more advanced programming topics, and project management, and even explored the fascinating tidyverse ecosystem designed for data science.

We didn't stop there. We ventured into the world of data visualization, a vital skill for interpreting and sharing your data findings effectively. You also gained insights into specific bioinformatics applications, from genomic data analysis to gene expression studies using Bioconductor packages.

The culmination of your journey was the capstone project, where you applied your newfound skills to tackle real-world biological challenges by analyzing RNAseq data.

In conclusion, you've equipped yourselves with a robust toolkit for programming, tailored to the intricate world of biology. These skills will serve as a strong foundation for your future endeavors in bioinformatics and computational biology.

Remember that your learning journey doesn't end here. Keep exploring, keep studying, and keep pushing the boundaries of what you can achieve in the dynamic field of biological sciences.

It takes effort to be good at anything. You've gained the basic but practical skills. Keep practicing so you have a solid foundation. We encourage you to take on a real-world project and apply what you have learned to it. The best way to learn is by doing it.

12.13.2 Materials to read to uplift your R skills to the next level

- What They Forgot to Teach You About R
- Advanced R
- Big book of R: The collection now stands at over 300 R books. Most of them are free.

- R inferion
- Awesome Quarto
- Orchestrating Single-Cell Analysis with Bioconductor. If you want to learn to use R for single-cell analysis. This is the book you should read.
- Rstudio cheatsheets

12.13.3 Coding Challenge

You have learned a lot! To challenge yourself after this course, you can try to finish this coding experience using TCGA data as well https://github.com/crazyhottommy/coding_exercise_TCGA_infiltration

It is a bit challenging and you can ask questions on the forum to get help.