**BIOL 5174 Omics & Systems Approaches in Biology**

**STUDENT ID: 2060576**

IDCGV Report

Part A

Type 2 diabetes mellitus (T2D) is complex disease that affects approximately 300 million individuals in the world’s population. It is characterized by insulin resistance and reduced secretion of insulin by β-cells (American Diabetes Association, 2010; Wheeler and Barroso, 2011). In addition to being influenced by environmental factors, T2D has also been shown to have a strong hereditary component (Ali, 2013). This section of the report attempts to explain the most suitable approach to identifying genetic loci predisposing to T2D.The most appropriate technique available for identifying predisposing loci for a common disease like T2D is the genome-wide analysis study (GWAS) (Kruglyak, 2008). This technique allows for the interrogation of the entire genome without being limited by the need of an established disease-loci relationship.

The way a GWAS is designed affects the meaningfulness of the results. The characteristic of a phenotype majorly affects the design of a GWAS. The T2D phenotype is conventionally categorical in nature. This type of phenotype warrants the use of a case-control study design, where individuals are either classified as affected (“case”) or unaffected (“control”). The success of this type of design is attributed to a well-defined phenotype, usually aided by clinical variables. Ideally, with the exception of disease status, the cases and controls should be similar. They should be of similar sample size, be selected from the same population, and have similar sex and age distribution. All these specifications for a case-control study are there for the purpose of minimizing bias that can be introduced by stratification and/or exposure differences. It is also worth mentioning that controls should be randomly selected, in order to remove selection bias. To detect small effects of common variants, the statistical power of the study should be very high (≥0.8). Having a large enough sample size can really improve the statistical power.

The next step in the GWAS pipeline would be to select and genotype the single-nucleotide polymorphisms (SNPs) that would be used in the analyses. SNPs would be the markers of choice because millions of them have been well characterised across the genome, their inheritance patterns have been defined (Kruglyak, 2008; Wheeler and Barroso, 2011). Ideally the repertoire of SNPs selected should contain randomly selected SNPs that span the entire genome, tag SNPs, and SNPs singled out for functionality from previous GWAS’s. There should also be a substantial amount (> 300, 000) of SNPs selected to minimise cofounding by locus. Once SNPs are selected, the next step is to get their genotypes. When using large samples, genotyping tends to happen in multiple batches. In such cases, we can minimise ascertainment bias by adopting a blind genotyping approach; where the identity of the sample as a case or control is unknown. Sometimes samples could have missing genotypes, and this can be bypassed by imputation.

Once the data for genotypes and phenotypes have been sorted, the next step would be to perform a range quality control (QC) checks. There are some software that allow for some these QCs before running the analyses, for instance PLINK. Filtering out SNPs that violate Hardy-Weinberg Equilibrium helps minimise genotyping errors. Individuals and/or SNPs can be removed if the fall below a threshold for percentage missing data. SNPs with a minor allele frequency below 1% can be filtered out. Pairwise comparisons of the genotype data for genetic relatedness can also be done to correct for cryptic relatedness.

Data that pass through the QC filters can then be used in the association analyses. The appropriate statistical test for the T2D would be a logistic regression. An additive model should ideally be the model for the genetic risk, as the effect of each of the three genotypes is assessed. The result of this regression can then be used to construct a Manhattan plot, which will allow us visualise loci across the genome that significantly associate with T2D. An ideal statistical threshold for significance is the Bonferroni-corrected type I error rate (p < 5 x 10-8). The result of the analyses should then subject to additional statistical tests to check for cofounders. These tests include stratified analysis, principal component analysis, ad quantile-quantile plot of the p-values.

The latest GWAS for T2D as conducted using samples from the Japanese population (Imamura *et al*, 2016). There are some similarities and differences between the design of their study and the design discussed in this report. In harmony with our design, their study used involved the use of a very large sample size (as it was a meta-analysis), and a large number of SNPs. They also characterised the T2D phenotype using the World Health Organisation criteria. Their study also involved our recommended QCs of HWE testing, the use of 1% MAF threshold, and data with ≥ 99% call rate. They also assessed cryptic relatedness for each sample. However, there was a major difference from ours in the design of their case-control selection. They did not match the case and control data sets for potential cofounders. They did however control for these potential cofounders by performing stratified analysis, principal component analysis, and quantile-quantile plots p-values (Imamura *et al*, 2016). Overall, the design of their study was ideal, and very similar to the one we have described.

Part B

(i) Hardy-Weinberg equilibrium (HWE) is a term generally used to describe a population, whose genotype frequencies remain stable from one generation to another. The stability of the genotype frequencies is maintained so far there is no mutation, selection, migration, or non-random mating. In this model, the population is also assumed to be infinitely large.

When a population is in HWE, the frequency of alleles present in the population can be used to determine the proportions of genotypes. This can be expressed mathematically.

Let’s say we have a single nucleotide polymorphism with two alleles, A and B.

If the frequency (or proportion) of A is p, then the frequency of B is 1-p.

Note: 1-p is commonly denoted as q (which we would use from now on).

Given this representation, the frequency of homozygosity for A is p2, and for B is q2.

The frequency of heterozygosity is 2pq, since the A and B can be combined in two different ways (i.e AB and BA). The genotype frequencies have to some up to 1. So we end up with the equation below:

P2 + 2pq + q2 = 1

We can use these mathematical expressions to help calculate whether genotypes at a SNP site are in HWE. We can use the allele frequencies observed in a sample population to calculate the genotype frequencies we would expect, if they were in HWE.

We can then use a chi-squared test to see if the genotype frequencies we have observed is in good agreement (good fit) with the one we calculated (the expected).

Doing this, we would get a chi-squared statistic, which we can look up in the table of chi-squared values to get the estimated p value. If the p-value is less than 0.05, then there is a statically significant difference between the observed and expected genotype frequencies. The case is otherwise if the p-value is greater than 0.05.

The chi-squared test can be run computationally with R, using a R- package called ‘genetics’. This package has a function (called HWE.chisq) that can perform a chi-squared test for genotypes of a particular SNP. This function returns the chi-squared value, as well as a simulated p-value (based on a thousand replicates).

To calculate whether genotypes at 100 different SNP loci are in HWE, we first need to create a file containing the information each SNP in separate columns. The first row of the columns would be the name of the SNP sites (which serves as the header), and the genotypes for each SNP will occupy the rows beneath. An example is shown below

|  |  |  |
| --- | --- | --- |
| Rs456 | Rs576 | Rs689 |
| aa | cg | tc |
| at | cc | cc |

Figure x: Example of file containing information for 3 SNP sites.

This file will then be read into a variable in R, and attached to the search path.

Assuming the variable is named SNPdata, this is how the code for the above would look:

SNPdata <-read.table("C:/Users/Paul/Documents/file.txt", header=T)

attach(SNPdata)

We can then use the ‘genotype’ function of the genetics package to create an object that stores the allele and genotype frequencies of the SNP. We can then use the HWE.chisq function on the object. The code for these steps are written below:

Object <- genotype(SNPname, sep=””)

HWE.chisq(Object)

We can re-iterate this task with the different SNP names using a loop.

(ii) Linkage disequilibrium is the term used describe alleles at different loci that exhibit a non-random association. They are in disequilibrium if their frequency of association is higher (positive LD) or lower (negative LD) than expected.

LD is used in association studies for narrowing down the number of significant hits in attempt to find the causal variant.

It can be measured by the correlation coefficient of LD (D’), or R2.

To calculate LD for the 100 SNPs, you can use the R statistical software. The first step would be to create data frame of genotype objects. Genotype objects are created using he genotype() function of the genetics R-package. The genotype function will convert the column of a SNPs information to a new column where the genotypes of that particular SNP have been formatted. An example of the command is shown below

Install.packages(“genetics”)

Library(genetics)

rs3457geno <- genotype (rs3457, sep=””)

This data frame can then be used by the LDheatmap() function of the LDheatmap R-package, to construct an LD heatmap. You can modify the measurement type to a D’ or R2 using the LDmeasure modifier. You can label the map using the SNP.name modifier to identify which part of your data frame holds the information of the names of the SNPs, and the text = TRUE modifier for those names to be displayed. For example:

LDheatmap(frameOfgenotypes, LDmeasure = “r”, SNP.name = colnames(frameOfgenotypes), text = TRUE)

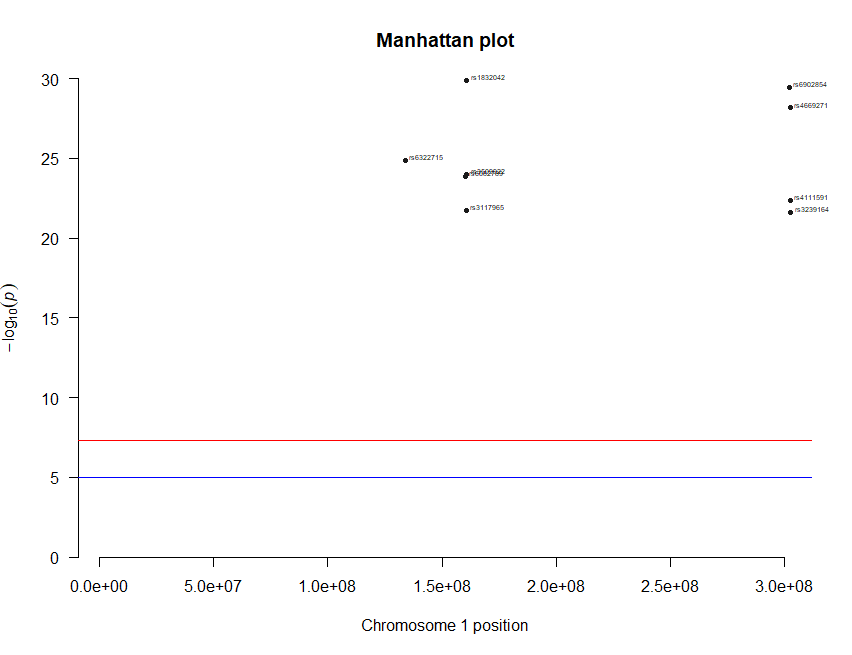
The LD heatmap would show you all the LD scores of each SNP.

(iii) Tag SNPs are SNPs that are used as a proxy for surrounding SNPs, as they sit in a region of the genome with high LD. We will be assessing 9 tag SNPs (located on chromosome 1) for association with a dichotomous phenotype using the PLINK software on a UNIX operating system. The statistical test run by PLINK for this analysis is a logistic regression under the additive model.

PLINK requires a bed, bim, and fam file run the analysis. These three files are made as per the specifications cited on the PLINK website. When running the analyses, we used some quality control filters to minimize the events of spurious association. These included a test for HWE (<1 x 10-6), and a minor allele frequency threshold (MAF, <1%)

The results of the analyses are summarised with plots and tables below. The minor alleles of these SNPs all appear to be risk alleles. The Manhattan plot below visualises the significance of their association to the phenotype. The red line is the Bonferroni-corrected type 1 error rate (5 x 10-8). The blue line indicates p-value of 0.0005.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CHR** | **SNP** | **BP** | **A1** | **TEST** | **NMISS** | **OR** | **SE** | **L95** | **U95** | **STAT** | **P** |
| 1 | rs6322715 | 1.34E+08 | C | ADD | 17105 | 1.753 | 0.05365 | 1.578 | 1.947 | 10.46 | 1.33E-25 |
| 1 | rs6082789 | 1.6E+08 | T | ADD | 17105 | 1.626 | 0.04746 | 1.481 | 1.784 | 10.24 | 1.32E-24 |
| 1 | rs3117965 | 1.61E+08 | A | ADD | 17105 | 1.49 | 0.04087 | 1.375 | 1.614 | 9.755 | 1.75E-22 |
| 1 | rs3509922 | 1.61E+08 | A | ADD | 17105 | 1.614 | 0.04659 | 1.473 | 1.768 | 10.27 | 9.56E-25 |
| 1 | rs1832042 | 1.61E+08 | T | ADD | 17105 | 1.626 | 0.04227 | 1.497 | 1.767 | 11.51 | 1.22E-30 |
| 1 | rs6902854 | 3.02E+08 | C | ADD | 17105 | 1.836 | 0.05321 | 1.654 | 2.038 | 11.42 | 3.33E-30 |
| 1 | rs4669271 | 3.02E+08 | A | ADD | 17105 | 1.831 | 0.05423 | 1.647 | 2.037 | 11.16 | 6.71E-29 |
| 1 | rs4111591 | 3.03E+08 | A | ADD | 17105 | 1.539 | 0.04352 | 1.413 | 1.676 | 9.9 | 4.15E-23 |
| 1 | rs3239164 | 3.03E+08 | T | ADD | 17105 | 1.507 | 0.04221 | 1.388 | 1.637 | 9.722 | 2.44E-22 |



Part C

This section of the report presents the findings, and investigative procedure, of genotypic association analyses for four sets of data. The results of these analyses are summarized using Manhattan plots, and quantile-quantile(QQ) plots (Figure 1, 2, 3, and 4). Each set of data contains the genotype of several single-nucleotide polymorphisms (SNPs) for an independent chromosome. The SNPs on two of these chromosomes are analysed for association to a quantitative trait, while those on the other two are analysed for association to a dichotomous trait.

**Association analyses**

The association analysis was run using the PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) software on a UNIX operating system. Various quality control (QC) filters were included as part of the analyses in an attempt to reduce bias introduced by poor quality DNA, and genotyping errors. SNPs with a minor allele frequency (MAF) < 0.01, a Hardy-Weinberg equilibrium p-value < 10-6, and a missing genotype rate > 0.1, were filtered from the analyses. Individuals with missing genotype rates > 0.05 were also filtered from the analyses. The statistical threshold for this study was set to the Bonferroni-corrected type on error rate (5 x 10-8).

Quantitative trait analyses

The quantitative trait was analysed using a linear regression under a model of additive allelic effect. A total of 22030 (10487 males, and 11543 females) individuals were used in this study, all of which passed the filters. 13430 SNP variants were analysed for chromosome 1, of which 13195 passed the filters (235 removed due to MAF threshold). 1626 SNP variants were analysed for chromosome 2, of which 1595 passed the filters (31 removed due to MAF threshold). After the running the analyses, PLINK provides a file with the results. This file was exported to the R- statistical software, and used to design a Manhattan plot (Figure 1a). This plot is the representation of the association analyses result. The R-software was also used to construct a QQ plot from the data (Figure 1b). This plot allows us to examine our association study results for systematic errors.

Five loci on chromosome 1 were observed to have significant associations with the quantitative trait (Figure 1). No statistically significant loci were observed on chromosome 2 (Figure 2a and b). There was a total of 130 statistically significant SNPs (hits) across chromosome 1 (table too large to be shown). The hits in the third and fifth locus and are amongst those with the lowest p-value and the highest effect size. The top hit in fifth locus (rs7278711) is also the top hit across the whole chromosome, with an effect size that falls within the confidence interval of the SNP with the highest effect size (rs7288230).

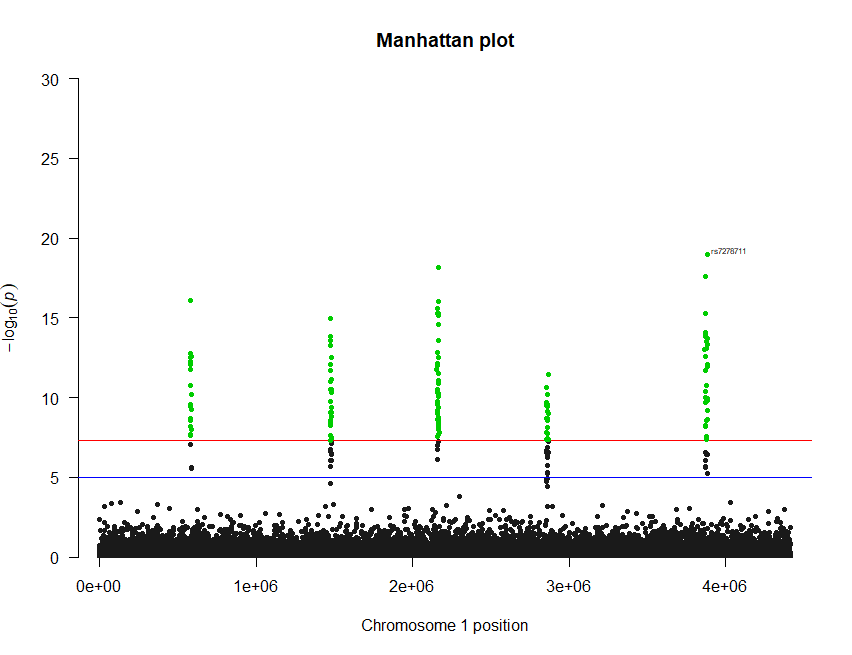


Figure 1a: Manhattan Plot of chromosome 1 for the quantitative trait analysis. The dots highlighted in green are those with p-values ≤ 5 x 10-8 (red line). The blue line represents the cut-off for the typically agreed p-value <0.00005.

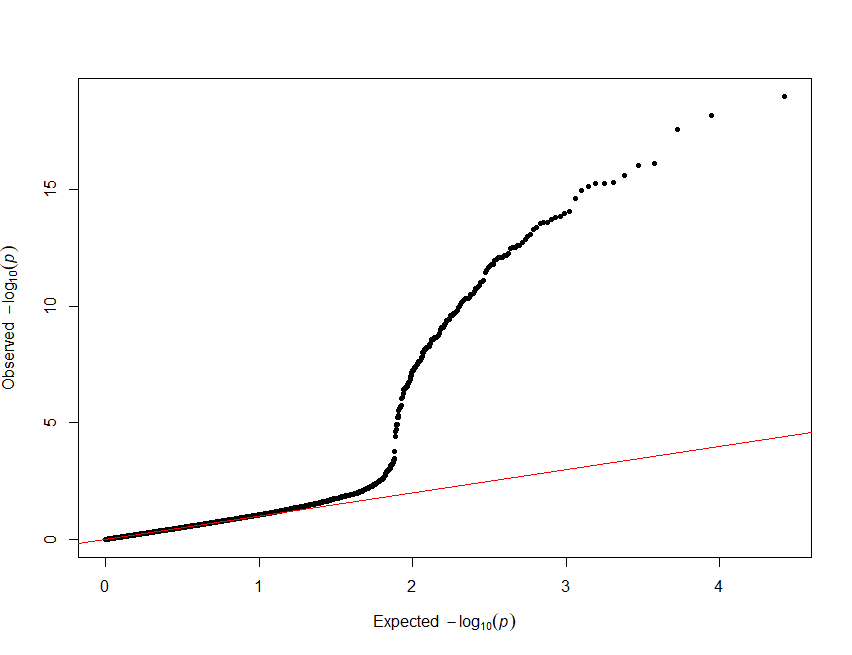


Figure 1b: a quantile-quantile plot of the p-values for the quantitative trait analyses (chromosome 1)

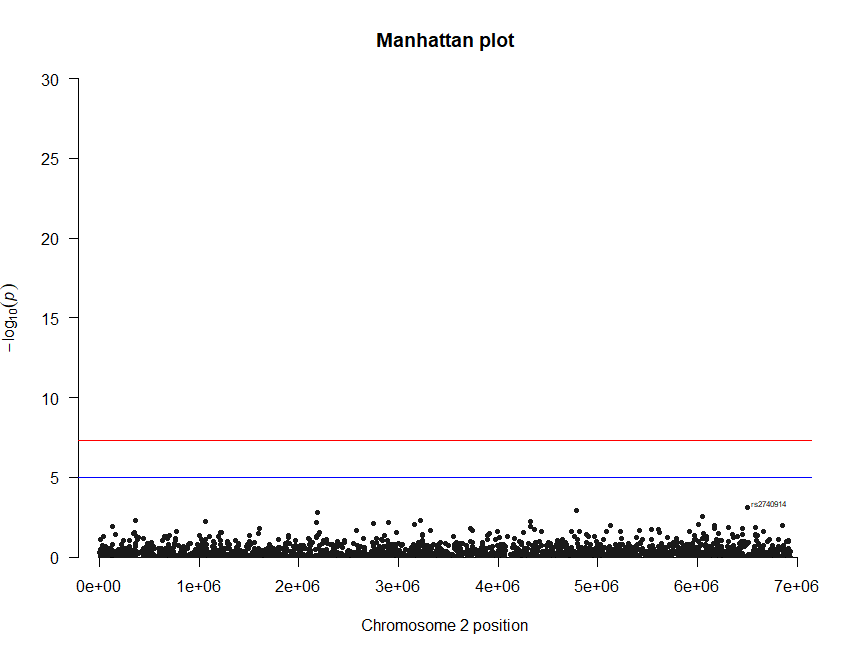


Figure 2a: Manhattan Plot of chromosome 2 for the quantitative trait analysis. The red line represents the cut-off with p-values ≤ 5 x 10-8. The black line represents the cut-off for the typically agreed p-value <0.0005.

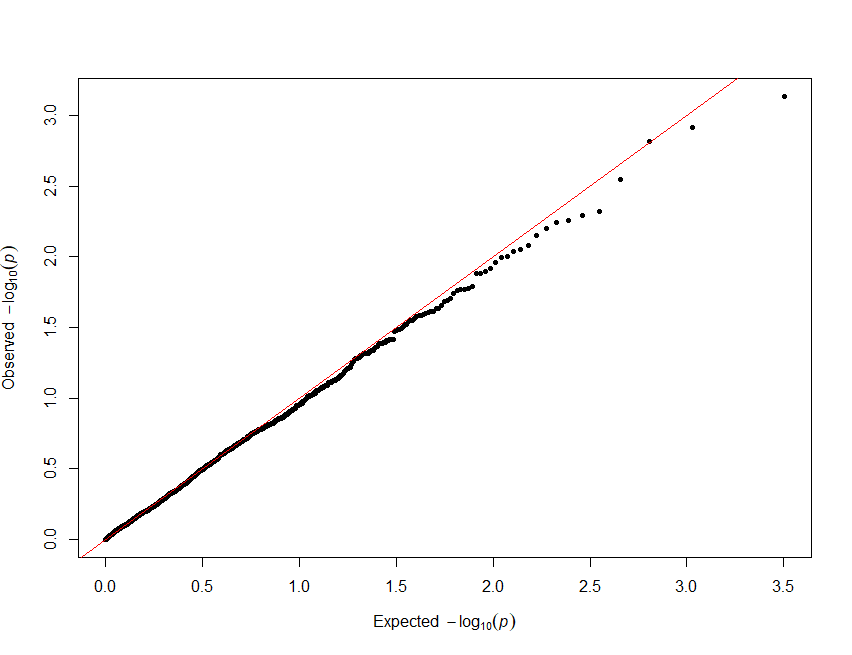


Figure 2b: a quantile-quantile plot of the p-values for the quantitative trait analyses (chromosome 2)

Dichotomous trait analyses

The dichotomous trait was analysed using a logistic regression under a model of additive allelic effect. A total of 17105 (11516 males, and 5589 females) individuals were used in this study, all of which passed the filters. Among these phenotypes, 3003 were cases, and 14102 were controls. 15744 SNP variants were analysed for chromosome 1, of which 15449 passed the filters (295 removed due to MAF threshold). 19075 SNP variants were analysed for chromosome 2, of which 18696 passed the filters (379 removed due to MAF threshold). After the running the analyses, PLINK provides a file with the results. The summary plots (Figure 3 and Figure 4) were generated in the same manner as for the quantitative trait study.

Three loci on chromosome 1, and two loci on chromosome 2, were observed to have significant association with the dichotomous trait. There was a total of 87 hits across chromosome 1, and a total of 59 hits across chromosome 2. The minor alleles of all the hits across both chromosomes were shown to increase the risk of the trait (table too large to be shown).

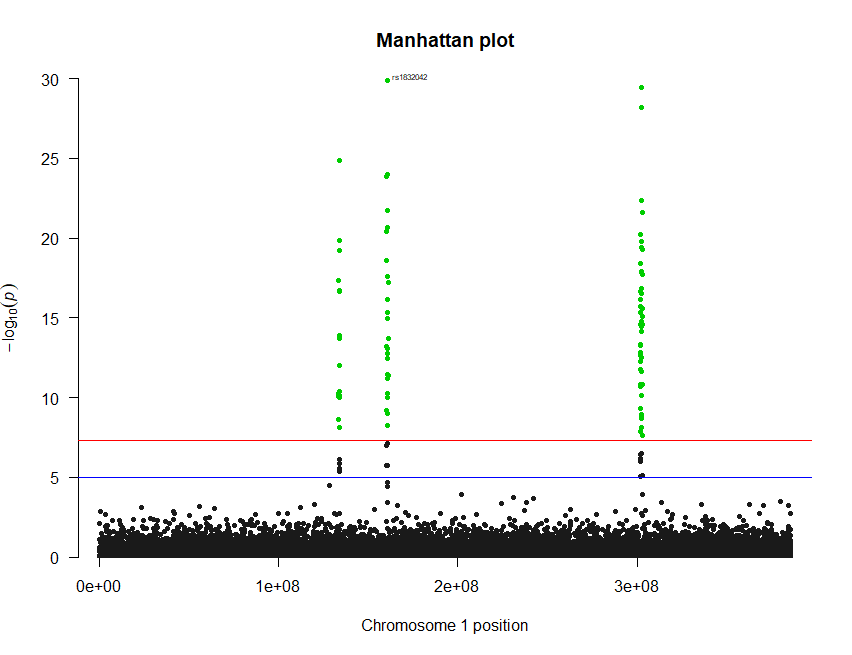


Figure 3a: Manhattan Plot of chromosome 1 for the dichotomous trait analysis. The dots highlighted in green are those with p-values ≤ 5 x 10-8 (red line). The blue line represents the cut-off for the typically agreed p-value <0.00005.

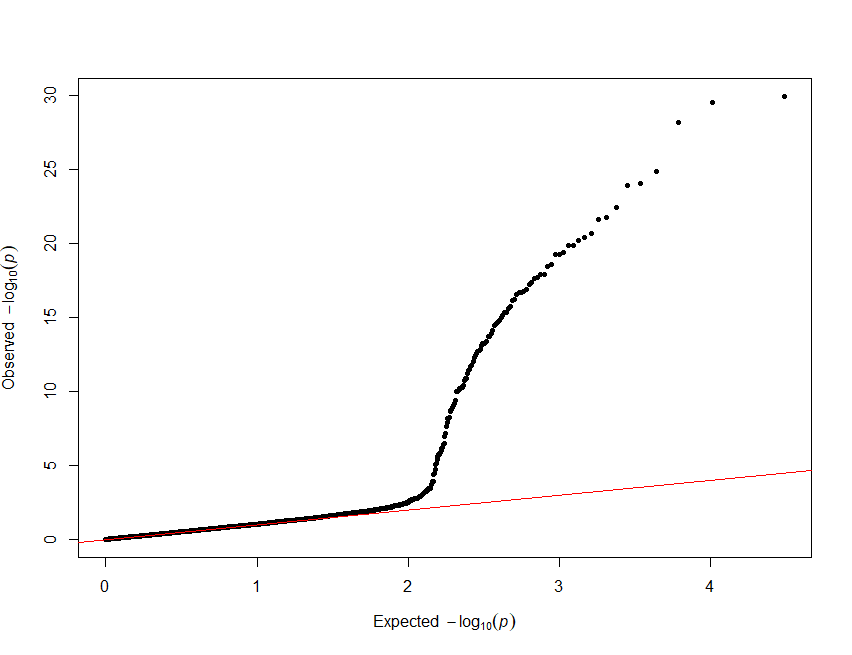


Figure 3b: a quantile-quantile plot of the p-values for the dichotomous trait analyses (chromosome 1)

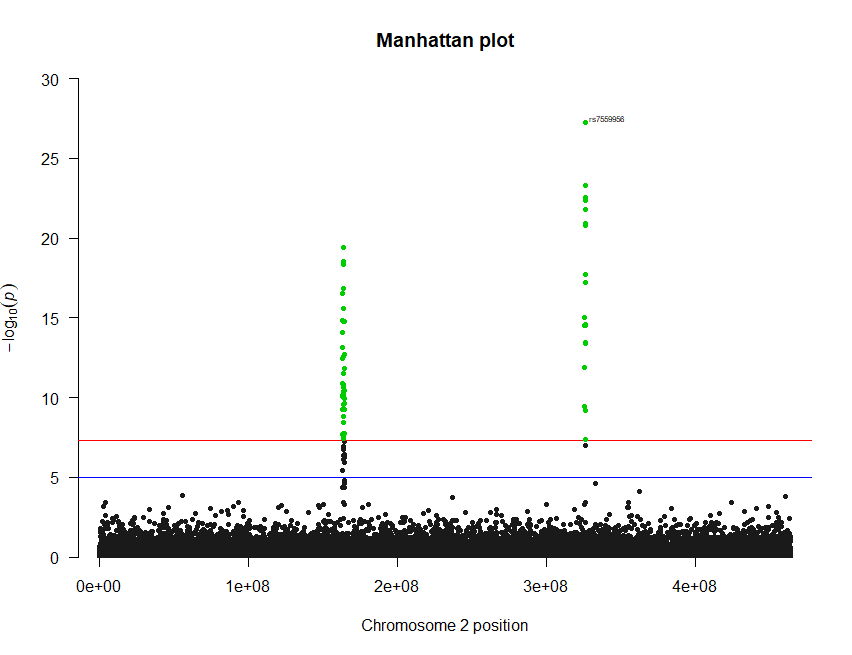


Figure 4a: Manhattan Plot of chromosome 2 for the dichotomous trait analysis. The dots highlighted in green are those with p-values ≤ 5 x 10-8 (red line). The blue line represents the cut-off for the typically agreed p-value <0.00005.

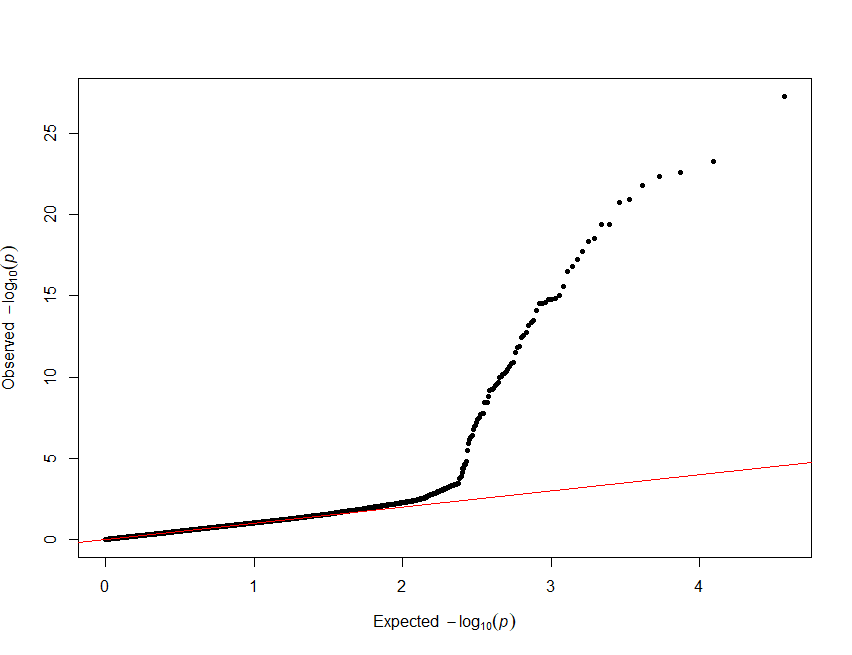


Figure 4b: a quantile-quantile plot of the p-values for the dichotomous trait analyses (chromosome 2)

**Single locus analysis**

Further analysis on the loci of interest was hindered by the fact that the data sets used were simulated. If this was not the case, single SNPs could have been selected from each locus to establish patterns of linkage disequilibrium around its region. The selection of the SNP for regional LD analysis would have been based on a combination of its p-value and effect size. The pattern of LD can be observed with the use of a LD heatmap (Figure 5). This diagram helps visualise the pairwise LD measurements between SNPs in a specified genomic region. It allows for the observation of blocks of LD across a region. Haploview (Barret *et al*, 2005) or the R-statistical software can be used to construct the heat map. To perform this in R, two files are needed. The first file is a matrix that contains the r2 values of the LD pairwise comparisons. The second file is one that contains the locations of the SNPs in the matrix. These can be used as parameters in the LDheatmap() function.

Assuming a reasonable heat map could be generated with our data, I would select (“tag”) SNPs in the region that have a strong LD (r2 ≥ 0.5), and for a set of tagged SNPs. I would also make use an online tool, called locus zoom (See figure 5), to create plots of the region. These plots help define the region of the tagged SNPs. These regions can be examined for functionality. Further analysis to help identify causal may also include calculating LD between the tagged SNPs and rare variants.

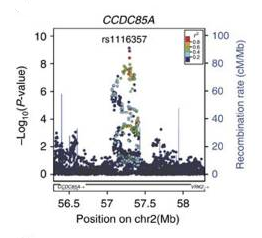
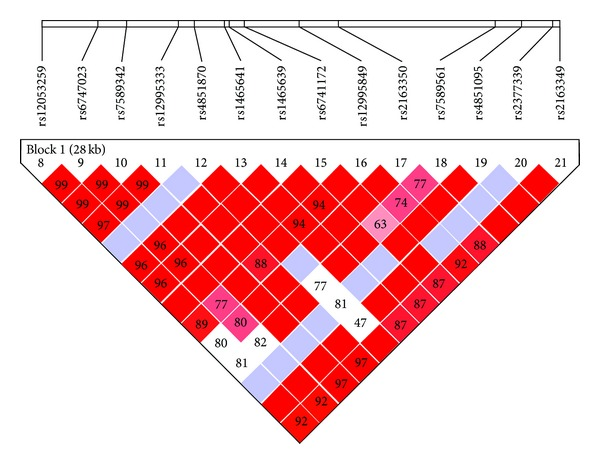


Figure 5: Example an LD heatmap (left) [adapted from Liu *et al*, 2013] and a locus zoom plot(right) [adapted from Imamura *et al*, 2016].

**Validity of Results**

Additional statistical tests were conducted to ensure that the results of the association analyses were valid. The main factor being curbed is the event of a spurious association. These types of association are more often than not the result of systematic errors. As mentioned earlier, quality control filters were utilised to minimize genotyping errors., while QQ plots were used to explore other sources of systematic error. The QQ plots for both analyses appear to be in agreement with their corresponding Manhattan plots. The nature of the deviation from the identity line, observed in figure 1, 2, 3, and 4, suggest that only a subset of the SNPs analysed contribute to the statistical significance. In figure 2a, where no hits occurred, very minute deviation from the identity line is observed. This signifies that no association was more statistically significant than we would expect.

There are however some limitations to the analyses. The population which was sampled for this study is unknown. If the sample was from a sub-population in an area where population stratification is at play, some of the hits in this analyses may not be consistent with other populations. The age of the individuals in this study is also unknown. If they were, we could observe if there is stratification according to median age. It may also have been useful to observe stratification according to sex.

Part D

In this section, four variants present on the interleukin 23 receptor (IL23R) gene were examined for functionality. These variants were selected using the Exome Aggregation Consortium (http://exac.broadinstitute.org) browser. As indexed by ExAC, these variants are 1:67685387 T/C, 1:67633812 G/T, 1:67633862 G/A, and 1:67666586 G/T. In this report we refer to them as var1, var2, var3, and var4, respectively. These SNPs were examined primarily with the use of the Ensembl Variant Effect Predictor (VEP). This tool was used to predict the functional effects of these variants on transcripts, protein sequence, and regulatory regions at or near the IL23R gene. The results of this analyses are summarised in table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SNP ID | SNP | Consequence | Impact | SIFT | PolyPhen-2 |
| Var 1 | 1:67685387 T/C | missense variant | Moderate | 1 | 0 |
| Var 2 | 1:67633812 G/T | missense variant | Moderate | 0.19 | 0.005 |
| Var3 | 1:67633862 G/A | stop-gained | High | N/A | N/A |
| Var 4 | 1:67666586 G/T | splice region/intron variant | Low | N/A | N/A |

Table 1: Summary of VEP analyses.

**Results of VEP and other secondary analyses**

Var1 and var2 are variants common in most of the world’s populations. Var1 was found to present on seven transcripts in two genes, IL23R and c1orf141. The consequence of this variant on the IL23R protein coding region is a missense mutation. VEP classified the severity of this missense variation as moderate. Var1 also appeared to be a non-coding transcript exon variant on an IL23R processed transcript; the severity of which was classified as a modifier. The missense variant causes a leucine to proline change at position 310. The possible impact of this change on the structure and function of the protein was assessed using PolyPhen-2 online tool (http://genetics.bwh.harvard.edu/pph2/). The mutation was predicted to be of benign effect, with a score of 0.000. The multiple sequence alignment (MSA), produced along with score, shows that proline at that position is well conserved across many species (data not shown). Another online algorithm, called SIFT (http://sift.jcvi.org/), validated the prediction by PolyPhen-2; predicting the mutation to be tolerable, with an equivalent maximum score of 1. Given that c1orf141 is an uncharacterized gene, no further analysis carried out beyond the VEP.

Var2 was found to be present on four transcripts in the same two genes as var1. The consequence of this variant on the IL23R protein coding region, as well as its severity, is the same as for var1. The missense variation of var2 causes a glutamine to Histidine change at position 3. Both PolyPhen-2 and SIFT predicted this change to be of little impact. The MSA presented by PolyPhen-2, however, appeared to be inconclusive (data not shown).

Var3 and var4 are quite rare variants. Var3 was to be present on four transcripts in the same genes as var1 and var2. The consequence of this variant on IL23R protein coding region is a nonsense mutation. The severity of this mutation was classified as high (very severe). This variant causes the change of a tryptophan codon to a stop codon at position 20. Var4 was present on four transcripts in three genes, IL23R, c1orf141, and RNU6-586P. The consequence on IL23R protein coding region is a splice region mutation located on an intron. The severity of this mutation was classified as low.

**Concluding Remarks**

Several genome-wide analysis studies (GWAS’s) have revealed an association between the IL23R gene and Inflammatory Bowel disease (IBD) (Yu *et al*, 2012; Fischer *et al*, 2017). The analyses that I have conducted has revealed predictions of the functional effects of four variants within this gene. Observing whether any of the variants could potentially explain the association of IL23R with IBD was in my prospective.

Var1 and var2 were predicted to have little to no effect on the function of IL23R. Association studies between these variants and IBD have actually been conducted by others (Yu *et al*, 2012; Fischer *et al*, 2017). These variants have been found to be associated with Ulcerative colitis (one of the two most common IBDs). However, only var1 shows any association with a Bonferroni corrected p-value (Yu *et al*, 2012). Even if the correction is overlooked, the associations show population stratification (Yu *et al*, 2012; Fischer *et al*, 2017). Thus, in combination with findings of these other studies, the results of my analyses suggest that var1 and var2 might play a redundant role in the association of IL32 to IBD. The same suggestion can be made about var4, as its effect on IL23R’s function is predicted to be low.

Var3 was predicted to have a big effect on the function of IL23R. This is no surprise, as stop codons within open-reading-frames (ORFs) tend to cause truncated proteins or no protein at all (as a result of nonsense-mediated decay). According to the information available on ExAC, var3 has only been recorded in populations of African descent, at very low frequencies. Given this, var3 is probably a lethal variant.

References

Ali O. (2013). Genetics of type 2 diabetes. *World Journal of Diabetes*, 4(4), 114—123.

American Diabetes Association. (2010). Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*, 33(Suppl 1), S62—S9.

Barrett JC, Fry B, Maller J, and Daly MJ. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. [PubMed ID: 15297300]

Fischer S, Kövesdi E, Magyari L, Csöngei V, Hadzsiev K, Melegh B, Hegyi P, and Sarlós P. (2017). IL23R single nucleotide polymorphisms could be either beneficial or harmful in ulcerative colitis. *World Journal of Gastroenterology*, 23(3), 447—454.

Imamura M, Takahashi A, Yamauchi T, Hara K, Yasuda K, Grarup N, Zhao W, Wang X, Huerta-Chagoya A, Hu C, Moon S, Long J, Kwak SH, Rasheed A, Saxena R, Ma RC, Okada Y, Iwata M, Hosoe J, Shojima N, Iwasaki M, Fujita H, Suzuki K, Danesh, Jørgensen , Jørgensen ME, Witte DR, Brandslund I, Christensen C, Hansen T, Mercader JM, Flannick J, Moreno-Macías H, Burtt NP, Zhang R, Kim YJ, Zheng W, Singh JR, Tam CH, Hirose H, Maegawa H, Ito C, Kaku K, Watada H, Tanaka Y, Tobe K, Kawamori R, Kubo M, Cho YS, Chan JC, Sanghera D, Frossard P, Park KS, Shu XO, Kim BJ, Florez JC, Tusié-Luna T, Jia W, Tai ES, Pedersen O, Saleheen D, Maeda S, Kadowaki T. (2016). Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nature Comunications*, 10531. doi: 10.1038/ncomms10531.

Kruglyak L. (2008). The road to genome-wide association studies. *Nature Reviews Genetics*, 9(4), 314—318.

Liu Z, Guo X, Jiang Y, and Zhang H. (2013). NCK2 Is Significantly Associated with Opiates Addiction inAfrican-Origin Men. *The Scientific World Journal*, (4): 748979.

Wheeler E, and Barroso I. (2011). Genome-wide association studies and type 2 diabetes. *Briefiengs in Functional Genomics*, 10(2), 52—60.

Yu P, Shen F, Zhang X, Cao R, Zhao X, Liu P, Tu H, Yang X, Shi R, and Zhang H. (2012). Association of single nucleotide polymorphisms of IL23R and IL17 with ulcerative colitis risk in a Chinese Han population. *PLoS One*, 7(9).