Association Mapping Approach into Type 2 Diabetes using Biomarkers and Clinical Data

B. Abdulaimma[1], A. Hussain[1], P. Fergus[1], D. Al-Jumeily[1], C. Aday Curbelo Montañez[1], J. Hind[1], N. Radi[2]

[1]Liverpool John Moores University, Applied Computing Research Group, Faculty of Engineering and Technology, Byrom Street, Liverpool, L3 3AF, UK.
B.T.Abdulaimma@2015.ljmu.ac.uk
{A.Hussain, P.Fergus, D.Aljumeily }@ljmu.ac.uk
C.A.Curbelomontanez@2015.ljmu.ac.uk
J.Hind@2012.ljmu.ac.uk
[2]Al-Khawarizmi International College, Abu Dhabi, UAE.
n.radi@khawarizmi.com

Abstract. The global growth in incidence of Type 2 Diabetes (T2D) has become a major international health concern. As such, understanding the aetiology of Type 2 Diabetes is vital. This paper investigates a variety of statistical methodologies at various level of complexity to analyse genotype data and identify biomarkers that show evidence of increase susceptibility to T2D and related traits. A critical overview of several selected statistical methods for population-based association mapping particularly case-control genetic association analysis is presented. A discussion on a dataset accessed in this paper that includes 3435 female subjects for cases and controls with genotype information across 879071 Single Nucleotide Polymorphism (SNPs) is presented. Quality control steps into the dataset through pre-processing phase are performed to remove samples and markers that failed the quality control test. Association analysis is discussed to address which statistical method can be appropriate to the dataset. Our genetic association analysis produces promising results and indicated that Allelic association test showed one SNP above the genome-wide significance threshold of 5×10^{-8} which is rs10519107 (Odds Ratio $(\mathrm{OR}) = 0.7409, \mathrm{P}-\mathrm{Value}\ (\mathrm{P}) = 1.813\times10^{-9}$), While, there are several SNPs above the suggestive association threshold of 5×10^{-6} these SNPs could worth further investigation. Furthermore, Logistic Regression analysis adjusted for multiple confounder factors indicated that none of the genotyped SNPs has passed genome-wide significance threshold of 5×10^{-8} . Nevertheless, four SNPs (rs10519107, rs4368343, rs6848779, rs11729955) have passed suggestive association threshold.

Keywords: Genetics, Genome-Wide Association Studies (GWAS), Logistic Regression Model, P Values, Single Nucleotide Polymorphism (SNP), Type 2 Diabetes (T2D).

adfa, p. 1, 2011.

1 Introduction

Currently, the prevalence and the incidence of Type 2 Diabetes (T2D) throughout the world are increasing at an alarming rate. The International Diabetes Federation (IDF) has estimated that the number of diabetic people is expected to rise from 366 million in 2011 to 552 million by 2030 worldwide [1]. Type 2 Diabetes is a multifactorial disorder and is the result of the complex interaction between genetic, environment and sedentary lifestyle [2], however, genetic susceptibility has been established as a key component of risk [3]. Twins studies have exposed that the concordance rate of T2D in monozygotic twins is approximately 70% compared with 20% to 30% in dizygotic twins [4].

The major tools for identifying disease susceptibility loci are genetic variations which are termed as Single Nucleotide Polymorphism (SNP). SNP is a single base-pair change in the genetic code and it is the main cause of human genetic variability [5].

Genome-wide association studies (GWAS) have been widely used and specifically developed for investigating the genetic architecture of human disease in the entire genome [6]. The ultimate aim of GWAS is to identify the genetic risk factors for common complex diseases such as Type 2 Diabetes, Schizophrenia, Epilepsy, Obesity, Cardiovascular Disease, and Hypertension [6]. GWAS becomes more routinely employed with increase the availability of less expensive genotyping technologies [7]. The identification of genetic markers that show evidence of increase susceptibility to T2D and related traits are important to advance and facilitate the translation of this genetic information into clinical practice [8]. This advance may help to improve risk prediction [9] of the disease and delay or prevention of disease onset and to mitigate cares expenditures [10]. However, to understand the aetiology of such complex diseases, genetic information solely would not be sufficient without considering the non-genetic factors [11]

There are several statistical methods for association mapping including allelic test, genotypic test, dominant test, recessive test, Cochran Armitage trend test, Fisher exact test, and Logistic regression test [12]. However, it is difficult to specify which association tests to use [13]. It would be ideal to design optimal analyses based on the knowledge about the penetrance patterns of predisposing variants such as additive effect, dominant or recessive effects. Lacking this knowledge forces investigators to use their judgment [13].

This paper considers a case-control study design to conduct several classes of association analysis including; chi-square test based on (Allelic test, Genotypic test, Dominant test, and Recessive test), and Logistic regression. Logistic regression is the preferred approach to perform association analysis as it can readily expand to include covariates such as clinical variable, sociodemographic and environmental factors. Using genetic association analysis would facilitate the investigation of genetic markers that manifest themselves as candidate to increase susceptibility to T2D. These findings provide starting points to researchers and professionals to investigate further and to provide better understanding to the disease onset and advance the development of medical therapies.

2 Background

Understanding the aetiology of complex diseases such as T2D that is caused by the contribution of genetic and non-genetic risk factors is challenging [14]. The development of genetic association mapping has facilitated the discovery of genetic markers predisposing to complex diseases as T2D. Recently, several GWAS studies accompanied with various statistical methods have been performed in different cohorts and/or ethnic groups, to measure the association of genetic variants (loci) to disease susceptibility and to test for statistical significant (p-value). A series of publications have addressed various aspects and strategies into T2D genetics studies to be available within the literature for further investigations.

In [15], the authors performed a case-control study to investigate the differences in association of peroxisome proliferator activated receptor, gamma, coac- tivator 1 alpha (PPARGC1A) gene with T2D risk among population with African origins. The study includes adults aged >30 years old from African Americans (cases = 124, controls = 122) and Haitian Americans (cases = 110, controls = 116). The statistical method used within this study was Chi-squared goodness-fit test that was employed to check genotype counts for each SNP for Hardy-Weinberg Equilibrium. Furthermore, the t-test was used to compare between cases and controls considering demographic (age, sex, BMI, smoking status) and clinical information. Logistic regression approach was also used to calculate adjusted and unadjusted Odds Ratio (OR) with 95% confidence interval (CI). The result indicated that SNP rs7656250 (OR = 0.22, p-value = 0.005) and rs4235308 (OR = 0.42, p-value = 0.026) showed protective association with T2D in Haitian Americans. While in African Americans, SNP rs4235308 (OR = 2.53, p-value = 0.028) showed significant risk association with T2D.

While, in [16] the association analysis was performed on a case-control study to investigate the role of the mutation of KCNJ11 gene (potassium inwardly-rectifying-channel, subfamily-J, member 11) particularly E23K polymorphism (rs5219) in susceptibility to T2D. In this study, 56,349 T2D cases, 81,800 controls, and 483 family trios were collected from 48 published studies. The statistical methods used within the approach included Standard Q-statistic test, subgroup analysis (ethnicity, sample size, BMI, age and sex) were utilized to explore whether the variation in these studies was due to heterogeneity. Furthermore, the odds ratio with its 95% confidence interval of KCNJ11 E23K polymorphism was calculated to measure the association with T2D. Dominant and Recessive genetic models were applied to examine the association of KCNJ11 E23K polymorphism and T2D risk. The result suggested that KCNJ11 E23K allele of rs5219 (OR = 1.12, $p < 10^{-5}$) was significantly associated with T2D risk. For heterozygous and homozygous allele with (OR = 1.09, $p < 10^{-5}$) and (OR = 1.26, $p < 10^{-5}$) respectively, significant increase of T2D risk was observed. For Dominant and Recessive genetic models, similar results were obtained. This study suggested that a modest but statistically effect of the 23K allele of rs5219 polymorphism in susceptibility to T2D, particularly in East Asians and Caucasians. However, the contribution of these genetic variations to T2D in other ethnic populations (e.g. Indian, African, American, Jews, and Arabian) appears to be relatively low.

Genetic association studies are becoming an important approach for identifying genes particularly SNPs conferring susceptibility to complex diseases. The findings of Disease-SNP associations have been reported consistently using various statistical analysis methods that calculate statistical significant of the SNPs and measure the strength of the association in the study.

3 Materials and Methods

This section provides description of the dataset that is used in this paper and illustrates the quality control steps taken to pre-process that dataset. This section also describes the concept of genetic association analysis and provides in depth information related to statistical methods that is used in this domain.

3.1 Data Description

The Nurses' Health Study (NHS) cohort data set is used in this paper and it is provided by the Database of Genotypes and Phenotypes (dbGap) [17]. The NHS was established in 1976. Participants were 121,700 female registered nurses between age 30 to 55 and residing in 11 U.S states. All nurses responded to mailed questionnaire requesting information related to their medical history and lifestyle characteristics. Since then, the Nurses have been requested twice a year to fill questionnaire and attain updated information (for instance information on newly diagnosed illness). Furthermore, all participants were requested to provide blood samples, in which 32,826 members responded. The cases and controls participants were selected form the NHS T2D study. DNA of cases and controls participants were genotyped using the Affymetrix Genome-Wide Human 6.0 array. The ultimate version of the dataset includes 3435 female subjects for cases and controls with genotype information across 879071 SNPs. Participants in this dataset are identified as Hispanic or non-Hispanic and each belong to one of four racial categories (White, African-American, Asia or Other). Most participants are White and non-Hispanic representing (97.4%) of the dataset. The NHS dataset also includes corresponding clinical and dietary data, such as age, gender, BMI, alcohol intake, smoking status, physical activity, medical and family history.

3.2 Data Preprocessing

In this paper, the accessed genetic data is in PLINK format. PLINK v1.07 [18] is a whole genome data analysis toolset which is developed for handling SNP data. The files in PLINK format are very large and could cause issues with computational performance. As such, we convert these files to binary format using PLINK 1.07 toolset. Transferring to a binary formatted file, resulting in a considerable reduction in file size and significantly enhancing computational efficiency. This step is important for pre-

paring the dataset for quality control and filtering procedures. We performed data quality control for individuals and genetic data to produce a subset of reliable genetic markers and samples to be used for association analysis phase. Firstly, this study has been restricted to White and non-Hispanic ancestry to reduce potential bias due to population stratification. We removed data samples which have been reported with discordant sex information and duplicated or related individuals. Quality control for genetic markers was considered to remove genetic markers (SNPs) with > 0.1 missing data and with Minor Allele Frequency (MAF) of <0.05. We further conducted Hardy-Weinberg Equilibrium (HWE) and discarded those SNPs with a p-value <0.001 in control samples. Following the quality control steps, 3255 individuals and 665092 markers remained in the study from the original sample of 3435 and 879071, respectively.

3.3 Association Analysis

An association analysis of a case-control study aims to compare the frequency of alleles or genotypes at genetic marker loci (SNP) between cases and controls from a given population. This analysis will detect if there are any differences in the frequency of alleles between individuals in the study. The testing leads to determine whether the difference in alleles' frequency is statistically significant. In this situation that alleles (genetic marker) can be recognized as to be associated with the phenotype (disease trait) [19]. In other words, association analysis is a series of single-locus statistics tests, exploring each SNP separately for association to the phenotype.

In a case-control design study, the association between a single SNP and disease status can be based on standard contingency table tests for independence [13]. contingency table is widely used to display genetic marker (SNP) in the format of genotype or allele frequency by disease status (case-control) [19]. Each single SNP consists of minor allele a and major allele A among case and control groups and these can be represented as a contingency table of the disease status by either genotype count (e.g. aa, Aa and AA) with dimension of 2 x 3 of 2 degrees of freedom (d.f.) or allele count (a and A) with dimension of 2 x 2 of 1 d.f. The genetic data can also be analyzed assuming a prespecified genetic model, as contingency table allows for different models of disease penetrance such as dominant model and recessive model. For example, the contingency table of dominant model of penetrance can be summarized as a 2 x 2 table with 1 d.f. of genotype count of AA versus Aa or aa as any number of copies of minor allele a increase the risk of disease. While to test for a recessive model of penetrance, the contingency table is represented as 2 x 2 table with 1 d.f. requiring two copies of minor allele a to increase the risk of disease as the genotype count of recessive model is aa versus the combined count of Aa and AA [20].

The calculation of degrees of freedom is based on the inheritance models in which representing by genotypic, allelic, recessive and dominant [20]. Therefore, the degrees of freedom of genetic model is calculated based on the (number of rows in the contingency table – 1) × (number of columns in the contingency table – 1) [21]. For example, for allelic test where the number of both rows and columns is 2, the degrees of freedom is $(2-1)\times(2-1)=1$.

The contingency table for case and control analyses using different genetic model of penetrance has been summarized in Table 1, where DF represents degrees of freedom.

Whereas O_{ij} refers to the observed frequency of individuals in cases and controls, i refers to row number and j to column number. For example, in genotypic model test O_{11} refers to the observed frequency of individuals in cases when genotype aa occurs.

Test	DF	Contingency table representation			
			aa	***Aa***	***AA***
Genotypic test	2	**Cases**	O_{11}	O_{12}	O_{13}
		Controls	O_{21}	O_{22}	O_{23}
			AA	***Aa* or *aa***	
Dominant model	1	**Cases**	O_{11}	O_{12}	
		Controls	O_{21}	O_{22}	
			aa	***Aa* or *AA***	
Recessive model	1	**Cases**	O_{11}	O_{12}	
		Controls	O_{21}	O_{22}	
			a	***A***	
Allelic test	1	**Cases**	O_{11}	O_{12}	
		Controls	O_{21}	O_{22}	

Table 1. Contingency Table for Different Genetic Models

Practically the association test within genetic data of case and control status is to test the null hypothesis of no association between the SNP and phenotype of interest (disease status) in the contingency table. Pearson's chi-squared test (x^2) can be used to test for association. The principle of chi-squared test (x^2) is to compare the distributions of observed and expected values of their contingency tables [22]. Chi-square test summaries the differences between the observed frequency values and the expected frequency values at a single genetic marker loci (SNP) across cases and controls.

The following equation presents the standard Chi-square test for independence of rows and columns in the contingency table considering genotypic model for association [20]:

$$x^2 = \sum_i \sum_j \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad (1)$$

Where E_{ij} is the expected frequency of allele or genotype in case and control and O_{ij} refers to observe frequency of individuals.

Following the calculation of Chi-Square test, the p-value for Chi-Square is determined based on the degrees of freedom of the test if it has 1 or 2 degrees of freedom. The p-value is a measure of the significance of the Chi-squared test. Formally, the p-value is defined as the probability of seeing a value of test statistic (chi-square statistic test) as equal to or larger than the one that was observed in a given dataset, assuming the null hypothesis (no association) is true [6]. More specifically, the p-value represents the degree of association between the SNP and the phenotype across the entire sample

set. This means that lower p-value indicates that it is unlikely for the results to occur under the null hypothesis of (no association) [6].

logistic regression is defined as a statistical method for predicting binary outcome [19]. Logistic regression model can be used to analyze the contingency table for independence, where disease status accounts as binary traits (0/1) for case and control.

Logistic regression approach can be easily expanded to allow for covariates including further SNPs, sociodemographic and clinical factors. In a case-control study, the strength of an association is measured by the odds ratio (OR) [19]. Odds ratio is the ratio of the odds of disease in exposed group (cases) compared with non-exposed group (controls) [19]. For example, based on the variables provided from Table 1, the allelic OR measure the association between disease and allele considering the odds of disease if allele A (major allele) is carried in compared to the odds of disease if allele a (minor allele) is carried. The following formula is used to estimate the allelic OR for allele A [23].

$$OR_A = \frac{odds\ of\ disease\ with\ A\ allele}{odds\ of\ disease\ with\ a\ allele} \tag{2}$$

The strength of association of allele A is estimated based on the value of OR. Therefore, OR's value equal to 1 indicates no association, more than 1 indicates an association, and less than one indicates protective association.

3.4 Association Analysis of Geneva NHS Dataset

We conducted a case-control association analysis in an unrelated, white and non-Hispanic racial subpopulation to compare the frequency of alleles or genotypes at genetic marker loci (SNP) between cases and controls of Geneva NHS Dataset. The association analyses were performed using PLINK v1.07. We calculated the odds ratio with its 95% confidence interval (95% CI) to evaluate the strength of association between SNPs and T2D. Pearson's chi-squared test (x^2) was used to test the null hypothesis of no association. We conducted Allelic association test to explore the association between single allele of the SNP and the disease trait (Type2 Diabetes). Furthermore, genetic associations were also assessed using logistic regression methods were performed to calculate adjusted odds ratio with its 95% CI to assess the association of all SNPs in the study with disease status of binary traits (0/1) for case and control. Logistic regression was adjusted for covariate including (age, BMI, smoking status and physical activity) to examine the differences in the results that occur when the test based on a model accounting for non-genetic risk factors.

4 Results

Allelic association test's result suggested that there are at least one SNP above the genome-wide significance threshold of 5×10^{-8} While, there are several SNPs above the suggestive association threshold of 5×10^{-6}. Manhattan plot has been used to visualize

the results of the association as represented in Fig. 1(a). Allelic test indicated that SNPs rs10519107 ($OR = 0.7409, P = 1.813\times10^{-9}$), rs809736 ($OR = 0.7461, P = 7.627\times10^{-7}$), rs810517 ($OR = 0.7904, P = 2.682\times10^{-6}$), rs12571751 ($OR = 0.7913, P = 2.975\times10^{-6}$), rs10181181 ($OR = 0.7738, P = 3.908\times10^{-6}$), rs1020731 ($OR = 0.7765, P = 4.882\times10^{-6}$) showed protective association with T2D. Significant associations were detected in allelic test with SNPs rs4368343 ($OR = 1.890, P = 9.916\times10^{-7}$), rs6848779 ($OR = 1.2760, P = 1.578\times10^{-6}$), rs11729955 ($OR = 1.2750, P = 1.812\times10^{-6}$), rs11701035 ($OR = 1.3480, P = 2.736\times10^{-6}$). Table 2, demonstrated SNPs above suggestive threshold $< 10^{-6}$ with their OR and the p-value.

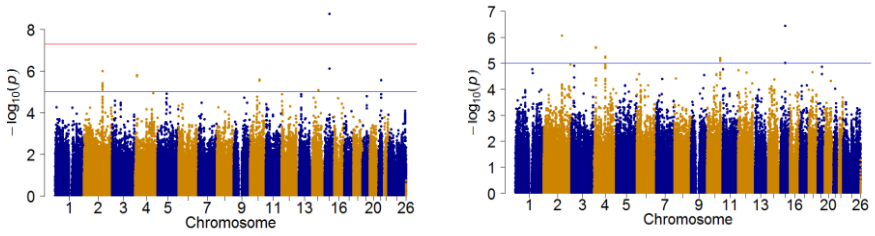
CHR	SNP	P-value	OR	Association
15	rs10519107	1.813×10^{-9}	0.7409	Protective
15	rs809736	7.627×10^{-7}	0.7461	Protective
2	rs4368343	9.916×10^{-7}	1.2890	Association
4	rs6848779	1.578×10^{-6}	1.2760	Association
4	rs11729955	1.812×10^{-6}	1.2750	Association
10	rs810517	2.682×10^{-6}	0.7904	Protective
21	rs11701035	2.376×10^{-6}	1.3480	Association
10	rs12571751	2.975×10^{-6}	0.7913	Protective
2	rs10181181	3.908×10^{-6}	0.7738	Protective
2	rs1020731	4.882×10^{-6}	0.7765	Protective

Table 2. SNPs with the Suggestive of Association From Allelic Test

Logistic Regression analysis adjusted for multiple confounder factors suggested that none of the genotyped SNPs has passed genome-wide significance threshold of 5×10^{-8} as represented in Fig. 1(b). Nevertheless, the result also indicated that the SNP rs10519107 ($OR = 0.7599, P = 3.583\times10^{-7}$) showed protective association with T2D whereas SNPs rs4368343($OR = 1.3210, P = 8.623\times10^{-7}$), rs6848779($OR = 1.3, P = 2.456\times10^{-6}$), rs11729955 ($OR = 1.3, P = 2.521\times10^{-6}$) detected significant association with T2D as shown in Table 3.

CHR	SNP	P-value	OR	Association
15	rs10519107	3.583×10^{-7}	0.7599	Protective
2	rs4368343	8.623×10^{-7}	1.3210	Association
4	rs6848779	2.456×10^{-6}	1.3	Association
4	rs11729955	2.521×10^{-6}	1.3	Association

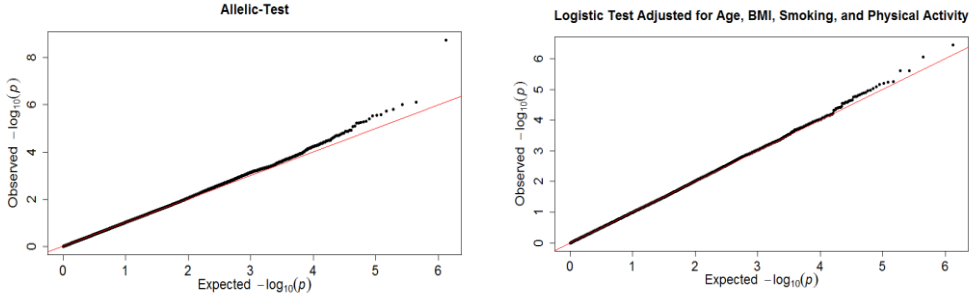
Table 3. SNPs with the Suggestive of Association from Logistic Regression Test



(a) (b)

Fig. 1. Manhattan plot demonstrated the $-log_{10}(p)$ for association of SNPs in a white racial subpopulation NHS data analysis. (a) Manhattan Plot for Allelic Association Test. (b) Manhattan Plot for Logistic Regression adjusted for confounders including age, bmi, smoking status and physical activity.

We used Q-Q plot as demonstrated in Fig. 2 to visualize the relationship between the expected distribution of p-value (null) and observed distribution of p-value of the association test. Allelic test showed that there is a slight deviation in the upper right tail from the y=x line, this suggests the existence of some form of association in the NHS dataset. Logistic regression adjusted for covariates suggested satisfactory and promising outcomes are observed between the expected p-values and calculated p-values, also showed less possibility of systematic bias (population stratification). As most observed SNPs in the study showed no statistical significance than would be expected, however for a number of observed SNPs statistical significance are above the expected and this indicates true association between these SNPs and T2D.



(a) (b)

Fig. 2. Q-Q plot showing the expected (null) vs. observed p-value. The red line represents the null hypothesis of no association. While the black dot refers to the observed $-log_{10}(p)$. (a) Q-Q plot for Allelic test. (b) Q-Q plot for logistic test adjusted for confounders

5 Discussion

In this paper, our discussion of the results presented in Section 4 is based on the consideration of the genetic association analysis that is performed to investigate genetic variations that show evidence of increase susceptibility to T2D. These findings may serve as a rigorous ground to advances the improvement of early predication, and prevention of the disease onset. We focused on two widely used association analysis including allelic test and logistic regression model. It is assumed that allelic test has an additive effect and so it is commonly used. However, logistic regression is the preferred approach due to its flexibility to allow for covariates effects including further SNPs, clinical and sociodemographic risk factors.

Of the list of SNPs obtained from allelic association test, only rs10519107 in chromosome 15 passed the genome-wide significance threshold of 5×10^{-8}. However, rs10519107 showed protective association to T2D with respective odds ratio of 0.7409. The location of rs10519107 is in the Retinoic Acid Receptor-Related Orphan Receptor Alpha (RORα) gene region. RORα gene has known to play an important role in the regulation of lipid and glucose metabolism and insulin expression that are involved in the development of T2D. Researchers in [24] suggested that the genetic variation in RORα gene might be an indication to the individual's susceptibility to T2D. This indicates that the effect of rs10519107 to the susceptibly to T2D could show risk association if it is investigated in another ethnicity populations. Furthermore, nine SNPs have passed the suggestive association threshold of 5×10^{-6}. However, the risk association of rs11701035 could not reach statistical significance. This is probably due to small sample size effect.

Unlike other study, information obtained from logistic regression model have considered the use of non-genetic risk factors such as age, Body Max Index (BMI), smoking status, and physical activity. The effects of these factors on the association analysis have shown promising results however, less SNPs have reached the suggestive association threshold and none of the genotyped SNPs has passed genome-wide significant threshold. Nevertheless, rs10519107 has shown protective association with statistically significance while the remaining (rs4368343, rs6848779, rs11729955) have shown risk association indicating probably with larger sample size these SNPs could worth further investigation.

Although our analysis generated promising results there are other approaches could be considered to model the complexity of non-linearity of genotype-phenotype interactions. Logistic regression has limited power for modelling such interactions. The non-linearity approaches are necessary in discovering the aetiology of complex diseases as T2D. Machine learning algorithms have shown considerable promise. Using machine learning techniques will allow to model the relationship between combinations of SNPs, environmental and clinical factors with disease susceptibility and thus to provide an advanced measurement to the aetiology of T2D. Moreover, considering the correlations between gene-environment interactions and the effects of epistasis (gene-gene interactions) are fundamental to advance researchers and scientists understanding of disease mechanisms as genetic factors (single SNP) do not act independently to increase disease risk.

6 Conclusions

Association analysis tests have been performed to explore the significant association loci (SNPs) that show evidence of increase susceptibility to T2D. Several genetic models have been chosen for association test, more specifically association under logistic regression adjusted for confounders particularly clinical and environmental factors has been examined to measure the strength of association and significant level within genotype-phenotype information. The analyses revealed satisfactory and promising results with significance level (p-value) were observed. The preliminary results that have been obtained are encouraging however, further exploration insights into this dataset remains.

7 References

[1] D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw, "IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030," *Diabetes Res. Clin. Pract.*, vol. 94, no. 3, pp. 311–321, 2011.

[2] J. Gulcher and K. Stefansson, "Clinical risk factors, DNA variants, and the development of type 2 diabetes.," *N. Engl. J. Med.*, vol. 360, no. 13, p. 1360; author reply 1361, 2009.

[3] R. B. Prasad and L. Groop, "Genetics of type 2 diabetes—pitfalls and possibilities," *Genes (Basel).*, vol. 6, no. 1, pp. 87–123, 2015.

[4] F. Medici, M. Hawa, A. Ianari, D. A. Pyke, and R. D. G. Leslie, "Concordance rate for type II diabetes mellitus in monozygotic twins: Actuarial analysis," *Diabetologia*, vol. 42, no. 2, pp. 146–150, 1999.

[5] D. Altshuler, E. Lander, and L. Ambrogio, "A map of human genome variation from population scale sequencing," *Nature*, vol. 476, no. 7319, pp. 1061–1073, 2010.

[6] W. S. Bush and J. H. Moore, "Chapter 11: Genome-Wide Association Studies," *PLoS Comput. Biol.*, vol. 8, no. 12, 2012.

[7] S. Behjati and P. S. Tarpey, "What is next generation sequencing?," *Arch. Dis. Child. Educ. Pract. Ed.*, vol. 98, no. 6, pp. 236–238, 2013.

[8] V. Lyssenko and M. Laakso, "Genetic Screening for the Risk of Type 2 Diabetes Worthless or valuable?," *Diabetes Care*, vol. 36, no. supplement, pp. S120–S126, 2013.

[9] X. Wang, G. Strizich, Y. Hu, T. Wang, R. C. Kaplan, and Q. Qi, "Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction," *J. Diabetes*, vol. 8, no. 1, pp. 24–35, 2016.

[10] N. Hex, C. Bartlett, D. Wright, M. Taylor, and D. Varley, "Estimating the current and future costs of Type1 and Type2 diabetes in the UK, including direct health costs and indirect societal and productivity costs," *Diabet. Med.*, vol. 29, no. 7, pp. 855–862, 2012.

[11] V. S. Samsom M, Trivedi T, Orekoya O, "Understanding the Importance of

Gene and Environment in the Etiology and Prevention of Type 2 Diabetes Mellitus in High-Risk Populations.," *Oral Heal. case reports*, vol. 2, no. 1, pp. 1–10, 2016.

[12] A. Cortes, S. E. Medland, and M. E. Renterı, "Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis," in *Genome-Wide Association Studies and Genomic Prediction*, vol. 1019, Springer Science and Business Media, 2013, pp. 193–213.

[13] D. J. Balding and D. J. Balding, "A tutorial on statistical methods for population association studies.," *Nat. Rev. Genet.*, vol. 7, no. 10, pp. 781–91, 2006.

[14] S. Tudies, M. Murea, L. Ma, and B. I. Freedman, "Genetic and environmental factors associated With type 2 diabetes and diabetic vascular complications," *Rev. Diabet. Stud.*, pp. 6–22, 2012.

[15] A. K. Cheema, T. Li, J. P. Liuzzi, G. G. Zarini, M. T. Dorak, and F. G. Huffman, "Genetic associations of PPARGC1A with type 2 diabetes: Differences among populations with African origins," *J. Diabetes Res.*, vol. 2015, 2015.

[16] L. Qiu, R. Na, R. Xu, S. Wang, H. Sheng, W. Wu, and Y. Qu, "Quantitative assessment of the effect of KCNJ11 gene polymorphism on the risk of type 2 diabetes," *PLoS One*, vol. 9, no. 4, 2014.

[17] K. A. Tryka, L. Hao, A. Sturcke, Y. Jin, Z. Y. Wang, L. Ziyabari, M. Lee, N. Popova, N. Sharopova, M. Kimura, and M. Feolo, "NCBI's database of genotypes and phenotypes: DbGaP," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 975–979, 2014.

[18] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, 2007.

[19] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, and P. Andrew, "Basic statistical analysis in genetic case-control studies," *Nat. Am.*, vol. 6, no. 2, pp. 121–133, 2011.

[20] X. Wang, C. Baumgartner, D. C.Shields, H.-W. Deng, and J. S.Beckmann, *Application of Clinical Bioinformatics*, vol. 11. Dordrecht: Springer Netherlands, 2016.

[21] M. Bland, *An Introduction to medical statistics*, 4th ed. OXFORD UNIVERSITY PRESS, 2015.

[22] Z. Chen, H. Huang, and H. K. T. Ng, "An improved robust association test for GWAS with multiple diseases," *Stat. Probab. Lett.*, vol. 91, pp. 153–161, 2014.

[23] W. Li, "Three lectures on case-control genetic association analysis," *Brief. Bioinform.*, vol. 9, no. 1, pp. 1–13, 2008.

[24] Y. Zhang, Y. Liu, Y. Liu, Y. Zhang, and Z. Su, "Genetic Variants of Retinoic Acid Receptor-Related Orphan Receptor Alpha Determine Susceptibility to Type 2 Diabetes Mellitus in Han Chinese," *Genes (Basel).*, 2016.