MA5106 – GWAS ASSIGNMENT
Rowan Allan – 21470276

The GWAS began with 4 zip files which were unzipped to produce:

gwas.bed – Contains the genotype bed and requires plink to handle it
gwas.bim – Contains the SNP info for the GWAS

"head gwas.bim
1 rs3934834 0 995669 T C
1 rs3737728 0 1011278 A G ..."

"1" is the chromosome number, "rs3934832" is the specific SNP ID, "0" is the genetic distance, "995669" is the base pair position, "T" is allele one and "C" is allele two.

gwas.fam – Contains the sample info

"head gwas.fam
0 A2001 0 0 1 2
1 A2002 0 0 1 2 ...."

"0" is the family ID, "A2001" is the individual ID, "0" is the paternal ID, "0" is the maternal ID, "1" is the sex (1=male, 2=female) and 2 is the phenotype (2=case, 1=control), Cases and controls are necessary to identify SNPs associated with disease, the disease is unknown for this dataset.

gwas.covar – Contains covariate information (contains age of individuals)

Q1. Overview

```
plink --bfile gwas --freq --out summary
```

This produced two files:

summary.log – A .log file records the execution details of the command.

"……306102 variants loaded from .bim file.
4000 people (2000 males, 2000 females) loaded from .fam.
4000 phenotype values loaded from .fam. ……."

summary.frq – The allele frequency data per SNP is recorded in the .frq file.

"CHR SNP A1 A2 MAF NCHROBS
1 rs3934834 T C 0.09966 7636 ……"

Upon analysis of the summary.log file, 306,102 variants were loaded, 4000 people (2000 male, 2000 female) and there were 4000 founders and 0 non-founders present.

4000 phenotypes were loaded from the .fam file. However, this didn't specify the amount of cases and controls. In the .fam file column 6 indicates the phenotype:

2 = Case
1 = Control

```
awk '{print $6}' gwas.fam | sort | uniq -c
```

This command printed the number of 1s and 2s from column 6 in gwas.fam

2 = 2000
1 = 2000

Therefore, there were 2000 cases and 2000 controls

Analysis: 306,102 SNPs and 4000 individuals is a large size for carrying out a well powered GWAS for detecting traits associated with common variants.

The high genotyping rate of 98.3% listed in the summary.log file suggests there is minimal missing data which is ideal for statistical analysis.

The dataset is gender balanced allowing for the study to be adjusted for sex if needed.

0 non-founders indicates the data is likely from un-related individuals and the balance of cases and controls makes it a robust cohort.

Q2. QC Tests

i) How many SNPs is individual A2038 missing data for? (hint: use the grep linux command)

```
plink --bfile gwas --missing --out missing.stat
```

This produced three files:

miss.stat.lmiss – This file summarises missing data for SNPs

"CHR SNP N_MISS N_GENO F_MISS
1 rs3934834 182 4000 0.0455 ……"

miss.stat.imiss – Missing data for each individual

"FID IID MISS_PHENO N_MISS N_GENO F_MISS
0 A2001 N 5168 306102 0.01688 ……"

miss.stat.log – Recorded execution details of the command

"……. --missing: Sample missing data report written to miss.stat.imiss, and variant-based missing data report written to miss.stat.lmiss. ……"

The grep linux command can be used on the imiss file as it contains missing SNPs for individuals.

```
grep A2038 miss.stat.imiss
```

" 37 A2038 N 5211 306102 0.01702"

A2038 is the individual and **5211** represents the number of SNPs missing data

(5211/306102)*100 = 1.7%

The missing data for this individual is low and can be retained for the GWAS analysis.

ii) For how many individuals is the SNP rs2493272 data missing?

We can use the grep command again on the lmiss file

```
grep rs2493272 miss.stat.lmiss
```

" 1 rs2493272 111 4000 0.02775 "

1 refers to the position of the SNP which is on chromosome 1 and there are **111** individuals with missing data for this SNP.

(111/4000)*100 = 2.8%

The overall missingness for this SNP is 2.8% which is relatively low and acceptable for the GWAS analysis

iii) Create two plots which summarise the overall missingness for the data, one for SNPS and one for individuals. Would you consider the missingness rates in general to be high or low? What might this indicate about the data?

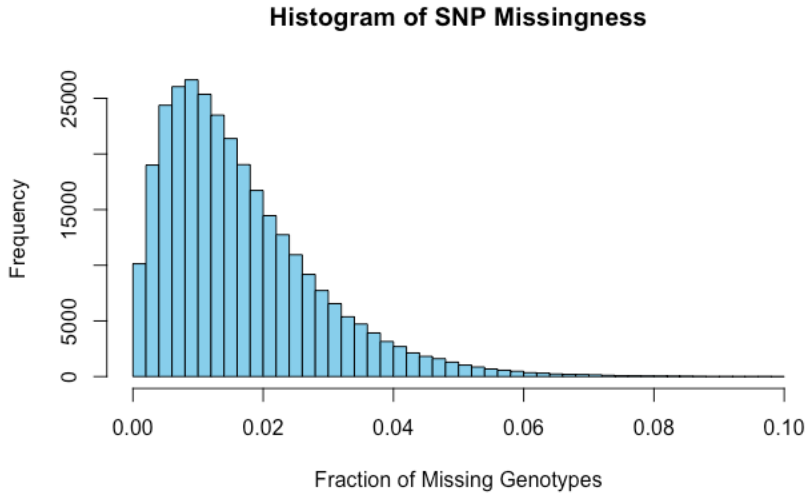
Load lmiss and imiss files into R

```
lmiss_data <- read.table("miss.stat.lmiss", header = TRUE)
imiss_data <- read.table("miss.stat.imiss", header = TRUE)
```

For the SNP missingness plot:

```
hist(lmiss_data$F_MISS, main = "Histogram of SNP Missingness",
     xlab = "Fraction of Missing Genotypes", col = "skyblue", breaks = 50)
```

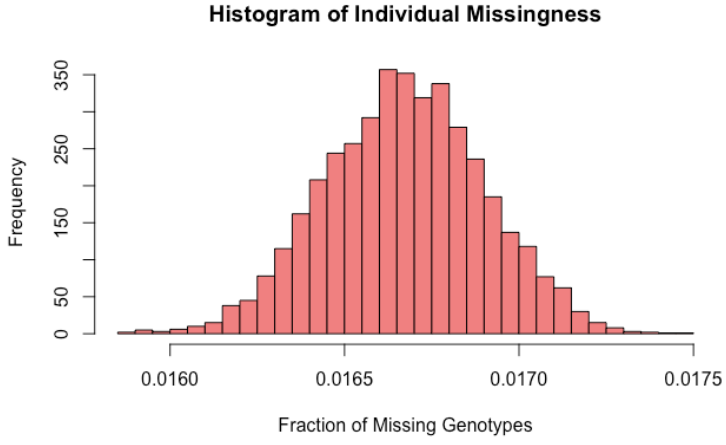
This produced:



For the Individual missingness plot:

```
hist(imiss_data$F_MISS, main = "Histogram of Individual Missingness",
     xlab = "Fraction of Missing Genotypes", col = "lightcoral", breaks = 30)
```

This produced:



Analysis:
The histogram of SNP missingness demonstrates a low rate of missingness as the peak of the distribution is around 0.01, which means most of the SNPs have about 1% of their genotypes missing. However the distribution does snake to the right meaning some SNPs have a high missingness. QC steps can be implemented to remove these SNPs

The individual missingness histogram is roughly bell-shaped, indicating a normal distribution. Most individuals have the same missingness which is between 0.0160 and 0.0175, this is relatively low as generally a missingness below 5% is acceptable.

The low missingness in this data set means the data is of high quality and is reliable.

3. Allele frequencies (10 marks)
i) Which is the minor allele for SNP rs4970357 and what is its frequency?

Taking the summary.frq file from Q1, the grep command can be used

```
grep rs4970357 summary.frq
```

"1 rs4970357 C A 0.05028 7856"

C denotes the minor allele and **0.05028** is the minor allele frequency. This frequency is greater than 5% meaning it's a common allele.

ii) Create a plot which shows the overall distribution of MAF.

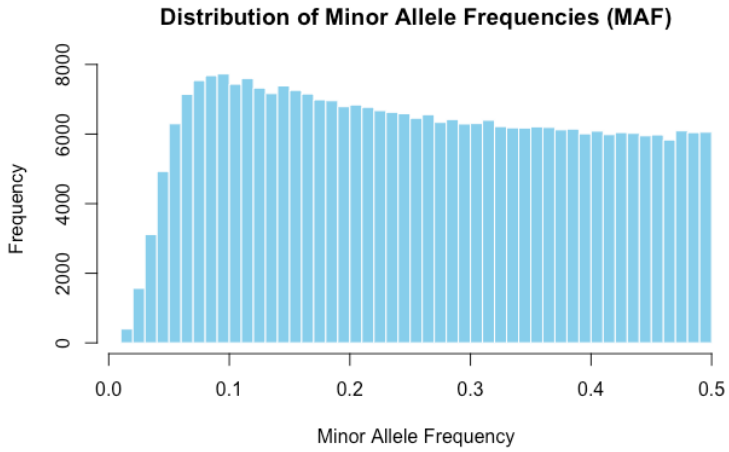
Load the summary.frq file into R

```
allele_freq <- read.table("summary.frq", header = TRUE)
```

Created histogram plot:

```
hist(allele_freq$MAF, main = "Distribution of Minor Allele Frequencies (MAF)", xlab = "Minor Allele Frequency", col = "skyblue", breaks = 50, border = "white" )
```

This produced:



Analysis:
The histogram reflects the distribution of the MAF. A minor allele having a frequency of greater than 0.05 is deemed a common allele while those with less than 0.01 are termed a rare allele. The majority of the distribution is to the right side of the histogram meaning most SNPs have relatively common alleles. There doesn't appear to be any rare alleles present in the dataset, SNP frequency sharply decreases as we move towards lower MAF values, indicating a limited presence. QC steps can be carried out to remove rare alleles to reduce the noise making it easier to detect true associations.

4. Other QC steps
i) Choose two additional QC steps to carry out on the data explaining your choice and visualising any results as appropriate.

QC Step 1:

Missingness filter for SNPs with missing rates greater than 5%. This QC step was chosen as looking at the histogram of SNP missingness you can see the distribution snake to the right displaying SNPs with high missingness over 5%.

```
plink --bfile gwas --geno 0.05 --make-bed --out SNP.miss
```

This produced four files:

SNP.miss.log - This records the execution details of the command

".....Total genotyping rate is 0.983323.
5552 variants removed due to missing genotype data (--geno).
300550 variants and 4000 people pass filters and QC. ....."

SNP.miss.bed – Binary genotype file minus the filtered SNPs
SNP.miss.bim – Variant information file
SNP.miss.fam – Sample information file

After applying the SNP filter 5552 variants were removed, this improves the overall quality of the dataset allowing for more robust results.

QC Step 2:

Further QC steps were carried out including filtering bad individuals with high genotype missingness rates (>5%), variants with rare alleles (MAF < 1%) and HWE test. However, 0 variants were removed by these after the first QC step suggesting the data set is of high quality.

```
plink --bfile SNP.miss --maf 0.01 --make-bed --out QC.data
```

This command removes low MAF SNPs which have values of less than 1%. These rare variants are removed to focus on more common variants and reduce noise in the dataset. The command produced four files:

QC.data.log – This records execution of the command

"….Total genotyping rate is 0.984114.
0 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
300550 variants and 4000 people pass filters and QC. ……"

0 variants were removed. Therefore, the .bed, .bim and .fam files remained the same.

QC.data.bed
QC.data.bim
QC.data.fam

To visualise the results a histogram conveying the updated SNP missingness was compared against the histogram produced before QC.

```
plink --bfile QC.data --missing --out QC.vis
```

QC.vis.lmiss – Summarising updated SNP missingness

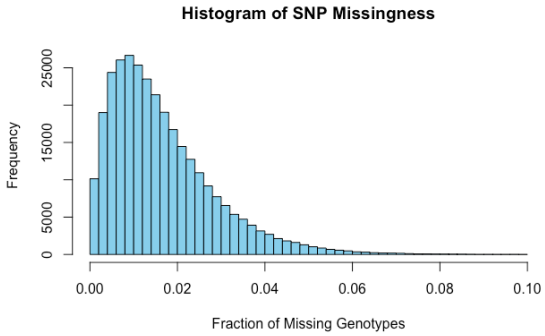
This file was loaded into R

```
lmiss_data2 <- read.table("QC.vis.lmiss", header = TRUE)
```

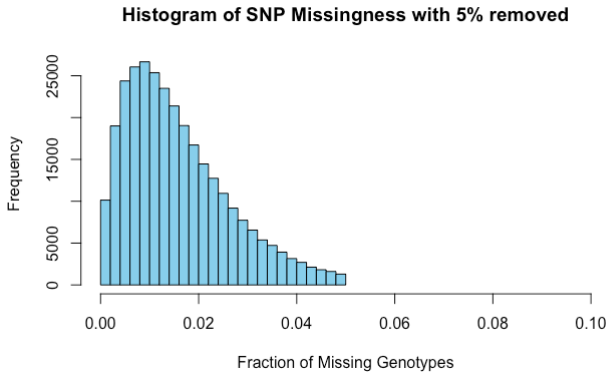
Histogram plot:

```
hist(lmiss_data2$F_MISS, main = "Histogram of SNP Missingness with 5% removed", xlab = "Fraction of Missing Genotypes", col = "skyblue", breaks = 30,  xlim = c(0, 0.1))
```

Graph before QC:



Graph after QC:



These graphs demonstrate the effectiveness of the QC tests in removing SNPs with high missingness rates. The remaining data has a lower overall missingness which will improve the calibration for upcoming statistical tests, ensuring downstream accuracy and reliability.

5. Basic association under different genetic models (DOM/REC etc.)

i) Under which genetic model does SNP rs9651273 show the smallest p-value? Note that the --model command may fail if any individual cells in the contingency table are less than 5, you can change this to a lower threshold using the --cell option (or use the --fisher option).

```
plink --bfile QC.data --snp rs9651273 --model --out model.results
```

This command produced:

model.results.model – This contains the association test results for the SNP rs9651273 under 5 different models. The p value represents the SNPs association with the phenotype which is case/control, whether the SNP is associated with an affected individual or not.

"CHR	*SNP*	*A1*	*A2*	*TEST*	*AFF*	*UNAFF*	*CHISQ*	*DF*	*P*
1	*rs9651273*	*A*	*G*	*GENO*	*322/1013/650*	*305/939/714*	*6.085*	*2*	*0.04773*
1	*rs9651273*	*A*	*G*	*TREND*	*1657/2313*	*1549/2367*	*3.995*	*1*	*0.04563*
1	*rs9651273*	*A*	*G*	*ALLELIC*	*1657/2313*	*1549/2367*	*3.892*	*1*	*0.04853*
1	*rs9651273*	*A*	*G*	*DOM*	*1335/650*	*1244/714*	*6.029*	*1*	*0.01407*
1	*rs9651273*	*A*	*G*	*REC*	*322/1663*	*305/1653*	*0.3062*	*1*	*0.58"*

The SNP shows the smallest p-value under the **DOM** model (0.01407). This is below the 0.05 threshold suggesting there is a significant association between the SNP and the phenotype.

6. Association testing with p-value correction for multiple testing (10 marks)
i) How many SNPs show a significant (<0.05) p-value under all of the multiple testing correction approaches?

First step is to carry out an association adjusted test on the data. This includes P-values from GWAS association steps after applying all of the multiple testing approaches

```
plink --bfile QC.data --assoc --adjust --out assoc.adjusted
```

This produced:

assoc.adjusted.log – This recorded execution details of the command
assoc.adjusted.assoc.adjusted – This file included the P-values corrected for multiple testing and the GC p-values.

```
"CHR     SNP       UNADJ       GC       BONF       HOLM   SIDAK_SS    SIDAK_SD    FDR_BH    FDR_BY
  3    rs6802898  2.327e-20  3.485e-20  6.994e-15  6.994e-15  6.994e-15      6.994e-15     6.994e-15    9.225e-14 …"
```

Multiple testing correction approaches include Bonferroni and FDR.

Load association results into R:

```
data <- read.table("assoc.adjusted.assoc.assoc.adjusted", header=TRUE)
```

Apply corrections:

```
significant_snps <- subset(data, BONF < 0.05 & HOLM < 0.05 & SIDAK_SS < 0.05 & SIDAK_SD < 0.05 & FDR_BH < 0.05 & FDR_BY < 0.05)
```

```
print(significant_snps)
```

This command printed:

```
   CHR       SNP     UNADJ        GC      BONF      HOLM  SIDAK_SS  SIDAK_SD    FDR_BH    FDR_BY
1    3 rs6802898 2.327e-20 3.485e-20 6.994e-15 6.994e-15 6.994e-15 6.994e-15 6.994e-15 9.225e-14
2   10 rs7901695 6.563e-12 8.217e-12 1.973e-06 1.973e-06 1.973e-06 1.973e-06 9.863e-07 1.301e-05
3   16 rs8050136 1.006e-08 1.177e-08 3.023e-03 3.022e-03 3.018e-03 3.018e-03 7.639e-04 1.008e-02
4   16 rs3751812 1.017e-08 1.190e-08 3.056e-03 3.056e-03 3.051e-03 3.051e-03 7.639e-04 1.008e-02
5   10 rs7904519 2.478e-08 2.878e-08 7.449e-03 7.449e-03 7.421e-03 7.421e-03 1.397e-03 1.843e-02
6    3 rs7615580 2.789e-08 3.235e-08 8.382e-03 8.382e-03 8.347e-03 8.347e-03 1.397e-03 1.843e-02
7   10 rs7903146 3.889e-08 4.498e-08 1.169e-02 1.169e-02 1.162e-02 1.162e-02 1.409e-03 1.859e-02
8    3 rs6768587 3.966e-08 4.586e-08 1.192e-02 1.192e-02 1.185e-02 1.185e-02 1.409e-03 1.859e-02
9    3 rs2028760 4.220e-08 4.877e-08 1.268e-02 1.268e-02 1.260e-02 1.260e-02 1.409e-03 1.859e-02
```

Therefore, 9 SNPs are identified as significant under all corrections. The Bonferroni controls for false positives while the FDR controls the false discovery rate. SNPs identified are likely to be genuinely associated with the phenotype.

ii) Is there evidence for population structure which may be confounding the analysis? Explain your answer.

Genomic control (GC) can be used to test for population structure. In GWAS population structure can cause inflation of test statistics. If there is population structure confounding the analysis it'll be represented by a lambda factor which inflates the test statistics.

Using the adjust command from earlier:

```
plink --bfile QC.data --assoc --adjust --out assoc.adjusted
```

Examine the log file:

assoc.adjusted.log – This recorded execution details of the command

"..... --adjust: Genomic inflation est. lambda (based on median chisq) = 1.00943. ....."

The lambda value of 1.00943 is very close to one meaning there is very little to no population stratification. The case and control groups are largely homogenous in respect to ancestry and that any population structure present is not impacting results. The results are likely to reflect true associations between the SNPs and the disease.

7. Logistic regression

i) Carry out a logistic regression test for association which includes both sex and age as covariates. Note that the --sex option can be used to include information on gender (which is already included in the binary files). Additional covariates can be included by using the --covar command and then providing the filename which contains the covariates. Like other types of association, the --adjust option will provide multiple testing correction. Has the lambda value changed with the covariates?

First lambda value was checked for logistic regression test for association without covariates.

```
plink --bfile QC.data --logistic --adjust --out no
```

This produced three files:

no.log – This recorded execution details of the command.

"... Writing logistic model association results to no.assoc.logistic ... done.
--adjust: Genomic inflation est. lambda (based on median chisq) = 1.01077. ..."

no.assoc.logistic – Contained association results for each SNP tested.
no.assoc.logistic.adjusted – Contained same results but with adjustments for multiple testing.

Lambda values was 1.01077 after logistic regression. Now the lambda value was check after a logistic regression association test including additional covariates like sex and age. The gwas.covar file was used. However, in order to run the command with covariates the .covar file had to be converted into text:

```
mv gwas.covar gwas.covar.txt
```

Then the test could be run:

```
plink --bfile QC.data --logistic --covar gwas.covar.txt --sex --adjust --out yes
```

This produced three files again:

yes.log

"…Writing logistic model association results to yes.assoc.logistic … done.
--adjust: Genomic inflation est. lambda (based on median chisq) = 1.0108. …."

yes.assoc.logistic
yes.assoc.logistic.adjusted

The lambda value of 1.0108 basically remained the same after carrying out the association test including the covariates. This means the population structure is not related to age or sex and they don't have a significant impact on the phenotype being tested for. This may be because the age distribution between cases and controls is balanced and doesn't drive association with the disease.

8. Plot the results
Use the R qqman package to produce a Manhattan plot visualising the results of the association test from step 7 above.

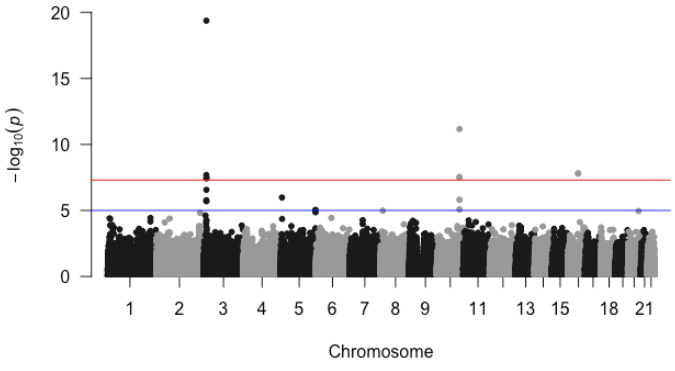
First R qqman package was installed in R:

```
install.packages("qqman")
library(qqman)
```

Association test results were loaded in and plotted:

```
results <- read.table("yes.assoc.logistic", header = TRUE)
manhattan(results, chr="CHR", bp="BP", snp="SNP", p="P" )
```

This produced:



This manhatten plot visualises the results of the association test. There are two horizontal lines signifying thresholds. The blue line marks the threshold for suggestive significance at a p value of -log10(5), while the red line is for genome-wide significance at -log10(7.3). The dots identified above the thresholds may be investigated further for their association with the trait. The x-axis denotes their chromosomal position, chromosomes 3 and 10 have peaks of significant SNPs highlighting a higher genetic involvement in the phenotype.

Significant SNPs identified:
rs7615580, rs6768587, rs2028760, rs6802898, rs7901695, rs7903146, rs7904519, rs8050136, rs3751812

The significant SNPs identified are linked to various diseases, with Type 2 diabetes (T2D) emerging as a recurrent theme. The rs6802898 SNP is strongly associated with T2D through its role in regulating glucose metabolism which influences fasting glucose levels [1]. The SNPs rs7901695 and rs7903146 were also identified as conferring risk of T2D by genotyping analyses across multiple cohorts [2]. Located within the intronic regions of the *TCF7L2* gene, these variants are linked with expression in pancreatic islets affecting B cell proliferation and diminishing insulin secretion [3]. A study targeting the *TCF7L2* rs7901695 variant found that its effects can vary based on specific macronutrient consumption [3]. Further investigation may help tailor precise dietary interventions to mitigate T2D risk. The rs8050136 SNP in the *FTO* gene has been significantly associated with insulin resistance and obesity, positioning it as a key candidate for understanding the pathogenesis of T2D [4]. The remaining SNPs were either implicated in T2D or their importance was unclear. The likely case phenotype in the GWAS was for T2D patients, given the common association among the significant SNPs.

References:

1. Mercader JM, Puiggros M, Segre AV, Planet E, Sorianello E, Sebastian D, Rodriguez-Cuenca S, Ribas V, Bonas-Guarch S, Draghici S, et al.: **Identification of novel type 2 diabetes candidate genes involved in the crosstalk between the mitochondrial and the insulin signaling systems**. *PLoS Genet* 2012, **8**:e1003046.

2. Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, et al.: **Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes**. *Nat Genet* 2006, **38**:320-323.

3. Bauer W, Adamska-Patruno E, Krasowska U, Moroz M, Fiedorczuk J, Czajkowski P, Bielska D, Gorska M, Kretowski A: **Dietary Macronutrient Intake May Influence the Effects of TCF7L2 rs7901695 Genetic Variants on Glucose Homeostasis and Obesity-Related Parameters: A Cross-Sectional Population-Based Study**. *Nutrients* 2021, **13**:1936.

4. Bego T, Causevic A, Dujic T, Malenica M, Velija-Asimi Z, Prnjavorac B, Marc J, Nekvindova J, Palicka V, Semiz S: **Association of FTO Gene Variant (rs8050136) with Type 2 Diabetes and Markers of Obesity, Glycaemic Control and Inflammation**. *J Med Biochem* 2019, **38**:153-163.