Homework - 3

Almog Angel

In this homework assignment, we will analyze genotyping data of 9 dog breeds.

Go to the study “The Shepherds’ Tale: A Genome-Wide Study across 9 Dog Breeds Implicates Two Loci in the Regulation of Fructosamine Serum Concentration in Belgian Shepherds” by Forsberg et al - https://doi.org/10.1371/journal.pone.0123173 (https://doi.org/10.1371/journal.pone.0123173)

- Read the abstract, introduction, discussion and the GWAS results section and answer the following questions in brief:

1. What is Fructosamine? and what disease is it associated with?
2. Which dog breeds are at low risk of developing diabetes?
3. What do the authors aim to find in this study?
4. Why are domestic dogs useful models for genetic studies of human complex diseases?

```
# Write yourt answers here:
(1)
Fructosamine is a protein that binds to the sugar molecules in the blood. It is a biomarkers of glycaemia (The blood sugar level) by ireversible reaction between glucose and free amino groups on serum proteins. Fructosamine reflects the average blood sugar concentration over the past 2–3 weeks. It can therefore be used to determine more short-term changes in a patient’s glucose cont rol and monitor the degree of balance in the blood sugar level of diabetic patients.

(2)
German shepherds and Golden retrievers are mentioned as dog breeds at low risk of developing dia betes.

(3)
The primary aim of the study is to investigate the genetic factors influencing variation in seru m fructosamine concentrations in healthy dogs across different breeds, with an emphasis on ident ifying specific genetic loci associated with fructosamine levels. Through breed-specific analyse s, the study seeks to find genetic associations that might explain the breed-specific prevalence and risk factors for diabetes. By focusing on the regulation of serum fructosamine concentration and identifying genetic associations, particularly on Belgian shepherd dogs. This understanding could provide insights into protective mechanisms against the disease in certain breeds.

(4)
Domestic dogs are useful models for genetic studies of human complex diseases for several reason s.
The first, the domestic dog has been accompanying humans for thousand years, as a result they sh ared the samd environment as humans. Also, there was a strong selection for certain traits that have created dog breeds with unique diversity among mammalian species.
By all those reason and beacuse dogs and humans share many common and complex diseases, dogs are usful model for studying the genetic basis of complex diseases in humans.
```

- Load packages

```
library(statgenGWAS)
```

```
## Warning: package 'statgenGWAS' was built under R version 4.3.3
```

```
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ lubridate 1.9.3     ✓ tibble    3.2.1
## ✓ purrr     1.0.2     ✓ tidyr     1.3.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

- Set working directory

```
setwd("C:/technion/Bio/hw3")
```

- Load data and take a look:

```
geno <- read.table("dogs.geno") # SNPs matrix of samples from nine different dog breeds
map <- read.table("dogs.map") # Map table with the SNPs ID, chromosome and position
pheno <- read.table("dogs.pheno", header = T) # Phenotypic and metadata for each dog sample (gen
otype)
View(geno)
View(map)
View(pheno)
```

Part 1: Bobby and population genetics ——

Last month you adopted a 2.5 years old mixed-breed dog named Bobby. He is very cute, friendly and quite big (weighs 30kg). Therefore, you are pretty confident he is a mix of big dog breeds.

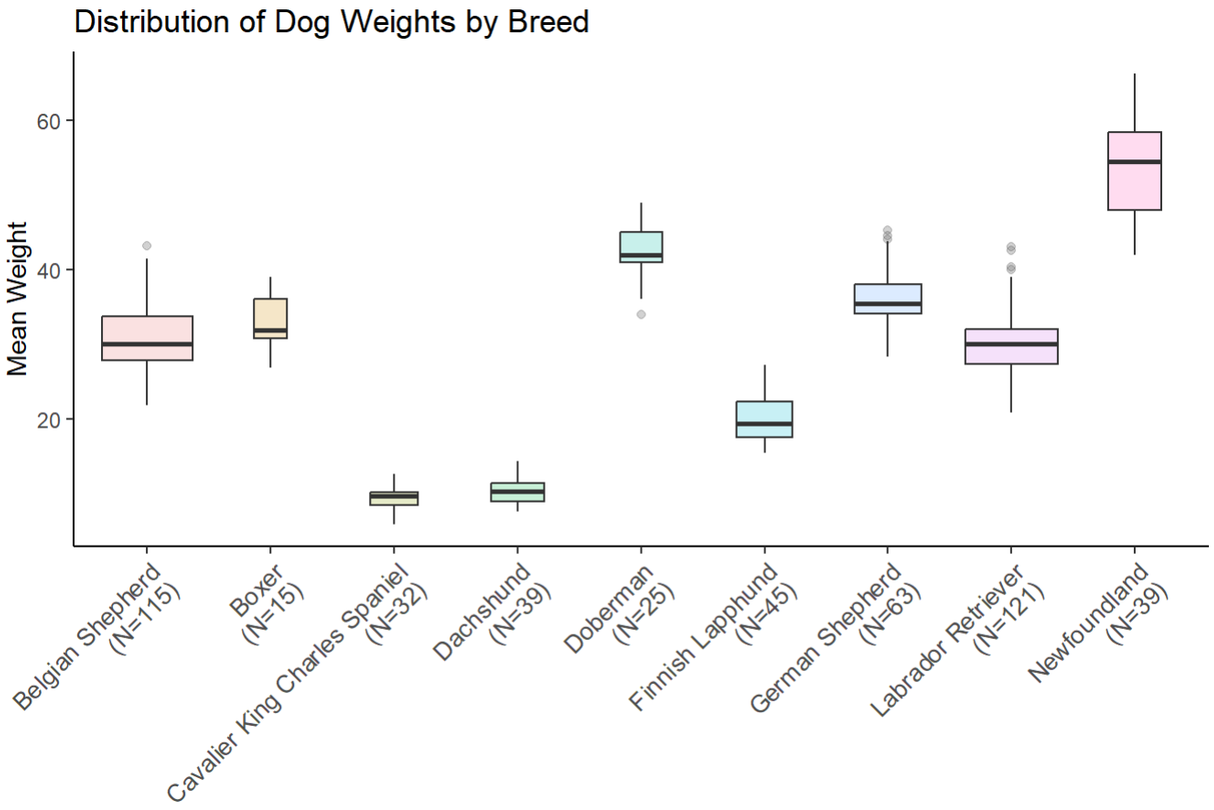
To find out what dog breed Bobby most likely related to, you did what any reasonable person would do - asked your friend from the faculty of biology to genotype a sample of Bobby.

For part 1, we would like to analyze only “big dogs” (over 20kg) as candidates for their relation to Bobby.

- Use the data in `pheno` to make a variable called `candidate_breeds` that holds a vector of dog breed names

1. Make a boxplot of dogs breeds vs. weight*
2. Based on the boxplot, use only dog breeds that reach 20kg and higher as candidates

```
candidate_breeds <- c(unique(pheno$Breed))

dogs_breeds <-  c(na.omit(pheno)$Breed)
dogs_weights <- c(na.omit(pheno)$Body_weight)

data <- data.frame(dogs_breeds,dogs_weights)
data$dogs_breeds <- factor(data$dogs_breeds)
my_xlab <- paste(levels(data$dogs_breeds),"\n(N=",table(data$dogs_breeds),")",sep="")

ggplot(data, aes(x=dogs_breeds, y=dogs_weights, fill=dogs_breeds)) +
    geom_boxplot(varwidth = TRUE, alpha=0.2) +
    theme_classic() +
    theme(legend.position="none",
        axis.text.x = element_text(angle = 45, hjust = 1, size = 10)) +
    labs(x="", y="Mean Weight") +
    scale_x_discrete(labels=my_xlab) +
    ggtitle("Distribution of Dog Weights by Breed")
```



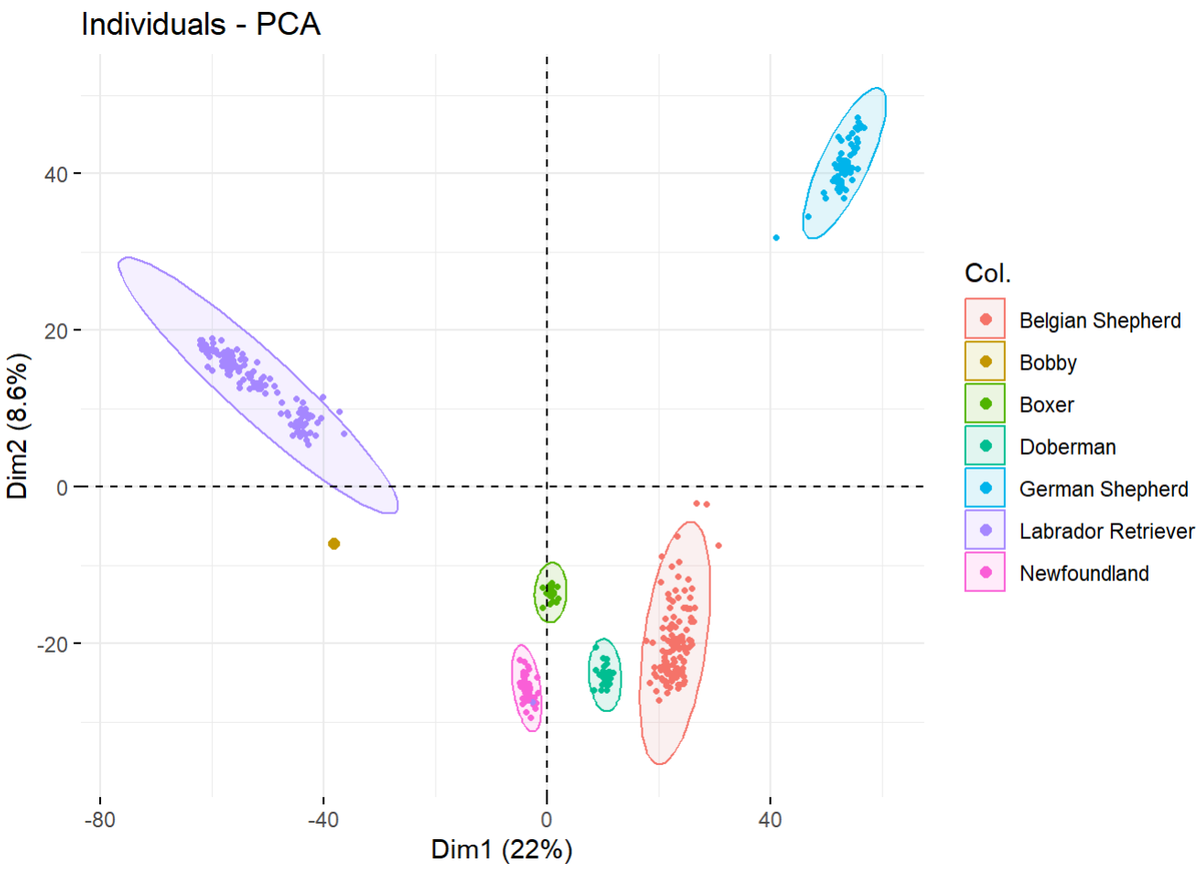

```
above_20 <- data %>%
  group_by(dogs_breeds) %>%
  summarise(mean_weight = mean(dogs_weights)) %>%
  filter(mean_weight > 20)

candidate_breeds <- c(candidate_breeds[candidate_breeds %in% above_20$dogs_breeds])
```

- Think of a computational method that we learned in class that will help you visualize and decide what dog breed Bobby is most likely related to.

1. Use only data from `geno` that belongs to `candidate_breeds`
2. Load Bobby's genotyping results file: `bobby.geno`
3. Do not forget to plot (pretty graphs get extra points!) TIP: use one of `factoextra` functions for the visualization and color the different dogs breeds

```
bobby.geno <- read.table("bobby.geno")

# filters pheno by candidate_breeds
filtered_pheno <- pheno[pheno$Breed %in% candidate_breeds, ]
# filters geno by the row names in the filtered_pheno
geno_filtered <- geno[rownames(geno) %in% rownames(filtered_pheno), ]

# creates matching vector for the pca and add bobby's data
indices <- match(rownames(geno_filtered), rownames(filtered_pheno))
breeds_for_pca <- filtered_pheno$Breed[indices]
breeds_for_pca <- c(breeds_for_pca, "Bobby")
geno_filtered <- rbind(geno_filtered, bobby.geno)

# do the pca and plot
pca <- prcomp(geno_filtered)
fviz_pca_ind(pca,
             col.ind = factor(breeds_for_pca),
             label="none",
             geom.ind=c("point"),
             pointshape=20 ,
             addEllipses=TRUE,
             ellipse.level=0.99,
             col.ind.sup = c("#1B9E77", "#D95F02", "#7570B3", "#E7298A", "#66A61E", "#E6AB02",
"#A6761D"))
```

```
## Too few points to calculate an ellipse
```



- Answer the following questions:

1. Which dog breed is the most similar to Bobby?
2. Look at the clusters of Labrador, Belgian and German Shepherd and imagine lines that connect the centroids of those cluster which results in an equilateral triangle. Is it true to say that the similarity between Belgian Shepherd to German Shepherd is equal to the similarity between Belgian Shepherd to Labrador? Explain.

```
# Write yourt answers here:
(1)
The dog breed most similar to Bobby is Labrador Retriver.
(2)
No, They dont share equal similatry.
This is true due to two main reasons:
  * The axis doesn't represent the same similatry. The x axis hold 22% precent of the similarit
y, the y axis holds 8.6%. Therefor point that is     in equal distance in the two-dimensional sp
ace but don't share the same distance in the x axis and the y axis separately, dont share equal
similarity.
  * The PCA graph only shows the two fist dimention that holds the higher precentage of diversit
y. This PCA holds 390 dimension. therefor the       similatry  doesnt represented only by this 2
dimentions.
```

Part 2: GWAS ——

- Make sure geno, map and pheno left unchanged.

```
geno <- read.table("dogs.geno")
map <- read.table("dogs.map")
pheno <- read.table("dogs.pheno", header = T)
```

- Create a gData object and call it `gDataDogs` :

1. Make sure that your data match the instructions in `?createGData()`
2. Make a list called `dogsPhenoList` of different dog breeds out of `pheno` and use only the column `genotype` and `FRUCTO`

```
new_geno <- geno
new_pheno <- pheno
new_map <- map

colnames(new_map)[2:3] <- c("chr", "pos") # Rename Chromosome and Position columns
rownames(new_map) <- new_map[["V1"]] # Use genotypes as row name

new_pheno <- rownames_to_column(new_pheno, var = "genotype")
dogsPhenoList <- split(x = new_pheno[c("genotype", "FRUCTO")],
                       f = new_pheno[["Breed"]])

gDataDogs <- createGData(geno = new_geno, map = new_map, pheno = dogsPhenoList)
```

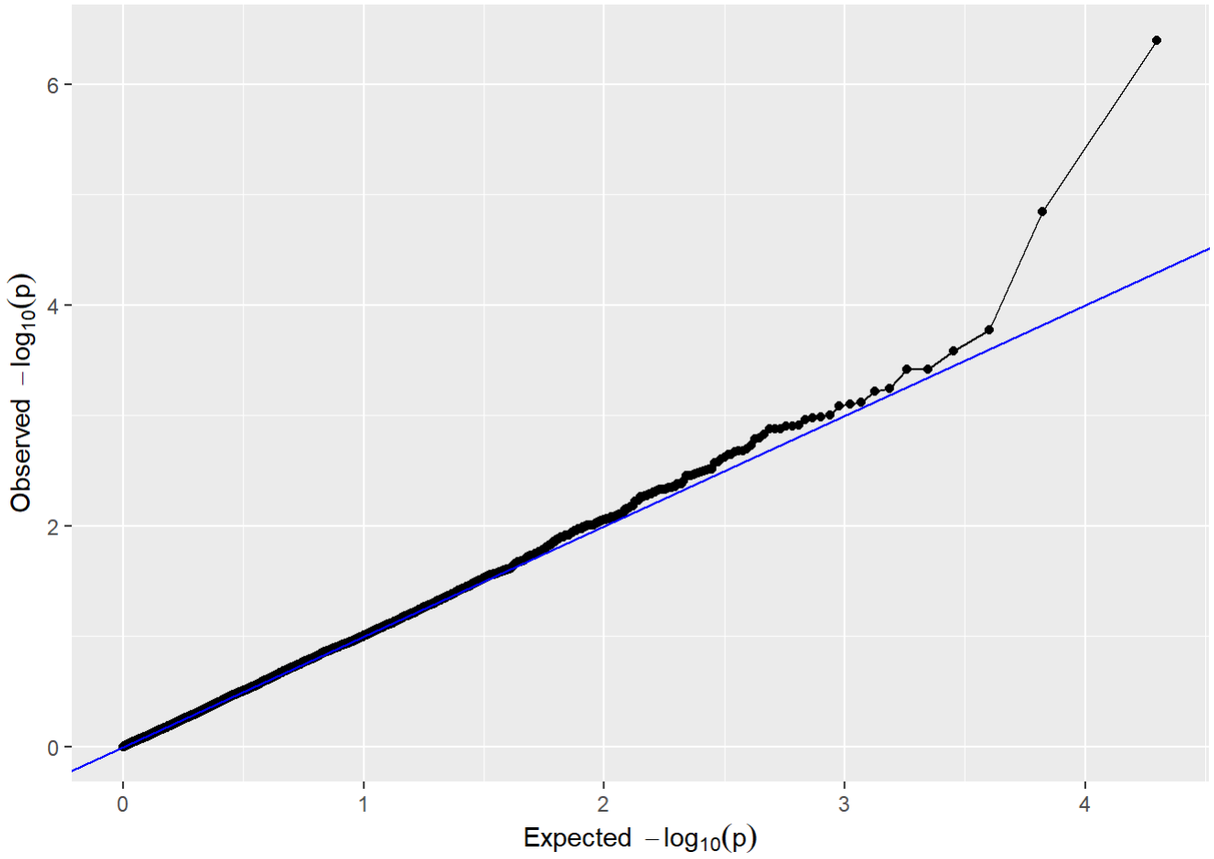
- Run a GWAS analysis for Belgian Shepherd with fructosamine concentrations:

1. Show the QQ- and Manhattan plots
2. Print significant SNP(s)

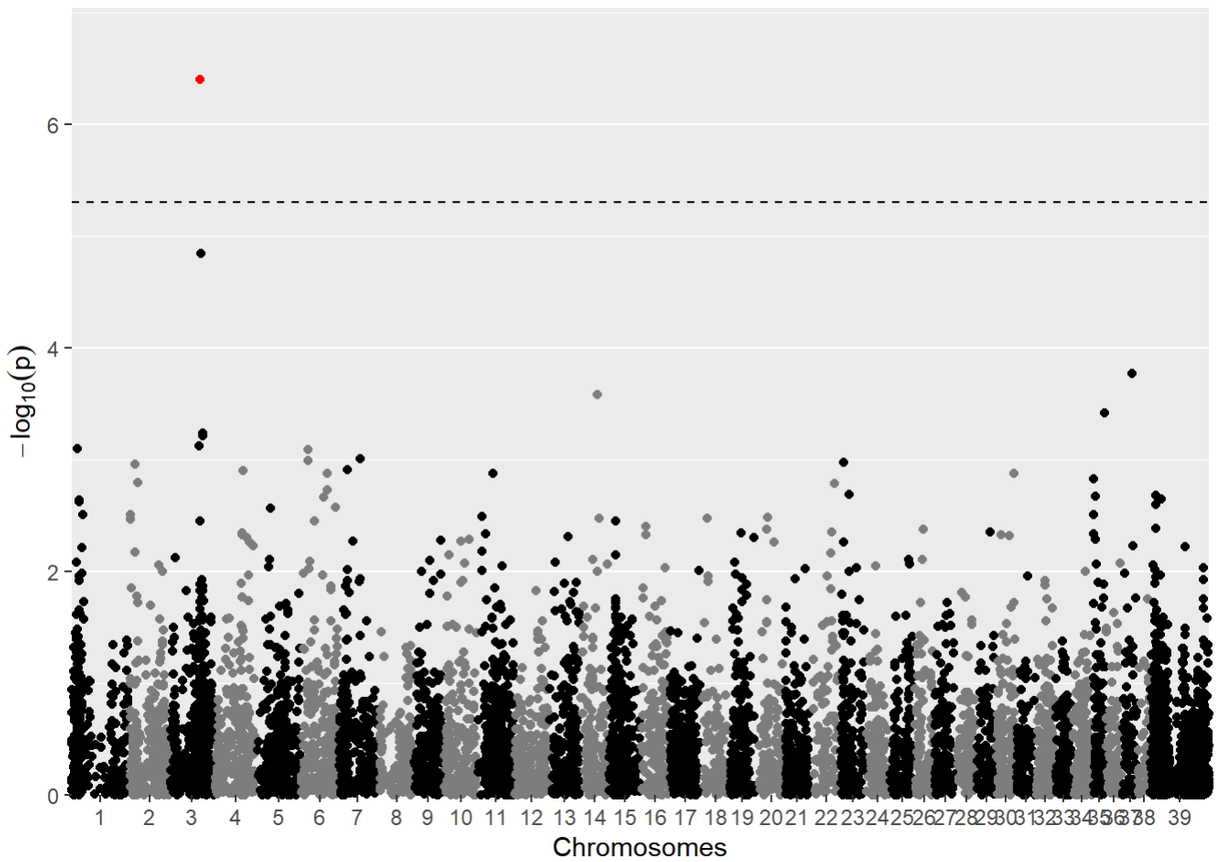
```
GWAS <- runSingleTraitGwas(gData = gDataDogs,
                                trials = "Belgian Shepherd",
                                traits = c("FRUCTO"))

summary(GWAS)
```

```
## Belgian Shepherd:
##  Traits analysed: FRUCTO
##
##  Data are available for 10000 SNPs.
##   20 of them were not analyzed because their minor allele frequency is below 0.01
##
##  GLSMethod: single
##  kinshipMethod: astle
##
##  Trait: FRUCTO
##
##      Mixed model with only polygenic effects, and no marker effects:
##      Genetic variance: 396.1156
##      Residual variance: 350.0582
##
##      LOD-threshold: 5.300161
##      Number of significant SNPs: 1
##      Smallest p-value among the significant SNPs: 4.005049e-07
##      Largest p-value among the significant SNPs: 4.005049e-07 (LOD-score: 6.397392)
##
##      No genomic control correction was applied
##      Genomic control inflation-factor: 0.991
```

```
# QQ plot
plot(GWAS, plotType = "qq", trait = "FRUCTO", main = "QQ Plot")
```




```
# Manhattan plot
plot(GWAS, plotType = "manhattan", trait = "FRUCTO", main = "Manhattan Plot")
```




```
gwas_significant <- subset(GWAS$GWAResult$`Belgian Shepherd`,  pValue< 5e-6)
gwas_significant$snp
```

```
## [1] "BICF2S2344808"
```

- Answer the following questions using the QQ- and Manhattan plot:

1. While presenting your GWAS results in an international dog-lovers confrence, an elderly woman with a well-groomed Pekingese dog in her bag, challenge your interpretation from the QQ plot. She claims that the

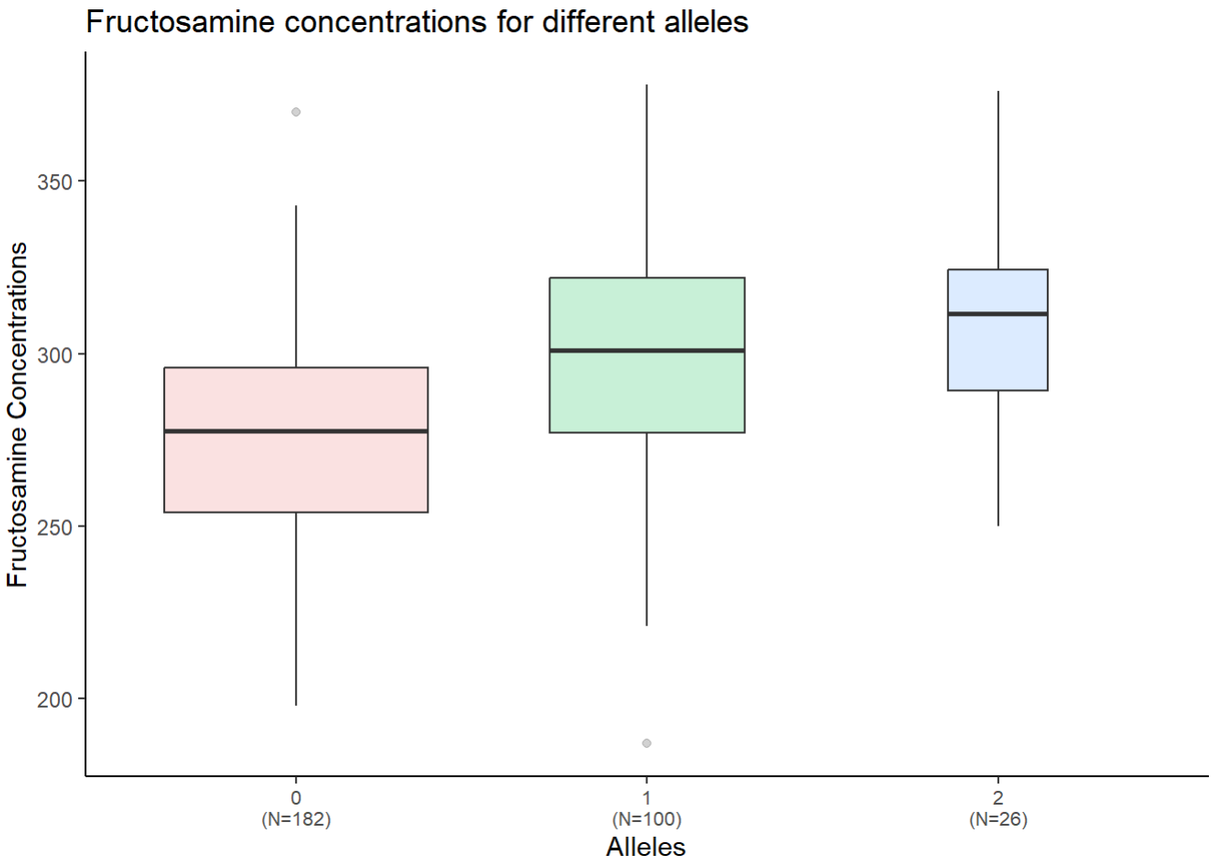
deviation observed in the QQ plot could be due to difference in the population structure rather than true genetic associations. Briefly describe how you would respond to her concerns given your QQ plot results.
2. A curious breeder from the audience, intrigued by the implications, posed the another question: “Could you explain how the location of these significant SNP(s) on your plot help us understand genetic associations with the trait?”

```
# Write yourt answers here:
(1)
The QQ plot compares the observed distribution of p-values against the expected distribution. De
viations from the blue line suggest that some SNPs have more significant associations with the t
rait. This can also represent difference in the population structure. In our examination we can
see in that in most of the region there is exact correlation. The discorellation (to the blue li
ne) is extremly further than nost of the points. if there was difference in the population struc
ture we would except more disscorelation in bigger regions than in one/two points in the graph.
Also by the menhatten plot we can see clearly one significant SNP above others. The results are
very specific and not spread over a wide spectrum of genes, what may indicates of true genetic a
ssociations ruther the opposite.

(2)
The location of significant SNP on the Manhattan plot is important to understand the genetic ass
ociations of the SNP and it's influnce on diabetes mellitus. In our GWAS, we identified a signif
icant SNP on chromosome 3 at position 65209415. By pinpointing the exact location of significant
SNP, we can narrow down the regions of the genome to investigate. In particular this region is a
ssociated with LETM1 and GAPDH that are importent protein in glucose metabolism and have previou
sly been implicated in the aetiology of diabetes mellitus. Also this region also harbours some c
andidate genes and regulatory regions but the exact mechanisms underlying the interaction are st
ill unknown. So understanding the excat region can help us diagnosys connections that are relate
d to the process but have a indirect connection to the SNP.
```

- Make a boxplot of Fructosamine concentrations for different alleles in Belgian and German Shepherds:

1. Use only the genotypes of Labrador Retriever, Belgian and German Shepherd
2. The X-axis should be the different alleles (0, 1, and 2) of the most significant SNP from the Belgian Shepherd GWAS results
3. Remove rows with NAs in fructosamine concentrations

```
# The most significant SNP: BICF2S2344808
breeds = c("Labrador Retriever", "Belgian Shepherd", "German Shepherd")
allels = c(0,1,2)

filtered_pheno <- pheno %>% filter(pheno$Breed %in% breeds)
filtered_geno <- geno %>% filter(rownames(geno) %in% rownames(filtered_pheno))

data <- data.frame(filtered_geno["BICF2S2344808"] ,filtered_pheno$FRUCTO)
data$BICF2S2344808 <-factor(data$BICF2S2344808)

my_xlab <- paste(levels(data$BICF2S2344808),"\n(N=",table(data$BICF2S2344808),")",sep="")

ggplot(data, aes(x=BICF2S2344808, y=filtered_pheno.FRUCTO, fill=BICF2S2344808)) +
    geom_boxplot(varwidth = TRUE, alpha=0.2) +
    theme_classic() +
    theme(legend.position="none",
        axis.text.x = element_text(size = 8)) +
    labs(x="Alleles", y="Fructosamine Concentrations") +
    scale_x_discrete(labels=my_xlab) +
    ggtitle("Fructosamine concentrations for different alleles")
```



- Answer the following questions:

1. Describe the results from the box-plot with respect the three allels (0, 1 and 2) and the three dog breeds.
2. Name the most significant SNP, indicates the chromosome and positions.

```
# Write your answers here:
(1)
The box plots shows us different concentrations of fructosamine depending on the genetic data in
the significant SNP of Labrador Retriever, Belgian and German Shepherd.
We can see that for allel 0 the mean fructosamine concentrations is ~275, for allel 1 it is ~300
and for allel 2 ~310. In allel 0 there is more significant difference of concentration expressed
that 1 and 2. Also by the reaserch data allel 2 is less commmon in the breeds population compare
d to the other allels (7.4%). It is the highest mean concentration and indicates of higher blood
sugar level in this individuals.
(2)
The most significant SNP is BICF2S2344808, chromosome 3 position 65209415
```

- Go to - “http://genome-euro.ucsc.edu/cgi-bin/hgGateway (http://genome-euro.ucsc.edu/cgi-bin/hgGateway)”.
- Select “Dog” from the list of species in the left.
- Select dog assembly: “May 2005 (Broad/canFam2)”.
- Select and use the chromosome and position of the SNP from the first question and click “GO”.
- Click on the last layer in the genome browser (Simple Nucleotide Polymorphism - rs23514694)
- The first part of this video can be useful: https://www.youtube.com/watch?v=8U5NhHofPI0 (https://www.youtube.com/watch?v=8U5NhHofPI0)

1. Write the nucleotides combinations of the three different alleles

```
# Write your answers here:
(1)
We found the observed data is "C/T" and the reference allele is "C", we can deduce the possible
genotypes for this SNP:
C/C: Homozygous reference genotype, where both alleles are the reference allele.
T/T: Homozygous variant genotype, where both alleles are the variant allele.
C/T or T/C: Heterozygous genotype, where one allele is the reference allele and the other is the
variant allele.
```

- Go back to the genome browser
- Zoom out (x100) three times - in the top right.

- Identify human proteins that are mapped using tBLAST to this dog genome region

2. Choose 5 proteins and check if they are mentioned in the paper. If they are mentioned, explain how they are relevant to the trait.

```
# Write your answers here:
(2)
GAPDH (Glyceraldehyde 3-phosphate dehydrogenase) is an enzyme serves to break down glucose for e
nergy and carbon molecules. In the glycolysis, glucose is broken down to pyruvate through a chai
n of chemical reactions in the cytosol. During the sixth step of glycolysis, NADH is generated a
nd in this step is catalysed by GAPDH, So this protein is important in glucose metabolism. Also
is was previously been implicated in the aetiology of diabetes. the gene is ~170kb from the most
associated SNP.

SLBP (stem-loop binding protein) and FAM53A (family with sequence similarity 53) are 2 proteins
with strond LD, within the haplotype the most tightly linked to the fructosamine associated regi
on. They are estimated ~80kb from the most associated SNP.
FAM53A is thought to play an important role in neurodevelopment by specifying the fate of dorsal
cells within the neural tube.
SLBP is required for all aspects of replication dependent histone mRNA metabolism.
FGFR3 (fibroblast growth factor receptor 3) is a protein that regulates bone growth by limiting
the formation of bone from cartilage.
It is ~88kb from the leading SNP.
Those 3 proteins are not directly linked to the glycolysis process but might be affected due to
the linkage disequilibrium in the fructosamine associated region.

We also found proteins such as CAPN5 (signal transduction in a variety of cellular processes), M
AEA (required for ubiquitination and downmodulation of surface cytokine receptor expression via
autophagy) and more that did not appear in the article.
```