Assignement 1

Alkis

2023-03-08

#Task 1  
#2a. Wu and colleagues have presented in this article a single cell and spatially resolved transcriptomics analysis of the human breast cancers, advancing the field further towards an integrated cellular model for the classification of this disease. They boil down their work into 3 levels. Firstly, they defined a detailed cellular taxonomy, revealing at least 9 new cell types and 49 states. Using scSubtype, a tumor cell classifier of their own design, they were able to provide supporting evidence that instrinsic subtype heterogeneity exists within the majority of cancer and reflects functional differences. This major advancement might be used for prediction of innate resistance and relapse after treatment. Secondly, they generated new knowledge by creating a spatial map of cellular locations and interactions. In addition, they revealed new findings on the cellular and spatial relationship between a mesenchymal-like state and proliferation, as well as new insights in the immune phenotype of breast tumors. Lastly, they identified groups of tumors, in particular 9, that share cell type proportions and prognostic associations, which they have named ecotypes, which are partially associated with intrinsic subtypes and genomic classifiers will greatly support previous stratification methods.  
#2b. They used 10X Chromium single-cell RNA sequencing, mapping the data to the the GRCh38 reference genome using Cell Ranger Single Cell Software. Analysis of the results was conducted using Seurat v3.0.0 in R. They also used CITE-Seq antibody staining, specifically 10X chromium 3' mRNA capture compatible TotalSeq-A antibodies, and analyzed it with python and the CITE-seq-count v.1.4.3.. Spatial transcriptomics were conducted after processing of 10µm sections with Visium Spatial Gene Expression kit and alayzed with Space Ranger Software v1.0.0.  
#3a. Question 1: Can we use RNA velocity on the data generated to quantify possible cellular transitions and reveal transient cellular dynamics among these heterogeneous ecotypes? Tumor are charecterized by rapid proliferation. Could these ecotypes evolve and change different across the tumors expansion?  
#Question 2: Different treatment strategies are usually followed. After surgery chemotherapy or radiotherapy usually follows or, in some cases, hormone or targeted therapies. Can we integrate data from the remaining tissue and make similar predictions on response and survival?  
#Question 3: It is stated that they identified 29 or 49 cell states at mid- and high- resolution respectively. However, they only hypothesize that these states might represent dynamic states along a continuum of differentiation. Could the local interactions be used in order to place the different states in a the continuum and characterize them in such a wa to understand tumor progression and maybe optimal treatment?

#task 4  
#1  
sqrt(10)

## [1] 3.162278

#2  
log2(32)

## [1] 5

#3  
sum(1:1000)

## [1] 500500

#4  
sum(seq(2,1000, by=2))

## [1] 250500

#5  
100\*99/2

## [1] 4950

#6  
genes\_in\_triplets<-combn(100,3)  
sum(genes\_in\_triplets)

## [1] 24497550

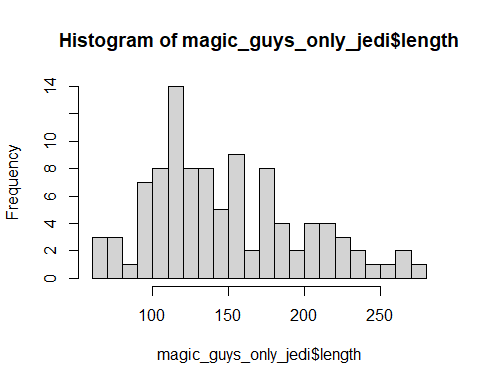
#task 5  
#1  
CO2  
#2  
 #Description  
#The CO2 data frame has 84 rows and 5 columns of data from an experiment on the cold tolerance of the grass species Echinochloa crus-galli.  
 #The CO\_2 uptake of six plants from Quebec and six plants from Mississippi was measured at several levels of ambient CO\_2 concentration. Half the plants of each type were chilled overnight before the experiment was conducted.  
#3  
Quebec<-subset(CO2, Type=="Quebec")  
ave(Quebec[,5])  
median(Quebec[,5])  
Mississippi<-subset(CO2, Type=="Mississippi")  
ave(Mississippi[,5])  
median(Mississippi[,5])

#Task 6  
#1  
example\_vector<-c(345,234,235,1945,374,745)  
mean(example\_vector)/median(example\_vector)  
#2  
mean(example\_vector[!example\_vector%in%c(max(example\_vector),min(example\_vector))])  
#3  
#Pipes are used to link together commands, minimizing your code and streamlinng analysis. To use pipes you need to add the %>% symbol from the magrittr package. Pipes can be slow and demand a lot of resources which makes them difficult to use in complex or large datasets, especially when they need debugging  
#4  
#The apply-family of functions will be useful in my work as I work with large datasets, particularly data frames, and they will allow me to apply a function to subsets of data at a time efficiently. This will not only save me time, but also simplify complex data processing tasks, making my work more efficient and accurate.

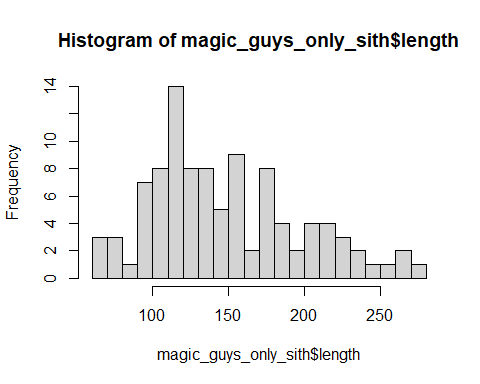
#Task 7  
  
#1a,b,c.   
  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 1.0.1   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.3.0 ✔ stringr 1.5.0   
## ✔ readr 2.1.3 ✔ forcats 1.0.0   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)  
  
magic\_guys <-read.csv(file="C:/Users/gezisi/OneDrive - KI.SE/PhD/Courses/Bioinformatics Analysis and Visualisation of Medical Genomics Data/magic\_guys.csv")  
magic\_guys  
magic\_guys\_only\_jedi <- magic\_guys[magic\_guys$length != "sith", ]  
magic\_guys\_only\_sith <- magic\_guys[magic\_guys$length != "jedi", ]  
  
png(file = "histogram.png", width=1600, height=1167) #saves as png, best for figures  
par(mfrow=c(1,2)) #makes two graphs next to each other  
hist(magic\_guys\_only\_jedi$length, breaks=10)   
hist(magic\_guys\_only\_sith$length, breaks=10)  
dev.off() #closes the png  
  
hist(magic\_guys\_only\_jedi$length, breaks=20)

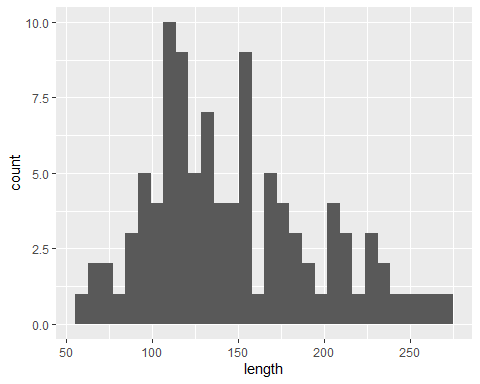


hist(magic\_guys\_only\_sith$length, breaks=20)

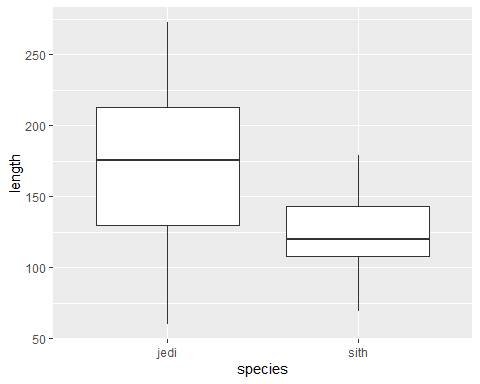


pdf(file = "histogram.pdf") #saves as pdf, for documents  
par(mfrow=c(1,2)) #makes two graphs next to each other  
hist(magic\_guys\_only\_jedi$length, breaks=10)   
hist(magic\_guys\_only\_sith$length, breaks=10)  
dev.off() #closes the pdf  
  
svg(file = "histogram.svg") #saves as svg, useful for illustrator, you an change the labels etc  
par(mfrow=c(1,2)) #makes two graphs next to each other  
hist(magic\_guys\_only\_jedi$length, breaks=10)   
hist(magic\_guys\_only\_sith$length, breaks=10)  
dev.off() #closes the svg  
  
ggplot(magic\_guys, aes(x=length)) + geom\_histogram() #histogram but with ggplot

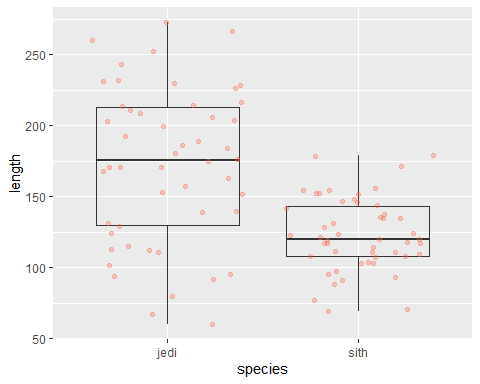
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



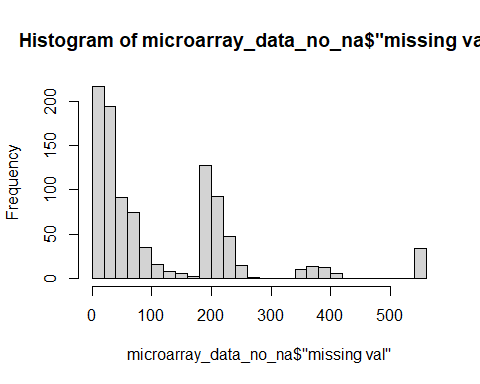
#the height of both species is normally distributed   
  
#boxplots  
png(file = "boxplot.png", width=1600, height=1167) #makes a png file  
par(mfrow=c(1,2))  
boxplot(magic\_guys\_only\_jedi$length, horizontal=TRUE, main="Jedi height")  
boxplot(magic\_guys\_only\_sith$length, horizontal=TRUE, main="Sith height")  
dev.off() #closes the png  
  
ggplot(data = magic\_guys, mapping = aes(x = species, y = length)) +  
 geom\_boxplot() #boxplot but with ggplot



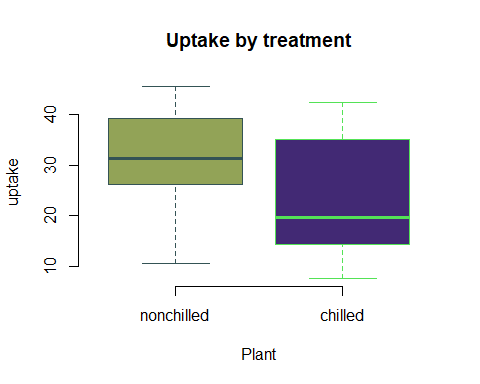
ggplot(data = magic\_guys, mapping = aes(x = species, y = length)) +  
 geom\_boxplot(alpha = 0) +  
 geom\_jitter(alpha = 0.3, color = "tomato") #example of different visualisation



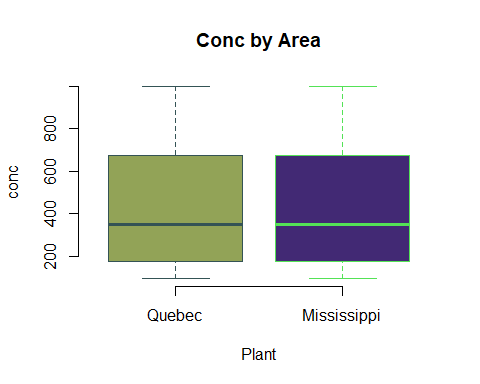
#2b  
  
sum(is.na (microarray\_data)) #there is a total of 117826 missing values in the matrix  
sapply(microarray\_data, function(x) sum(length(which(is.na(x))))) #finds n/a per column  
microarray\_data\_no\_na <- sapply(microarray\_data, function(x) sum(length(which(is.na(x)))))  
microarray\_data\_no\_na<-as.matrix(microarray\_data\_no\_na)   
microarray\_data\_no\_na <- cbind(rownames(microarray\_data\_no\_na), data.frame(microarray\_data\_no\_na, row.names=NULL))  
colnames(microarray\_data\_no\_na) <- c("gene", "missing val")  
microarray\_data\_no\_na  
hist(microarray\_data\_no\_na$"missing val", breaks=30) #visualization



#2c  
percent\_missing\_ten <- colMeans(is.na(microarray\_data))  
genes\_with\_more\_than\_10\_percent\_missing <- names(percent\_missing\_ten[percent\_missing\_ten > 0.1])  
genes\_with\_more\_than\_10\_percent\_missing  
  
percent\_missing\_twenty <- colMeans(is.na(microarray\_data))  
genes\_with\_more\_than\_10\_percent\_missing <- names(percent\_missing\_twenty[percent\_missing\_twenty > 0.2])  
genes\_with\_more\_than\_10\_percent\_missing  
  
percent\_missing\_fifty <- colMeans(is.na(microarray\_data))  
genes\_with\_more\_than\_10\_percent\_missing <- names(percent\_missing\_fifty[percent\_missing\_fifty > 0.5])  
genes\_with\_more\_than\_10\_percent\_missing  
  
#2d  
  
row\_means <- rowMeans(microarray\_data, na.rm=TRUE)  
row\_means  
microarray\_data\_replaced<-apply(microarray\_data, 2, function(x) {  
 x[is.na(x)] <- row\_means  
 x  
})  
microarray\_data\_replaced  
  
#3  
CO2  
  
boxplot(uptake ~ Treatment, data=CO2, breaks=30, frame=F, border=c("#345355", "#54E256"), col=c("#92A357", "#422974"),xlab="Plant", main="Uptake by treatment")#We see that nonchilled plants have more uptake of CO2 and that the chilled ones have bigger variation



boxplot(conc ~ Type, data=CO2, breaks=30, frame=F, border=c("#345355", "#54E256"), col=c("#92A357", "#422974"),xlab="Plant", main="Conc by Area")#The Conc is not dependent by the area

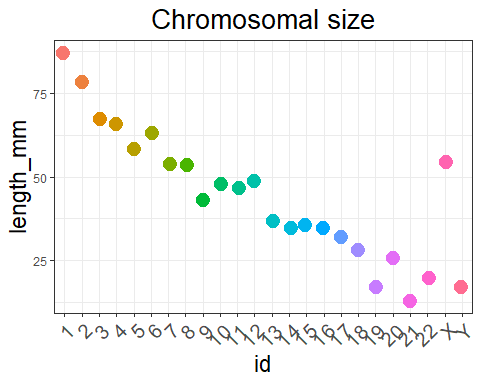


#Task 8  
  
#8a  
devtools::install\_github("hirscheylab/tidybiology")

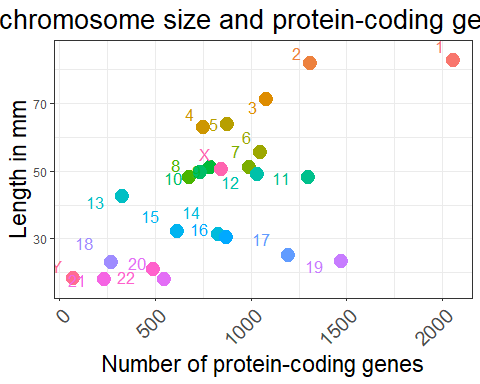
## WARNING: Rtools is required to build R packages, but is not currently installed.  
##   
## Please download and install Rtools 4.2 from https://cran.r-project.org/bin/windows/Rtools/ or https://www.r-project.org/nosvn/winutf8/ucrt3/.

## Skipping install of 'tidybiology' from a github remote, the SHA1 (d03a810a) has not changed since last install.  
## Use `force = TRUE` to force installation

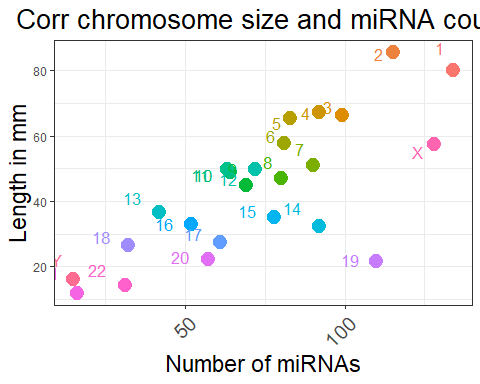
library(tidybiology)  
#a  
library(tidyverse)  
data(chromosome)  
chromosome  
Statistics<-summarise(chromosome,across(c(variations, protein\_codinggenes, mi\_rna),list(mean, median, max)))  
Statistics <- as.data.frame(matrix(unlist(Statistics), ncol = 3, byrow = TRUE))  
colnames(Statistics)<-c("Mean", "Median","Max")  
rownames(Statistics)<-c("Variations", "Protein-coding genes","miRNA")  
Statistics  
  
#8b  
library(ggplot2)  
Chromosomal\_size<-ggplot(chromosome)+ aes(x =id , y = length\_mm, color = id, size=1) +   
 geom\_point(position=position\_jitter(w = 0.1,h = 5)) +  
 theme\_bw() +  
 ggtitle("Chromosomal size") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1, size=15)) +  
 theme(axis.title.x = element\_text(size=17))+  
 theme(axis.title.y = element\_text(size=17))+  
 theme(plot.title = element\_text(hjust = 0.5, size=20))+  
 theme(legend.position = "none")  
Chromosomal\_size



#8c  
Chromosomal\_num\_pcg<-ggplot(chromosome)+ aes(x =protein\_codinggenes , y = length\_mm, color = id, size=1) +   
 geom\_point(position=position\_jitter(w = 0.1,h = 5)) +  
 theme\_bw() +  
 geom\_text(aes(label = id), hjust = 2, vjust = 0)+  
 ggtitle("Corr chromosome size and protein-coding gene count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1, size=15)) +  
 theme(axis.title.x = element\_text(size=17))+  
 theme(axis.title.y = element\_text(size=17))+  
 theme(plot.title = element\_text(hjust = 0.5, size=20))+  
 theme(legend.position = "none") +  
 labs(x = "Number of protein-coding genes", y = "Length in mm")  
Chromosomal\_num\_pcg



Chromosomal\_num\_miRNA<-ggplot(chromosome)+ aes(x =mi\_rna , y = length\_mm, color = id, size=1) +   
 geom\_point(position=position\_jitter(w = 0.1,h = 5)) +  
 theme\_bw() +  
 geom\_text(aes(label = id), hjust = 2, vjust = 0)+  
 ggtitle("Corr chromosome size and miRNA count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1, size=15)) +  
 theme(axis.title.x = element\_text(size=17))+  
 theme(axis.title.y = element\_text(size=17))+  
 theme(plot.title = element\_text(hjust = 0.5, size=20))+  
 theme(legend.position = "none") +  
 labs(x = "Number of miRNAs", y = "Length in mm")  
Chromosomal\_num\_miRNA



#8d  
data(proteins)  
proteins  
Protein\_mass\_length<-ggplot(proteins)+ aes(x =mass , y = length, color = uniprot\_id, size=1) +   
 geom\_point(position=position\_jitter(w = 0.1,h = 5)) +  
 theme\_bw() +  
 geom\_text(aes(label = protein\_name), hjust = 1.4, vjust = 0)+  
 ggtitle("Corr chromosome size and miRNA count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1, size=15)) +  
 theme(axis.title.x = element\_text(size=17))+  
 theme(axis.title.y = element\_text(size=17))+  
 theme(plot.title = element\_text(hjust = 0.5, size=20))+  
 theme(legend.position = "none") +  
 labs(x = "Protein mass", y = "Length")  
Protein\_mass\_length

