Investigating Genetic Association with Disease Susceptibility – SNP Analysis and Risk Score

(210928411)

INTRODUCTION

Genome-wide association studies (GWAS) identify associations between alleles and traits (Shaffer, Feingold and Marazita, 2012). Single nucleotide polymorphisms (SNPs) are an example of this, where individual DNA base changes occur at a frequency of 1 in 1000 base pairs (Sparks and Costenbader, 2014). These SNPs contribute to the populations' genetic diversity but can also contribute to the development or progress of diseases (Sparks and Costenbader, 2014).

The aim of the study was to investigate the genetic association between the provided SNPs (SNP1-SNP5) and disease susceptibility in a sample size of 100 individuals. Since the research assumed the trait was monogenic, there is an expectation that only one SNP will show significance for the disease risk score, which is the goal of the study.

To confirm this hypothesis, utilizing data visualisation and statistical testing like linear regression and Bonferroni correction (suggested by claude.ai) helped obtain the results to conclude the study. Once the result is obtained, the significant SNP can be used for further analysis to identify its potential role in disease development, helping the development and improvement of individuals' healthcare, prevention strategies and medical plans.

METHODS

The dataset (assignment.1.csv) provided consisted of 100 individuals' genetic information. This includes their ancestry (a categorical variable), disease risk score (a continuous variable) and their allele presence at five single nucleotide polymorphisms (SNP1 to SNP5). These alleles are labelled 0 (wild card) or 1 (mutated). To begin investigating the dataset for genetic association with disease susceptibility, the data was summarised using histograms, density plots and boxplots to visualise the relationship between the variables. For example, the plots show the distribution of the risk scores for all individuals, how those risk scores are compared across different ancestry groups and the spread/central tendency of those risk scores from each ancestry group.

The data was modelled using linear regression to determine the association between the risk scores and SNPs while considering ancestry. However, since the ancestry variable was coding for arbitrary groups, it was treated as a factor variable. Once the p-values were obtained from the linear regression, the values were adjusted using

Bonferroni correction to account for possible Type 1 errors (Armstrong, 2014), resulting in clear identification of the significant SNP.

The data was reshaped from a wide format to a long format to help plot a boxplot and bar chart, displaying the comparison of risk scores across the two SNP alleles and comparing the values before and after the Bonferroni correction using the “tidyverse”, “ggplot2”, “dplyr” and “RColorBrewer” packages provided by R/Posit Cloud.

RESULTS

The assignment.1.csv dataset containing all the genetic information on 100 individuals was analysed. Below are the key findings from each analysis carried out.

Distribution of Risk Scores:

The histogram (Figure 1) displays the distribution of risk scores in a population of 100 individuals. There is a high frequency between the 2-4 risk score scale, indicating that it is the most common score observed compared to the rest of the population. Along with this, the right-skewed distribution shown on the histogram indicates that there are more people with a higher risk score compared to those who have recorded a low-risk score, indicating that the majority of the population is at risk of disease susceptibility and may require intervention to help prevent this or close monitorsation.

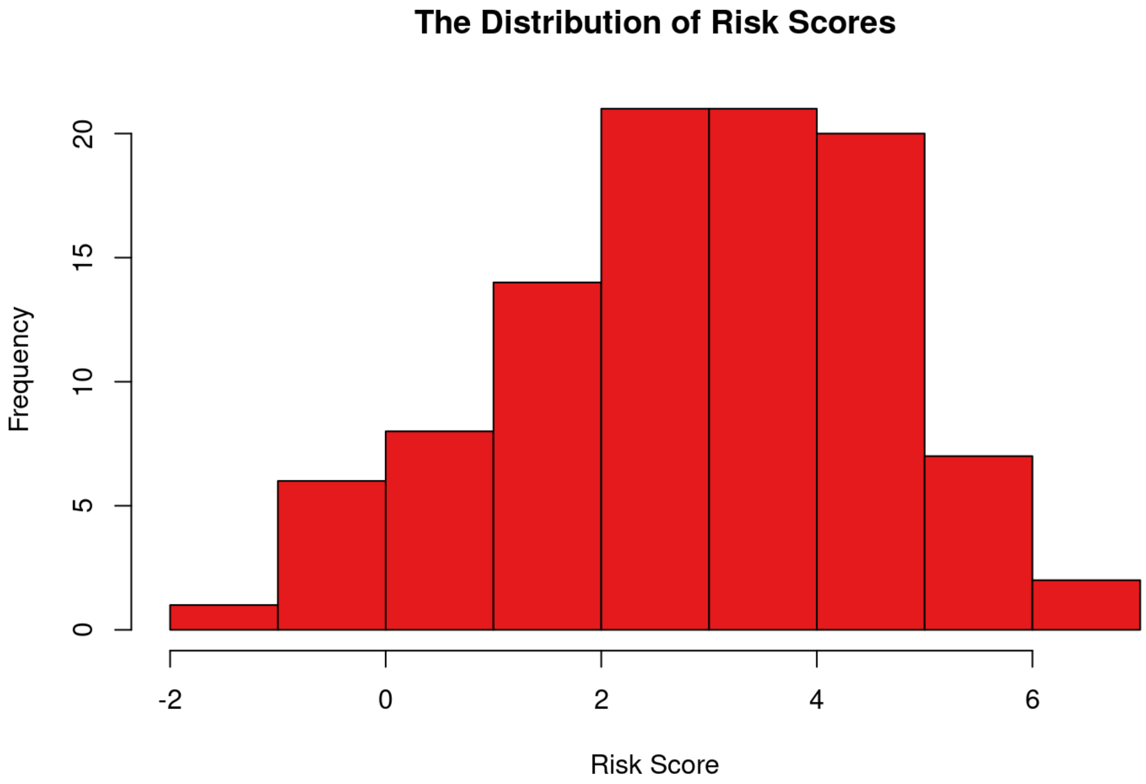


Figure 1. A histogram showing the distribution of risk scores across all individuals. The x axis is the range of risk score, while the y axis is the frequency of those risk scores.

Risk Scores stratified by Ancestry:

The density plot (Figure 2) and the box plot (Figure 3) display the risk score distribution across the two ancestry groups from the population. From Figure 2, the taller orange peak from Ancestry type 2 highlights that those individuals have a higher risk score compared to those from the teal peak (which seems evenly distributed). This indicates that individuals with Ancestry type 2 are more likely to be susceptible to genetic disease compared to those with ancestry type 1. This is supported by Figure 3, which shows that there is a higher medium from Ancestry 2 (the orange box plot) compared to Ancestry type 1, providing the understanding that ancestral background could be a factor in genetic risk, aiding in possible prevention strategies and risk assessments that prioritising individuals with ancestry type 2 to efficiency get the help they need.

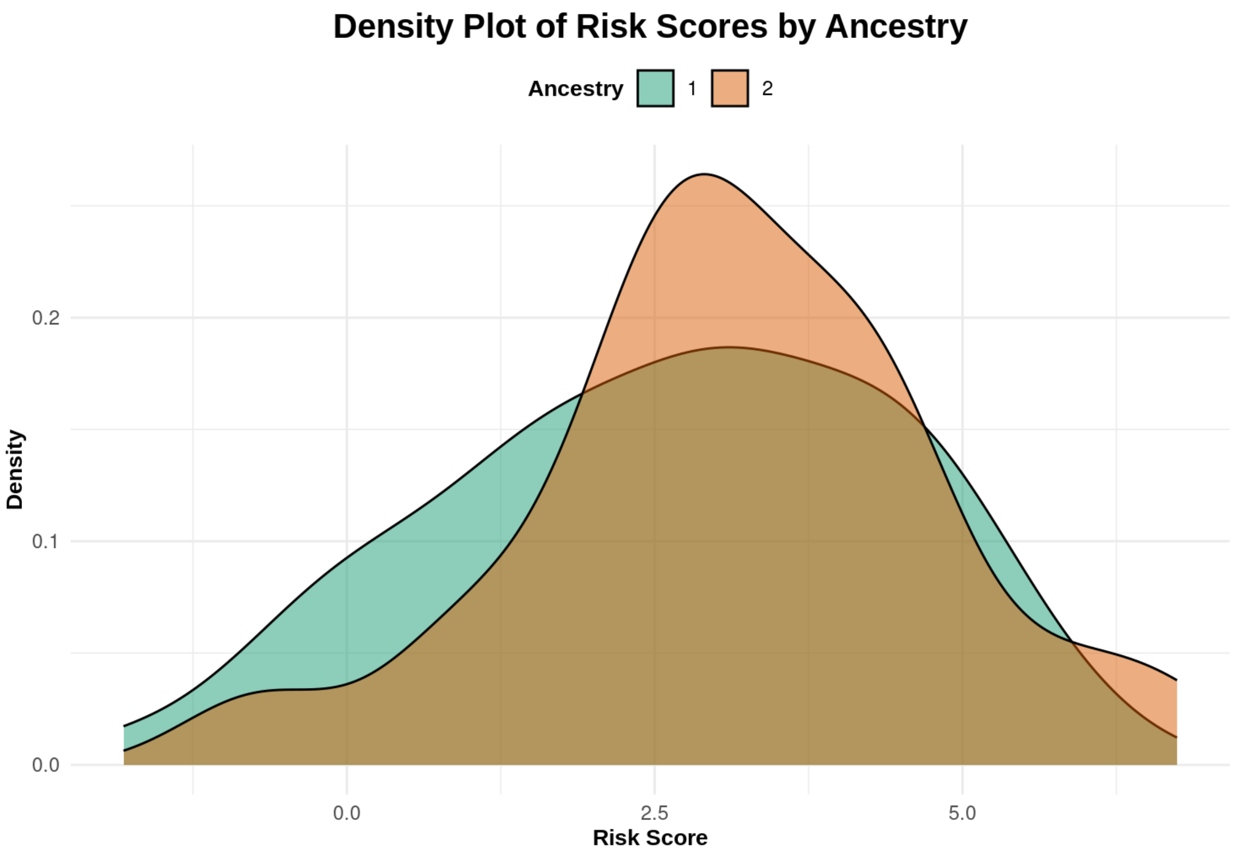


Figure 2. A density plot showing the distribution of risk scores across all individuals between two ancestry groups. The x-axis shows the distribution of risk scores, while the y-axis shows the density of the risk scores.

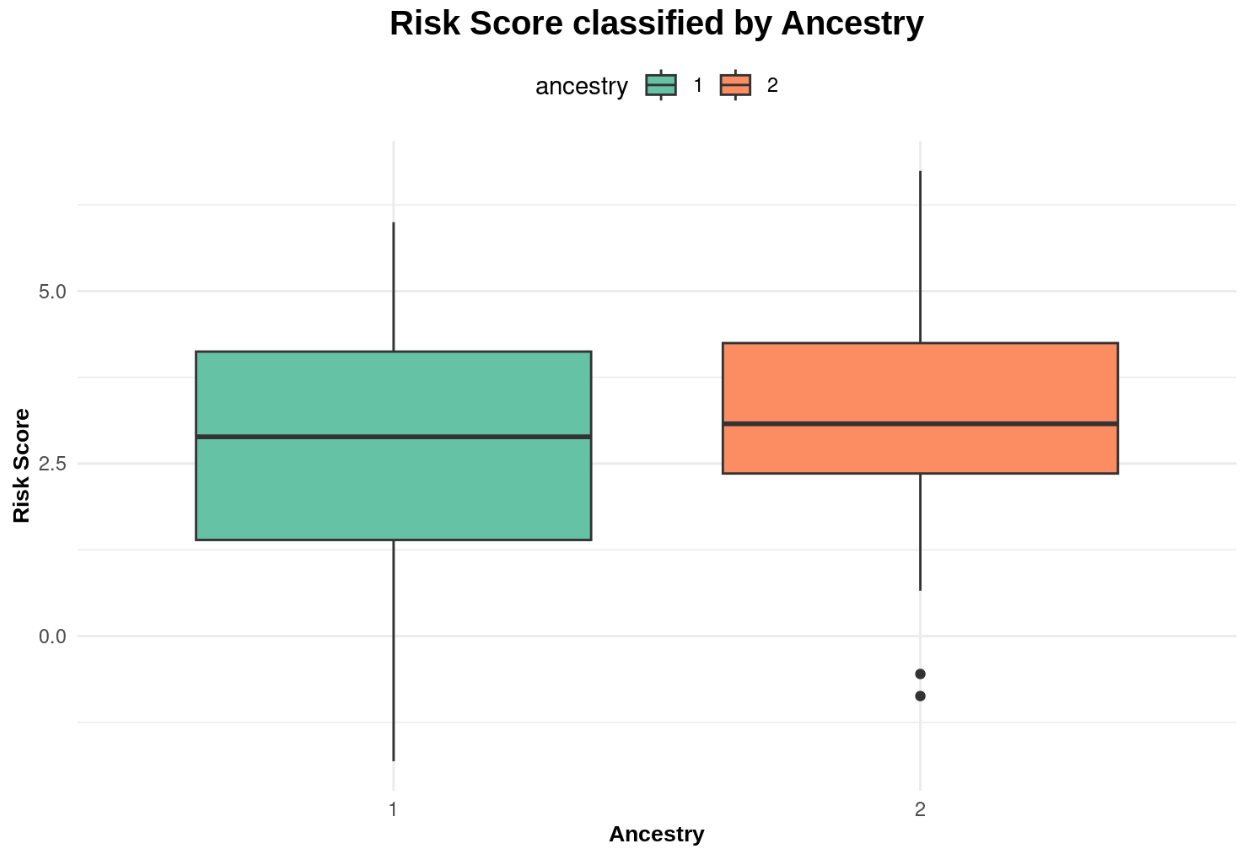


Figure 3. A box plot showing the spread and central tendency of risk scores across all individuals for each ancestry group. The x-axis shows the different ancestry groups, while the y-axis shows the risk scores.

The association between SNP and Risk Scores – Allele 0 and 1:

The box plot (Figure 4) displays how the risk scores vary across the different SNPs between individuals carrying (1) or do not carry (0) the risk allele. From Figure 4, SNP1, SNP3, SNP4 and SNP5 seem to have similar distributions and mediums for both alleles. However, there is a difference for SNP2, where the risk allele is seen to have a higher medium risk score compared to its counterparts, while the individuals who are not carrying the risk allele have a lower medium compared to the other SNPs. This indicates that SNP2 may have a genetic variant that causes a higher chance of disease susceptibilities. Although, this needs to be further investigated through statistical analysis to confirm that SNP2 is significant.

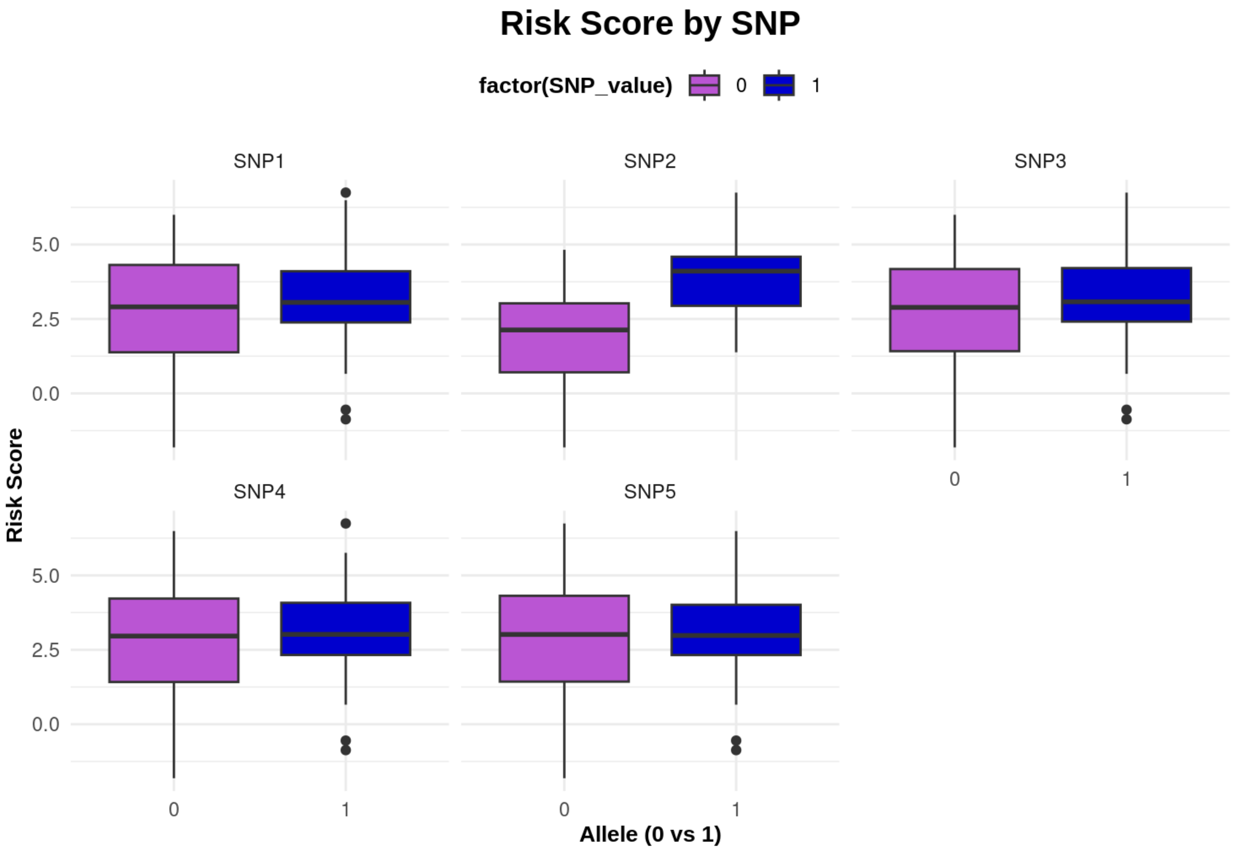


Figure 4. A box plot comparing the risk scores between all individuals that carry or do not carry the risk allele (0 vs 1) for each SNP.

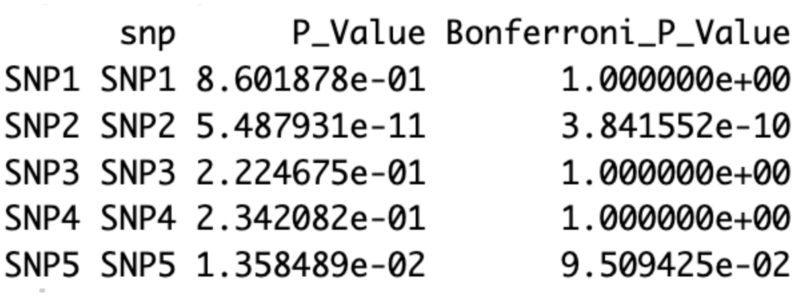
Linear Regression and Bonferroni correction:

To further understand the association between the disease risk score and the SNPs, while taking ancestry into account as a factor, a linear regression model was conducted. The p-values from the linear regression were then tested using Bonferroni correction to take into account any Type 1 errors that have occurred and further support what SNP was discovered to be significant.

The results (Figure 5) indicated that SNP2 was significant meaning it had a strong association with the risk scores, as the linear regression provided a p-value of 5.49 x 10^{-11}. This was confirmed after the Bonferroni correction, as SNP2 remained significant, providing a p-value of 3.84 x 10^{-10}, showing that both values were below 0.05. However, the same can not be said with the other SNPs (SNP1, SNP3, SNP4, SNP5) as they didn't show significance before or after the Bonferroni correction, given p-values that were above 0.05.

These findings strongly indicate that SNP2 is the genetic variant most likely associated with disease susceptibility, shown in Figure 4. The lack of significance from the other

SNPs, even after the correction, suggests that their contribution to the risk scores were either insignificant or needed further research to confirm with a larger sample size since SNP5 seems to be nearly close to the 0.05 threshold compared to the other SNPs with the p-values of 1.35 x 10^{-2} (linear) and 9.50 x 10^{-2} (Bonferroni correction).



```
      snp      P_Value Bonferroni_P_Value
SNP1 SNP1 8.601878e-01       1.000000e+00
SNP2 SNP2 5.487931e-11       3.841552e-10
SNP3 SNP3 2.224675e-01       1.000000e+00
SNP4 SNP4 2.342082e-01       1.000000e+00
SNP5 SNP5 1.358489e-02       9.509425e-02
```

Figure 5. A data frame/matrix showing the p-values from the linear regression and the Bonferroni correction.

Comparison of P-Values before and after Bonferroni Correction:

The bar chart (Figure 6) displays the comparisons of the p-values before (linear regression) and after the Bonferroni Correction. This visualises that SNP2 was the only significant SNP since the p-values were below 0.05. However, the bar chart indicates that SNP5 could be a possible candidate for further investigation as it was the closest amongst the other insignificant SNP to possibly reach/go below the significant threshold of 0.05.

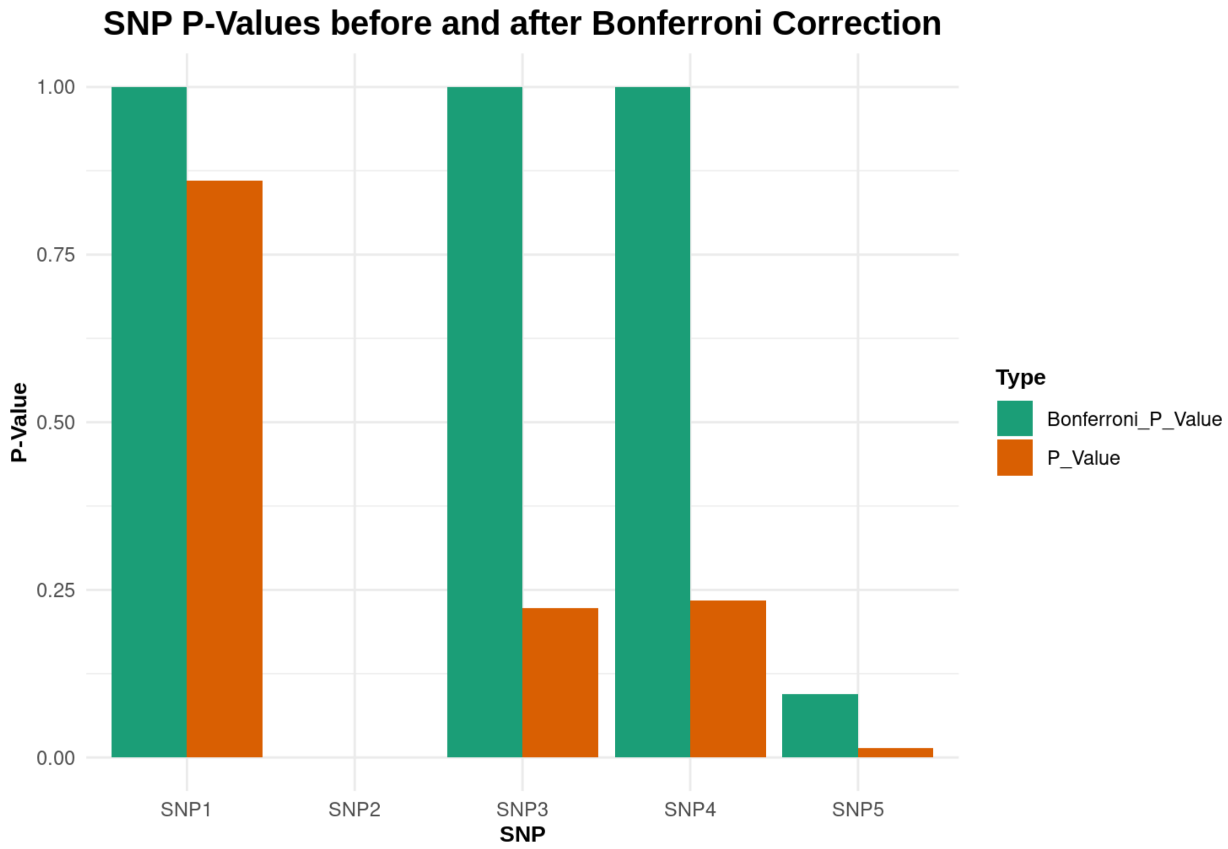


Figure 6. A bar plot showing the p-values for SNP and risk score association before and after Bonferroni correction.

DISCUSSION

In this research, the aim was to investigate the relationship between genetic association with disease susceptibility, focusing on specific single nucleotide polymorphisms (SNPs) from a GWAS of a cohort of 100 individuals. From various statistical analysis and plo visualisation, it was identified that SNP2 had displayed the most significant association with risk score. From this, we can further discuss what this result indicates, the possible limitations of this study and future consideration for further research.

Result interpretation:

The main finding from this study identified that SNP2 is strongly associated with the risk score, indicating that it is significantly associated with disease susceptibility. This is because both p-values were below the 0.05 threshold before (5.49 x 10^{-11}) and after (3.84 x 10^{-10}) the Bonferroni correction. This is supported by the linear regression conducted before the statistical test and the bar plot (Figure 6) coded to visualize the results to compare, showing the large difference in significance that SNP2 had compared to the other SNPs. However, it is worth considering SNP5 potentially being associated with disease susceptibility, even though it was concluded as non-significant, since when comparing to SNP1, SNP3 and SNP4, there was a noticeable difference between their p-values, with SNP5 being closest to the 0.05 significant threshold after SNP2.

Along with the SNPs, it is vital to consider ancestry as a control variable since it highlights further insight into the disease risk scores. This was visualised that people with ancestry type 2 were at higher risk of disease susceptibility compared to those with ancestry type 1 (Figure 2 and Figure 3). However, it doesn't provide as much significance compared to SNP2 meaning that the SNPs have more influence on the disease risk scores.

Limitations:

Despite the contributions of data visualisation and statistical testing highlighting the importance of SNPs' influence on the disease risk scores and disease susceptibility, the study has several limitations that need to be addressed.

1. Sample Size

The study was conducted with a sample size of 100 individuals. While a cohort of this size has provided evidence that SNP2 has significance in the association with the risk scores, it limits the statistical capacity to detect false positive results and other possible SNPs (in a GWS) that could also be significant, such as SNP5 (Wang et al., 2019). Having a larger sample size would solidify confidence in the statistical testing done while providing vital information that could be further investigated to construct possible prevention strategies and risk assessment for those that are affected.

2. Generalisation of Population

All the findings may not have generalized across a diverse population as there was no indication of demographic diversity within the dataset. Factors such as sex, age, environmental exposure and depth of ancestry were lacking from the dataset, giving insufficient understanding of what population is more susceptible to what disease. This should be taken into consideration as it was stated in Figures 2 and 3 that ancestry type 2 had a higher risk score overall compared to type 1, hinting that the variable could have some influence. This is demonstrated in Type 2 Diabetes (T2D) susceptibility, where population diversity led to the understanding of genetic association. This was conducted by a newly-found susceptibility locus (located at TOMM40-APOE – previously linked with Alzhemer's) showing significant heterogeneity in allelic effects between different ethnic groups, where East Asians were discovered to be the lowest risk of T2D (Cook and Morris, 2016). This finding emohasises the need to include a diverse demographic in GWAS, as factors like these can be overlooked in how genetic factors like ancestry can influence disease risk. By including a diverse sample size in a traditionally homogenous population testing style, it could help in developing more effective and tailored treatments.

3. Bonferroni Correction

Bonferroni correction is a type of statistical testing method used to reduce the possibility of Type 1 errors (false positives) being said as significant within a result. This is done when performing comparisons by adjusting your previous significant values (Kononenko and Matjaž Kukar, 2007). However, this method of statistical testing is known to be conservative (Kononenko and Matjaž Kukar, 2007). While it is effective in controlling the possibility of Type 1 errors occurring with our results, it leads to the risk of Type 2 errors (false negatives), yet it is still a popular of statistical testing (Armstrong, 2014). To counteract this limitation, the Bayesian Approach can be used as an alternative as it balances the possibility of false positives and negatives, making the data more interpretable and accepted compared to those using Bonferroni correction (Wakefield, 2012).

Future Implications for further investigation:

The investigation achieved the result of discovering SNP2 was significant in association with the disease risk score and susceptibility. Implications for future investigations should aim to address/reduce the limitations mentioned above. Having a larger sample size (>100), while including a diverse demographic in the study would enhance the reliability, understanding and confirmation of the results given. This could help provide further insight if ancestry is a vital factor to disease risk score and susceptibility. A suggestion of further investigation and statistical testing of SNP5 (with a larger population) would be recommended to confidently rule out the possibility that it is not significant.

Furthermore, the use of advanced statistic models/tests and possible machine learning approaches would provide confidence in results that are proven to be significant, lessening the risk of Type 1 and 2 errors from being identified as significant, allowing to provide better healthcare approaches and strategies.

CONCLUSION

To conclude, with the use of data visualization and statistical testing with linear regression and Bonferroni correction, the study identified that SNP2 had a strong association with disease risk scores, implicating that it had influenced on disease susceptibility. However, SNP5 had shown potential for further analysis with a larger sample size to rule out the possibility that it had a significant influence like SNP2. Additionally, it was found that ancestry had a role in disease risk, with individuals with type 2 ancestry having an overall higher risk score compared to type 1. Although, it doesn't have a high influence compared to the SNPs.

However, the findings highlighted a need for a stronger foundation for future research as a few limitations were addressed such as sample size, sample generalization and the lack of inclusion for multiple statistical testing. If these limitations are addressed, it would contribute to the understanding of how genetics play a critical role in disease susceptibility, helping construct and improve prevention methods and treatment plans.

REFERENCES

Armstrong, R.A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, [online] 34(5), pp.502–508.

Cook, J.P. and Morris, A.P. (2016). Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *European Journal of Human Genetics*, 24(8), pp.1175–1180.

Kononenko, I. and Matjaž Kukar (2007). Machine Learning Basics. *Elsevier eBooks*, [online] pp.59–105.

Shaffer, J.R., Feingold, E. and Marazita, M.L. (2012). Genome-wide Association Studies. *Journal of Dental Research*, 91(7), pp.637–641.

Sparks, J.A. and Costenbader, K.H. (2014). Genetics, Environment, and Gene-Environment Interactions in the Development of Systemic Rheumatic Diseases. *Rheumatic Disease Clinics of North America*, 40(4), pp.637–657.

Wakefield, J. (2012). Commentary: Genome-wide significance thresholds via Bayes factors. *International Journal of Epidemiology*, 41(1), pp.286–291.

Wang, Y., Li, Y., Hao, M., Liu, X., Zhang, M., Wang, J., Xiong, M., Shugart, Y.Y. and Jin, L. (2019). Robust Reference Powered Association Test of Genome-Wide