Modern Applied Science; Vol. 8, No. 6; 2014
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Using a Case-Control Genotypic Testing in Investigating the Association with Type-2 Diabetes

Hajah Norhakimah Haji Mohd Nor[1] & Masitah Shahrill[2]

[1] Sultan Bolkiah Vocational School, Ministry of Education, Bandar Seri Begawan, Brunei Darussalam

[2] Sultan Hassanal Bolkiah Institute of Education, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam

Correspondence: Masitah Shahrill, Sultan Hassanal Bolkiah Institute of Education, Universiti Brunei Darussalam, Jalan Tungku Link, Gadong, BE 1410, Bandar Seri Begawan, Brunei Darussalam. Tel: 673-246-3001; Fax: 673-246-1003. E-mail: masitah.shahrill@ubd.edu.bn

Received: July 10, 2014 Accepted: August 10, 2014 Online Published: August 19, 2014

doi:10.5539/mas.v8n6p1 URL: http://dx.doi.org/10.5539/mas.v8n6p1

Abstract

In the United Kingdom, Type-2 Diabetes (T2D) is the leading cause of blindness among working age adults. It is also known to cause kidney failure, amputations and cardiovascular diseases. In this study, genetic association tests were used to compare genetic variants carried by individuals against their disease status, with the aim to find genes that contributed to the risk of T2D. The identification of these genes could be of great importance especially in preventive healthcare measures. This study used a case-control genotypic test to find the association between Single Nucleotide Polymorphisms (SNPs) on chromosome 10 and T2D. SNPs are a type of polymorphism that occurs when a single nucleotide (A, C, G, and T) in the genome is substitute for another. At the beginning of the study, we had a total of 28,501 SNPs, however, 4,101 SNPs were removed after conducting both the Hardy Weinberg Equilibrium test and the control of Minor Allele Frequency in the preliminary analyses. These quality controls were done to remove SNPs that may lead to false associations. A total of 24,400 SNPs were left for association testing using the genotypic test of the 2×3 contingency table. Our testing revealed that there were a total of 12 SNPs that had potential association with the risk of T2D.

Keywords: Type-2 Diabetes (T2D), association testing, genotypic test

1. Introduction

1.1 Type-2 Diabetes

The number of people with diabetes mellitus in the world's population may be expected to double, from around 180 million to 300 million by 2025 (Zimmet et al., 2001). Diabetes Mellitus, also known as diabetes, is a long-term disorder caused by an absence or deficiency of insulin, which led to a significantly high level secretion of glucose (sugar) in the bloodstream (World Health Organization, 1999). Insulin, produced in the pancreatic β-cells, is the key hormone responsible in regulating glucose level in the bloodstream. As mentioned by the WHO, there are two main forms of diabetes: Type-1 and Type-2 diabetes.

The Type-1 Diabetes (T1D) is a condition where the pancreas is unable to secrete any insulin because the body's immune systems have destroyed the insulin-producing cells (β-cells). While, the Type-2 Diabetes (T2D) is a condition where the β-cells produces insufficient insulin to control the level of glucose in the blood, or the body's cells especially in muscle, fat and liver cells do not react effectively to insulin produced (known as insulin resistance). Both these diabetes have been known to result from the combination of both genetic and environmental risk factors. Thus, termed as 'complex' diseases.

In the United Kingdom, diabetes is the leading cause of blindness in people of the working age (Arun et al., 2003), and the main contributor to kidney failure, amputations and cardiovascular disease, including heart attack and stroke (Diabetes UK, 2014). The T2D usually develop at the age of 40 and it accounts for approximately 90% of all adults affected with diabetes, making it the most common form of diabetes. The T2D are known to be predominant for people of South Asian, African-Caribbean or Middle Eastern descendants, as it can start as early as the age of 25.

1.2 Genetics & Molecular Biology

In this section, some of the genetic terms that will be used in this study are described. In addition, most of the definitions are taken from the webpage for the National Human Genome Research Institute (NHGRI) (Note 1).

It is known that T2D have a strong hereditary component. The gene is the fundamental physical and functional unit of inheritance, passed from parents to offspring. In addition, the genes contain the information to explain specific traits. They are arranged in a linear manner, on structures called chromosomes. An allele is one of two or more versions of a gene. There are two alleles at each gene in every individual with one allele inherited from each parent. If there are two copies of the same alleles, the individual is homozygous for that gene. The individual is heterozygous if the two alleles are different. Phenotype refers to an individual's observable outcome or traits such as eye colour or the disease status. In our study, we refer the phenotypes as either having T2D or not. In contrast, genotype refers to the individual's unobserved genetic contribution to the phenotype. A locus (plural, loci) is the physical location of a gene of interest on a chromosome. The entire set of individual's genetic constitution is called genome. In humans, the genome usually consists of 23 pairs of chromosomes.

In general, the risk of developing a disease can either be the result of a single major gene, called 'monogenic' or the combination of small effects from multiple genes, called 'polygenic'. Consider the case where a disease is caused by a genetic variant in a locus in two distinct forms (alleles): *d*, the normal allele, and *D*, the disease susceptibility allele. There would be three possible genotypes at a single biallelic locus: *dd*, *dD* and *DD*, with *dd* representing the normal genotype. The penetrance function is the set of probability distribution functions for the phenotype given the genotype. This is given by,

$$Pr\left(Y \mid G\right) = Pr\left(Y = 1 \mid dd\right) = 0$$

where G is genotype and Y is phenotype assumed binary: 0 indicating unaffected and 1 indicating affected.

The concept of dominance can be categorized into three genetic risk models: dominant, recessive and codominant. In dominant model, allele *D* (disease susceptibility allele) is dominant over *d* which led to genotype *dD* and *DD* to have the same effect on the phenotype, (*Pr (Y|dD) = Pr (Y|DD))*. In recessive model, where allele *D* is recessive over *d*, genotype *dD* have the same effect on the phenotype as genotype *DD*, (*Pr (Y|dD) = Pr (Y|dd))*. When the allele is neither dominant nor recessive, each genotype has different effect on the phenotype (*Pr (Y|dd)* ≠ Pr (Y|dD) ≠ Pr (Y|DD)). In most cases, the heterozygote (*dD*) has an intermediate effect between that of the two homozygotes (*dd* and *DD*). This brings us to additive cases there *P (Y|dD)* is a midway between *P (Y|dd)* and *P (Y|DD)*. In this study, we will be assuming the genetic risk model of T2D to be codominant.

A nucleotide is the basic building block of nucleic acids, which consists of phosphate, sugar and bases. The four bases in the DNA are Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The most abundant genetic variations in the human genome are the single nucleotide polymorphisms. Single nucleotide polymorphisms or SNPs (pronounced 'snips') are a type of polymorphism that occurs when a single nucleotide (A, C, G, and T) in the genome is substituted for another. This is illustrated in Figure 1. There is about 3 billion nucleotide in the human genome and approximately 10 million of these will have SNPs occurring. If there are only two variants occurring at a single locus, it is called biallelic. Most SNPs will have no effect on health or development. However, some of these genetic variations (such as SNPs) have been proven to be harmful and may affect the risk of developing diseases. This is the interest of our study and we will only be dealing with biallelic SNPs. A haplotype is a group of single nucleotide polymorphisms (SNPs) residing on the same chromosome which tends to be inherited together.

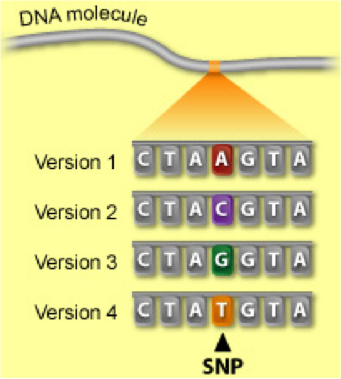


Figure 1. Illustration of SNP (taken from http://learn.genetics.utah.edu/content/pharma/snips/)

2. Objective of the Study

The completion of both the Human Genome Project in 2003 and the International HapMap Project in 2005 have made it possible for researchers to have the tools to detect genes that are associated with the risk of common diseases (NHGRI). In 2007, the Wellcome Trust Case-Control Consortium (WTCCC, 2007) reported their findings on the Genome Wide Association Studies (GWAS) they conducted on the 7 common diseases: coronary heart disease, T1D, T2D, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension. They were successful in finding many new disease genes which are associated with these diseases.

In this study, we are using data from the Wellcome Trust Case Control Consortium (WTCCC, 2007) but only focusing on the T2D. The data consists of 3499 individuals: 1999 individuals with T2D ('cases') and 1500 normal ('controls'). The T2D cases are identified as UK Caucasian. There were two control groups used in the original study but we will only be using one, which is from the National Blood Service (NBS). These are made of individuals living within England, Scotland and Wales. Due to our limited computer capacities and time constraint for data analysis, we will only be analysing SNPs in chromosome 10, which is a total of 28,501 SNPs. The original study had analysed the whole genome and had identified chromosome 10 to be the region which generates the strongest association signal for T2D. This is our main reason for choosing chromosome 10. Using these large numbers of bi-allelic SNPs, the objective of our study is to identify the SNPs that are associated with the risk of developing T2D for the British Caucasian population. Most of the statistical analyses were performed using the software package PLINK 1.07 (Purcell et al., 2007). Other software used were *R* (R Development Core Team, 2010) and Haploview (Barrett et al., 2005).

3. Preliminary Analysis

The quality control filtering process is a crucial first step in genetic association study to ensure the quality of the genotype data. We will need to remove low quality SNPs that may lead to false association. The two quality control processes are the Hardy Weinberg Equilibrium (HWE) test and Minor Allele Frequency (MAF).

3.1 Hardy Weinberg Equilibrium

The Hardy Weinberg law states that the genotype and allele frequencies of a large, randomly mating population remain constant from one generation to the next provided migration, mutation, and natural selection do not take place (Ziegler & Konig, 2006). Consider a biallelic locus with alleles *D* and *d*. Denote *f (D)* and *f (d)* as the allele frequencies of *D* and *d* respectively. There are three possible genotypes: *DD, Dd* and *dd*. For each locus, we can deduce two predictions by HWE. Firstly, the allelic frequencies in a population will not change from generation to generation: $p = f(D),\ q = f(d)$; and secondly, the genotype frequencies remain constant after one generation in the proportions. A significant deviation from HWE for a SNP could be due to non-random mating, selection or genotyping error. Our major concern is the genotyping error where heterozygotes are most commonly misclassified as one of the homozygotes. For example, *AT* are misclassified as *AA* or *TT*. Genotyping errors may lead to a false association.

We performed the HWE test for the control group only because we used a case-control test, where deviation from HWE in the case is interpreted as having an association with the risk of T2D. We tested HWE for each SNP by comparing the observed genotype frequencies with those expected under HWE. The null hypothesis, H_0 state there is no significant difference between the observed and the expected genotypic frequencies under HWE. The alternative hypothesis H_1 is that there is a significant difference between the observed and expected genotype frequencies. The commonly used approaches are the Pearson's chi-squared goodness-of-fit test and the Fisher exact tests (Agresti, 2013). Fisher exact test is preferred as it is computationally more demanding but can be computed easily in PLINK.

HWE test: Using chi-squared test

Consider a sample with *n* individuals, and denote the observed genotype frequencies of *DD, Dd,* and *dd* as *f (DD), f (Dd)*, and *f (dd)* respectively. Denote *f (D)* and *f (d)* as the allele frequencies of allele *D* and *d*, given by

$$f(D) = \frac{2f(DD)+f(Dd)}{2n},\ f(d) = \frac{2f(dd)+f(Dd)}{2n}$$

Under the null hypothesis, the expected genotype frequencies for *DD, Dd,* and *dd* are given by

$$f_{\exp}(DD) = nf(D)^2,\ f_{\exp}(Dd) = 2nf(D)f(d),\ f_{\exp} = nf(d)^2$$

The test statistic of the Pearson's chi-square goodness-of-fit test is given as

$$X^2 = \sum_{genotypes} \frac{(\text{Observed counts} - \text{Expected counts})^2}{\text{Expected counts}}$$

and the number of degrees of freedom, v are calculated by

$$v = \text{Number of expected genotype} - \text{Number of alleles}$$

The X^2 test statistic asymptotically follows a chi-squared distribution with one degree of freedom, where we can obtain the corresponding p-value and compare it with the threshold to make conclusion about HWE.

Using the example of SNP-A_4300367, consisting of 1500 individuals (controls), n =1500. The genotype *AA, AG* and *GG* have the observed genotypic frequencies of 328, 763 and 409 respectively. Denote *f (A)* and *f (G)* as the allele frequencies of *A* and *G* respectively, given by

$$f(A) = \frac{2f(AA) + f(AG)}{2n} = \frac{(2\times 328) + 763}{2\times 1500} = 0.473$$

$$f(G) = \frac{2f(GG) + f(AG)}{2n} = \frac{(2\times 409) + 763}{2\times 1500} = 0.527$$

Under the null hypothesis,

$$\begin{array}{llcl} f_{\exp}(AA): & nf(A)^2 & = & 1500\times 0.473\times 0.473 & = & 335.5935 \\ f_{\exp}(AG): & 2nf(A)f(G) & = & 2\times 1500\times 0.527\times 0.473 & = & 747.813 \\ f_{\exp}(GG): & nf(G)^2 & = & 1500\times 0.527\times 0.527 & = & 416.5935 \end{array}$$

The observed and expected genotype frequencies are summarized in Table 1.

Table 1. Distribution of the observed and expected genotypic counts

Genotype	AA	AG	GG	Total
Observed	328	763	409	1500
Expected	335.6	747.8	416.6	1500

The chi-squared test statistic is

$$\begin{aligned} X^2 &= \sum \frac{(\text{Observed counts} - \text{Expected counts})^2}{\text{Expected counts}} \\ &= \frac{(328-335.6)^2}{335.6} + \frac{(763-747.8)^2}{747.8} + \frac{(409-416.6)^2}{416.6} \\ &= 0.1718 + 0.3084 + 0.1384 \\ &= 0.6187 \end{aligned}$$

and the number of degrees of freedom,

$$\begin{aligned} v &= \text{Number of expected genotype} - \text{Number of alleles} \\ &= 3 - 2 = 1 \end{aligned}$$

The corresponding p-value for $X^2 = 0.6187$ on 1 degree of freedom is 0.4315. In this analysis, we had set the threshold of p-value at 5.7×10^{-7} . Since the p-value is higher than the significance level, there is not enough evidence to reject the null hypothesis. We can conclude that SNP_A-4300367 is in Hardy Weinberg Equilibrium. For the Fisher exact test, it is impossible to calculate this example without the help of statistical software. Hence, in our study, we conducted the Fisher's exact test in PLINK for the analysis of HWE.

3.2 Minor Allele Frequency

The Minor Allele Frequency (MAF) is the frequency of the less frequent allele (minor allele) over the total allele in that sample. In this study, we assume common SNPs with MAF greater than 0.01 are to be the cause of the common diseases like T2D. Therefore, common SNPs are more likely to reflect true associations than rare SNPs since there is a greater power to detect common SNPs. However, studies have shown that rare SNPs may have contributed to the risk of common diseases (Bodmer & Bonilla, 2008). We will not be analysing these rare SNPs in our study. In fact, we are excluding SNPs with MAF value less than 0.01, otherwise known as the 'rare SNPs'. For example, in SNP_A-4252987, we have 8 count for allele *T* and 6991 for allele *G*. Thus allele *T* is the minor allele and its minor allele frequency,

$$\text{MAF} = \frac{8}{6991+8} = 0.00143$$

Since the MAF value is less than 0.01, we will exclude this SNP from further analysis.

4. The Association Testing

The genetic association studies aim to find the relationship between the genetic variants carried by individuals and their phenotypes. In this association study, it requires the criteria that individuals to be tested is distant or of unknown relationship with each other. Subsequently, we used the biallelic single nucleotide polymorphism (SNP) as the genetic variant and look for its relationship with the risk of traits T2D. Specifically, we used the case-control test of association. It compared a single SNP with the disease status (cases or controls) and derive conclusion when a particular SNP is in significant abundance in cases or controls.

The most powerful test for detecting association is one which uses the true underlying genetic risk model. However, the true genetic risk model is not known and choosing the wrong genetic risk model will lead to a very low power of detecting association. In this study, we only used the genotypic test which assumed a codominant risk model. Codominant risk model can either be general or additive. We will not be performing tests based on dominant or recessive risk model.

4.1 Genotypic test, 2×3 table

In a case-control setting, we can represent the genotype data in a 2×3 contingency table where each individual is classified according to their disease status and the genotype they carry. Here, we denote *D* and *d* for the major and minor allele respectively. The three possible genotypes are *DD*, *Dd* and *dd* representing homozygotes, heterozygotes and rare homozygotes respectively. This construction is shown in Table 2. To test for association between a SNP and disease, we can carry out the usual χ^2 test for independence of rows and columns in contingency tables. The null hypothesis, H_0: The SNP has no association with the disease (no association between row and column of the table) against the alternative hypothesis H_1 : The SNP has an association with the disease. This is called genotypic test.

Table 2. The general 2×3 contingency table of the genotype counts

Genotype	DD	Dd	dd	Total
Cases	r_0	r_1	r_2	R
Controls	q_0	q_1	q_2	S
Total	N_0	N_1	N_2	N

The chi-squared test statistic is given by

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

and the number of degrees of freedom, *v* is given by

$$v = (k-1) = 3-1 = 2$$

where *k* is the number of genotypes. Here, X^2 has χ^2 distribution with 2 degrees of freedom under the null hypothesis. For example, SNP_A-2061203 has genotype counts of 751, 926, and 322 for *CC, CT* and *TT* respectively, in cases. A genotype counts of 571, 698, and 231 for *CC, CT* and *TT* respectively, in controls. The major allele is *C* and minor allele is *T*. The hypotheses are stated as follows:

$$H_0\text{:The SNP }(\text{SNP\_A}-2061203)\text{ has no association with the risk of T2D.}$$

$$H_1\text{:The SNP }(\text{SNP\_A}-2061203)\text{ has an association with the risk of T2D.}$$

The observed and expected frequencies are shown in Figure 2. The expected frequency is given by

$$\text{Expected frequency, } E_{ij} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Overall Total, } (N)}$$

Therefore, the expected value for the case and of genotype *CC* is calculated by

$$E = \frac{1999 \times 1322}{3499} = 755.2666$$

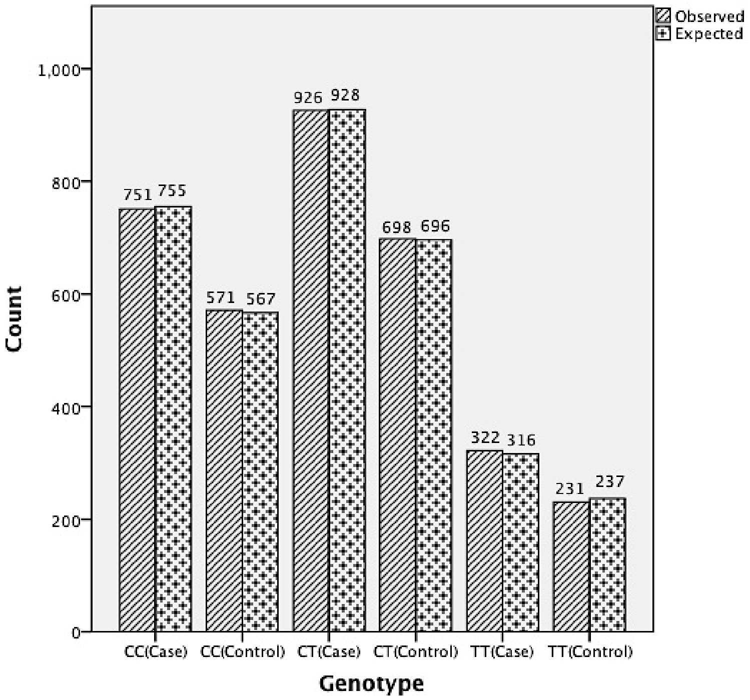


Figure 2. The genotype counts for SNP2061203

The chi-square test statistic,

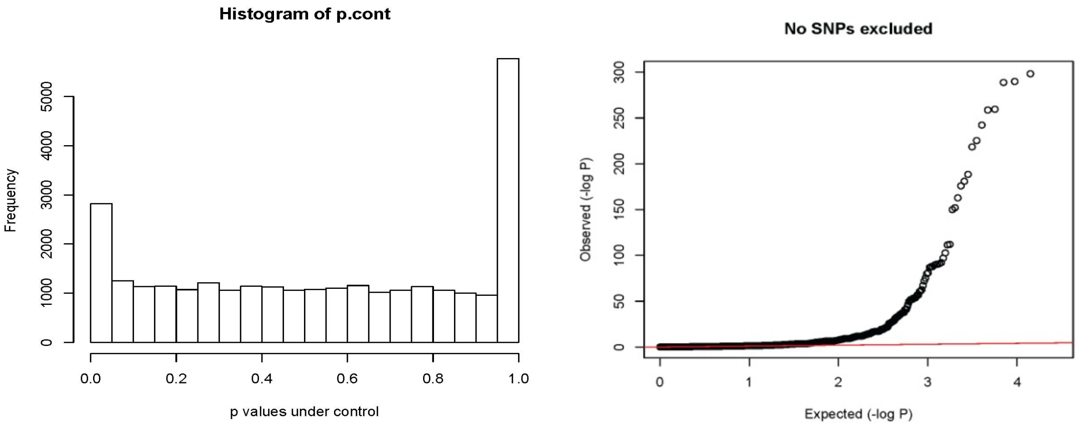
$$\begin{aligned} X^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(751-755.6)^2}{755} + 0.004 + 0.114 + 0.028 + 0.006 + 0.152 \\ &= 0.336225 \end{aligned}$$

and the corresponding *p*-value on 2 degrees of freedom is 0.8453. At a significance level of 5.7×10^{-7}, we do not have enough evidence to reject the null hypothesis and hence concluded there is no association between the SNP (SNP_A-2061203) and the risk of T2D.

5. Results

5.1 Hardy Weinberg Equilibrium (HWE) Test

Using PLINK, we ran the HWE test using Fisher's exact test on each and every 28,501 SNPs simultaneously to generate the observed HWE *p*-values. We only applied the HWE test on the control group. A histogram is used to show the distribution of HWE *p*-values across the whole SNPs. A quantile-quantile (QQ) plot is constructed by rearranging the observed negative log HWE *p*-value ($-\log_{10}p$) from the smallest to largest and plotted them against the expected negative log HWE *p*-value, the expected value are assumed to have a chi-squared distribution with 1 degree of freedom. A red line is added to the QQ plot to indicate the expected value under the null hypothesis. QQ plot can show indication of any deviation or as a measure of deviation from HWE. The histogram and QQ plot are shown in Figure 3.

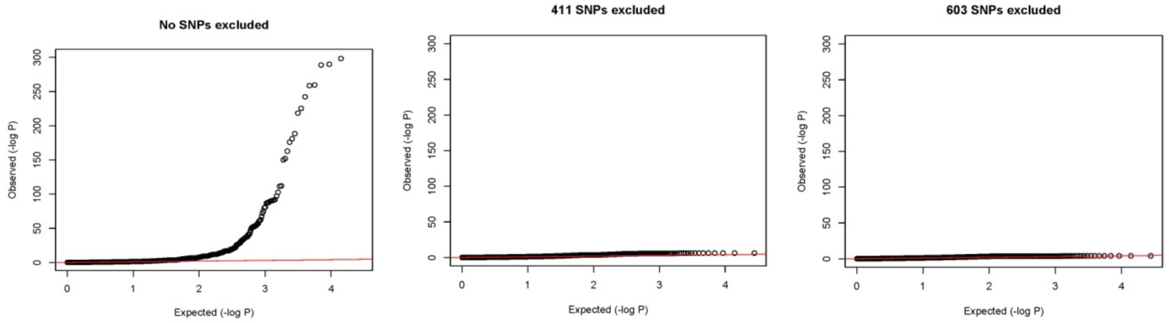


(a) (b)

Figure 3. (a) Histogram of the HWE p-values under null hypothesis (no SNPs excluded), (b) QQplot for HWE p-values (no SNPs excluded)

We can see from Figure 3(a), there is a peak at the very small HWE *p*-value, but overall the distribution is fairly uniform. For Figure 3(b), there are some SNPs showing big deviation from the expected value under the null hypothesis. These SNPs are not in HWE and thus will lead to false association. Using WTCCC threshold level of 5.7×10^{-7}, we will exclude all the SNPs with HWE *p*-value less than this value from further analysis.

A total of 411 SNPs have failed the HWE test and hence excluded. Now, we draw a QQ plot of the observed against expected $-\log_{10}p$ for the remaining SNPs. Then we can compare the plots of the before and after the exclusion of 411 SNPs. This is shown in Figure 4. We can see the SNPs that gave big deviation from the expected value have been removed. The remaining SNPs in the controls are in HWE.

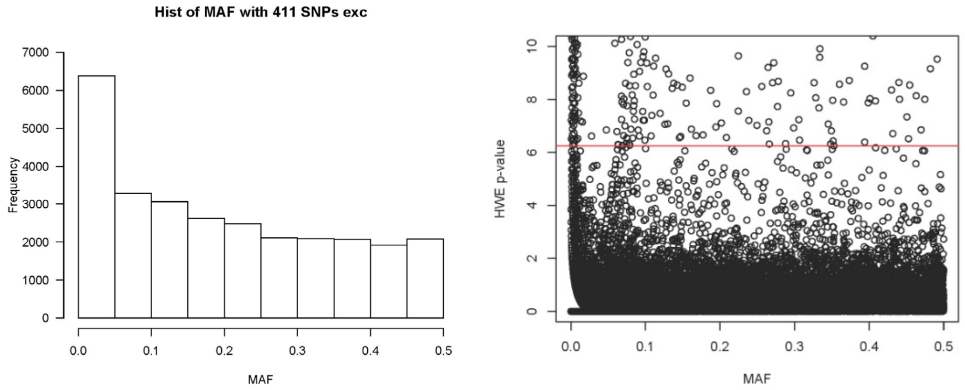


(a) (b) (c)

Figure 4. (a) QQplot for HWE p-values (no SNPs excluded) (b) QQplot with HWE $p \le 5.7\times10^{-7}$ excluded, having the same scale as graph (a), (c) QQplot with HWE $p \le 5.7\times10^{-7}$ excluded, having the graph zoomed in

5.2 Minor Allele Frequency

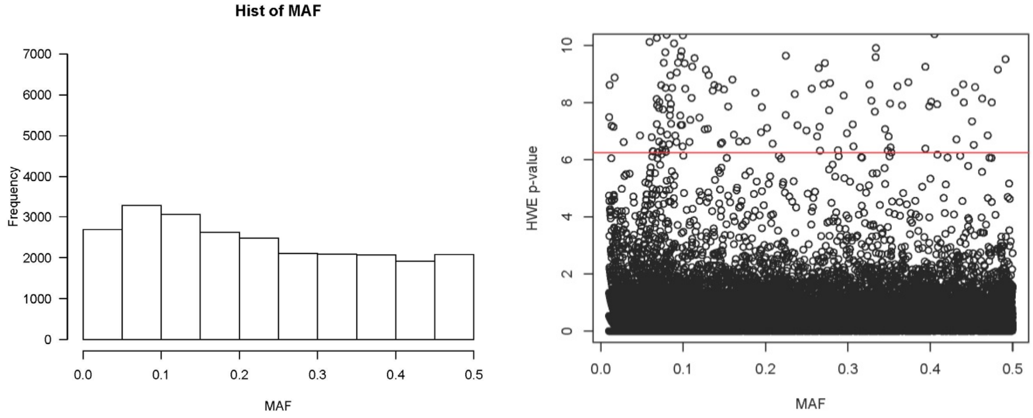
After 411 SNPs failed the HWE test, a total of 28,090 SNPs are left to be analysed. The second quality control is to remove a low MAF to avoid inflated false-positive results. We use PLINK to generate the MAF value of all the SNPs. A histogram is used show the distribution of the MAF values across the whole SNPs. A plot of HWE $-\log_{10}p$ of control and case combined against MAF value indicates how they are related. Both of these plots are shown in Figure 5. In the histogram plot of the MAF, there is a big peak in the range 0-0.05. In the second plot, we can see the inflated *p*-value caused by MAF. We will exclude SNPs that has MAF < 0.01, threshold we had set.



(a) (b)

Figure 5. (a) Histogram of the MAF and (b) Plot of the HWE p-value under control against the MAF value. Both graphs are using the data where only the exclusion by the HWE test is applied

A total of 3690 SNPs had MAF < 0.01 and they are excluded from further analysis. We repeate the same plots for the remaining 24,400 SNPs. This is shown in Figure 6. For the histogram, we can see the MAF distribution is now more uniform. For the second plot (b), the SNPs with a low MAF which caused inflated *p* value have been removed.



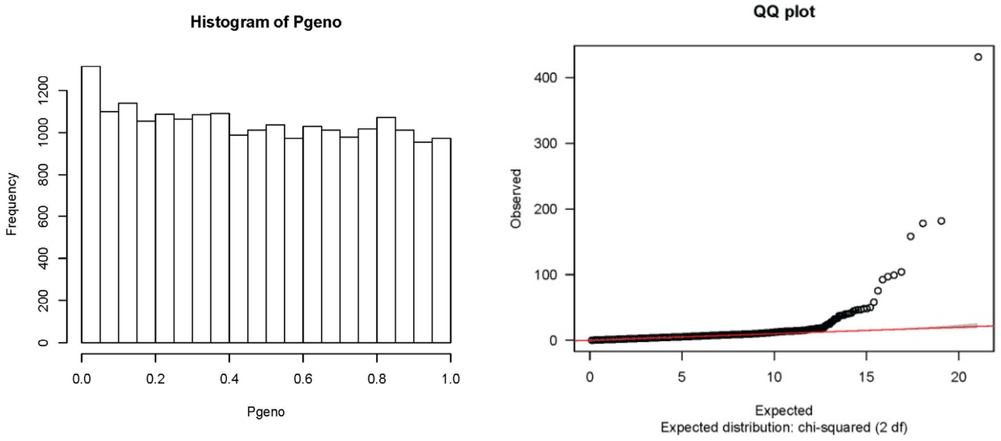
(a) (b)

Figure 6. (a) Histogram of the MAF and (b) Plot of the HWE p-value under control against the MAF value. Both graphs are using the data after excluding HWE p-value < 5.7×10^{-7} and MAF < 0.01.

5.3 The Genotypic Test

The genotypic test on a 2×3 contingency table is performed on the remaining 24,400 SNPs, which have passed both the HWE and MAF tests. A histogram of the genotypic *p*-values can show its distribution across the whole

SNPs. A quantile-quantile (QQ) plot is constructed by rearranging the observed negative log genotypic *p*-value ($-\log_{10}p$) from the smallest to largest and plotted them against the expected negative log genotypic *p*-value, the expected value are assumed to have a chi-squared distribution with 2 degrees of freedom. A red line is added to the QQ plot to indicate the expected value under the null hypothesis. The QQ plot can show SNPs that deviate from the null hypothesis and hence having an association with the risk of T2D. Both of these plots are shown in Figure 7. In Figure 7 (b), we can see some SNPs showing a big deviation from the expected value. The SNPs of interest are the ones with *p*-values less than 5.7×10^{-7}, where the distribution of the genotype is considered not random and has an association with the risk of T2D.



(a) (b)

Figure 7. (a) Histogram of the p-values under genotypic allelic test, (b) QQ plot of the observed against the expected log (p-value) under genotypic test

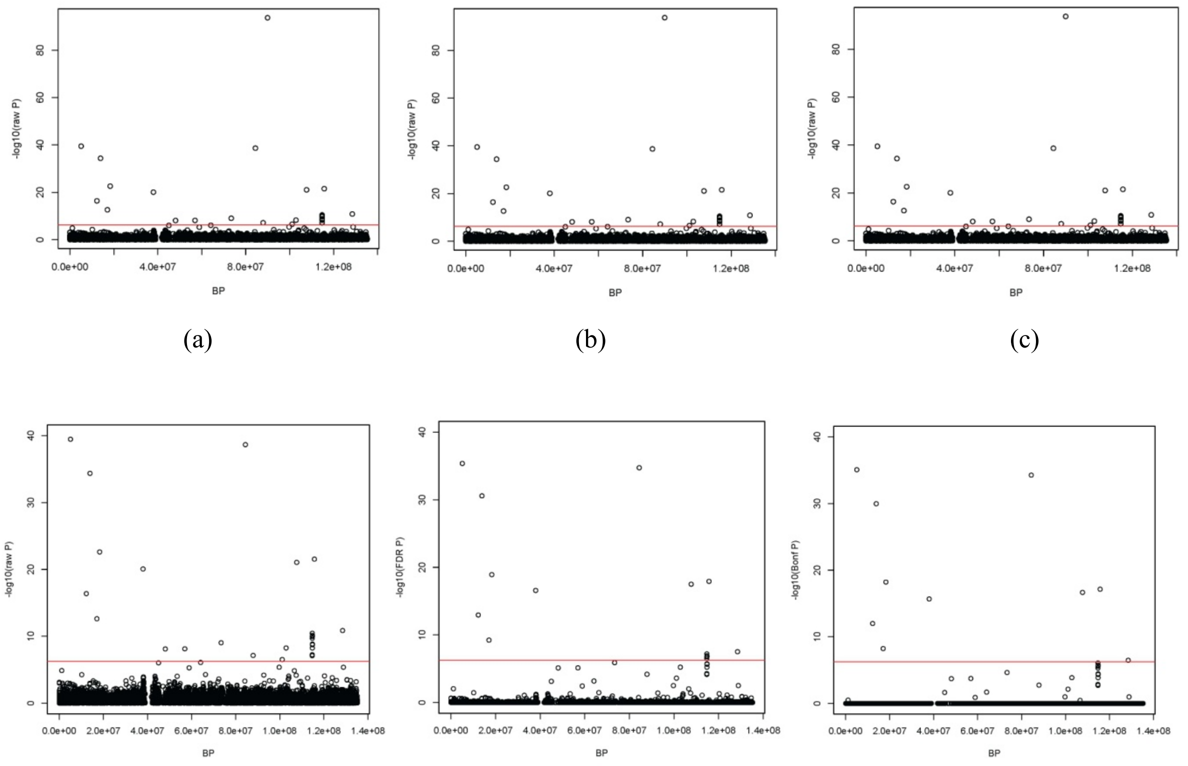
To address the multiplicity issues, we performed the Bonferroni, Sidak, Holm and the adjustment by False Discovery Rate (FDR) corrections to the genotypic *p*-value (Blakesley et al., 2009; Bretz et al., 2005). The SNPs are arranged according to the *p*-value, from the smallest to largest. The distribution of the top 17 *p*-values under the genotypic association test is shown in Table 3. In this table, '-' means the *p*-value is higher than 5.7×10^{-7}. Initially, there are 35 SNPs with genotypic *p*-value less than 5.7×10^{-7}. After applying the Bonferroni, Sidak, Holm and FDR corrections, there are 11, 11, 11 and 17 SNPs with *p*-values less than 5.7×10^{-7} respectively.

Table 3. Distribution of the raw and adjusted p-values under the genotypic association test

SNP	BP	P(UNADJ)	Bonferroni	Sidak	Holm	FDR_BH
1780818	89951510	2.093e-94	5.107e.90	0.0	5.107e.90	5.107e.90
1811809	5091939	3.473e-40	8.474e-36	0.0	8.474e-36	4.237e-36
2147541	84415869	5.485e-35	1.073e-31	0.0	5.485e-35	1.828e-35
1859018	13916693	4.373e-35	1.067e-30	0.0	1.067e-30	2.668e-31
4271922	18307982	2.490e-23	6.076e-19	0.0	6.075e-19	1.215e-19
1833194	115717273	2.930e-22	7.149e-18	0.0	7.148e-18	1.192e-18
2212963	107704329	8.968e-22	2.188e-17	0.0	2.188e-17	3.126e-18
4280587	37997426	8.613e-21	2.102e-16	0.0	2.101e-16	2.627e-17
2083541	12273828	4.184e-17	1.021e-12	0.0	1.021e-12	1.134e-13
2155029	17062946	2.406e-13	5.871e-09	5.870e-09	5.868e-09	5.871e-10
4273904	128482238	1.386e-11	3.382e-07	3.382e-07	3.380e-07	3.074e-08

2005462	114746031	3.499e-11	-	-	-	6.618e-08
1793595	114744078	3.526e-11	-	-	-	6.618e-08
2130702	114779067	7.354e-11	-	-	-	1.266e-07
1893548	114771287	7.784e-11	-	-	-	1.266e-07
2298644	114757761	1.106e-10	-	-	-	1.687e-07
1809965	114738487	2.004e-10	-	-	-	2.876e-07

As shown in Table 3, the p-values under the Bonferroni, Sidak and Holm corrections are almost similar. Only the FDR correction gives different p-values and added more significant SNPs. We plotted the unadjusted genotypic $-\log_{10}p$ values against the base position of the SNPs. We repeated the plot for the Bonferroni p-values as well as FDR p-values. This is shown in Figure 8. From the 3 plots, we can see the difference between the 3 types of p-values (before and after multiplicity correction) as well as the location of the significant SNPs. The significant SNPs are those found above the red line, having p-value less than 5.7×10^{-7}.

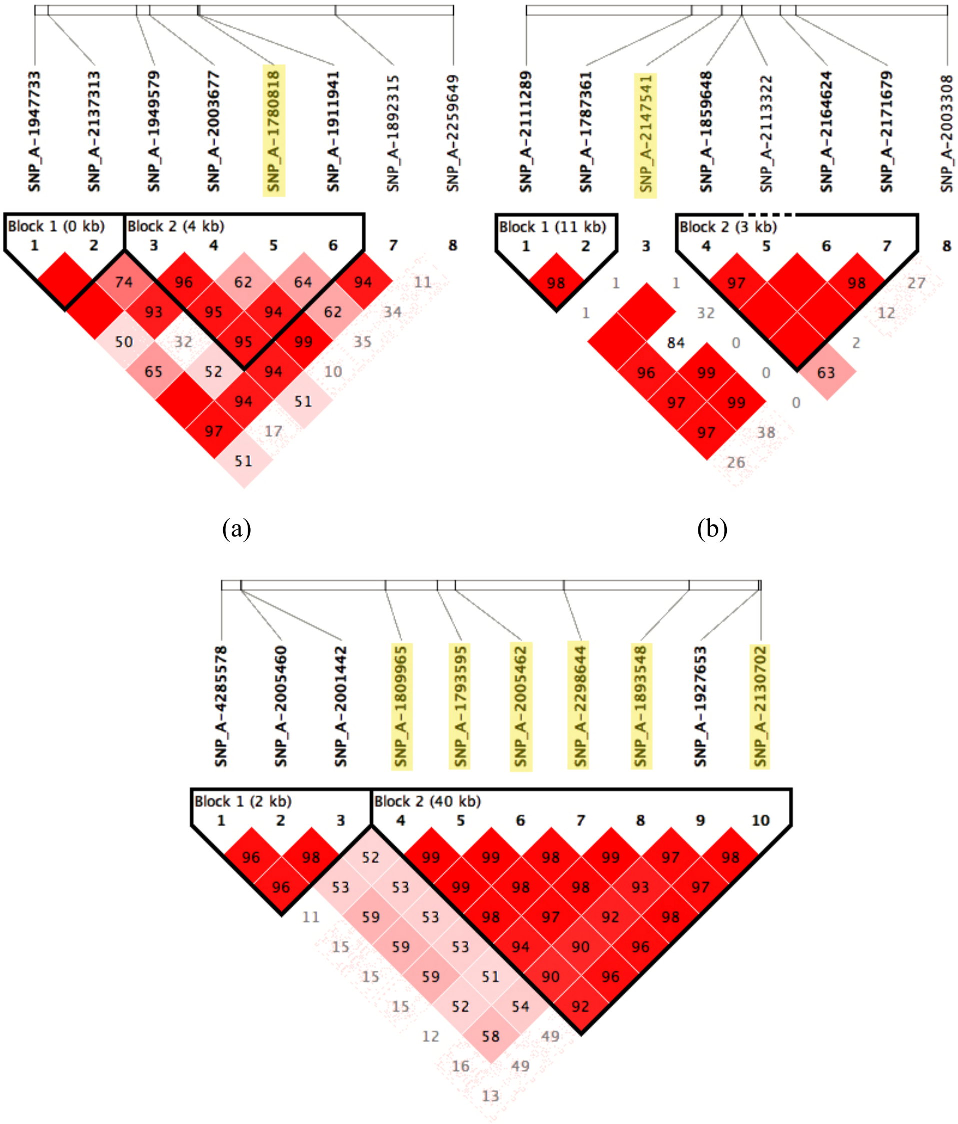


(a) (b) (c)

(d) (e) (f)

Figure 8. Under genotypic test: (a) Plot of raw $-\log_{10}p$-values against base position (BP), (b) Plot of adjusted $-\log_{10}p$ values using Bonferroni correction against BP, (c) Plot of adjusted $-\log_{10}p$ values using FDR correction against BP. The next 3 graphs (d)-(f) are similar to (a)-(c) respectively but at a smaller scale (more focus). The red line is $-\log_{10}p = -\log_{10}\left(5.7\times10^{-7}\right)$

For all the 17 significant SNPs, we constructed the Linkage Disequilibrium (LD) plot within the range of $\pm20\text{kb}$ (or 50 kb) base position of the SNP using Haploview. This is to identify the presence of LD between the SNPs. Figure 9 contain 3 separate LD plots. In Figure 9, (a) LD plot of SNP_A-1780818, it shows the alleles in that SNP is in high LD with alleles of SNP_A-1949579, (b) LD plot of SNP_A-2147541, the alleles are not in LD with any of the alleles of neighbouring SNPs, (c) LD plot containing 6 SNPs: SNP_A-180995, SNP_A-1793595, SNP_A-2005462, SNP_A-2298644, SNP_A-1893548, SNP_A-2130702. All of the alleles in the 6 SNPs are in high LD with each other.



(a)

(b)

(c)

Figure 9. (a) LD plot including SNP1780818, (b) LD plot including SNP2147541, and (c) LD plot including 6 SNPs: 2005462, 1793595, 2130702, 1893548, 2298644, 1809965. Shown in each box are the D' value expressed as a percentile, derived by Haploview software (v3.2)

6. Discussion

We had a total of 28,501 SNPs at the start of our study. The 411 SNPs in the control group had HWE p-value of less than 5.7×10^{-7} which could be the cause of genotyping error. Since they may lead to false association, they were excluded from further analysis. After excluding the 411 SNPs, we found 3690 SNPs had MAF of less than 0.01. Since we are assuming common SNPs (with MAF > 0.01) to be the underlying cause of T2D, these 3690 rare SNPs were also removed from further analysis. Therefore, in the preliminary analyses, we have removed a total of 4101 SNPs, which may potentially lead to false association. A total of 24,400 SNPs are left for association testing. The first test of association we did was the genotypic test of the 2×3 contingency table. This was assuming that the risk of T2D is general or non-additive. Here, we found 35 SNPs that were significant at a threshold of 5.7×10^{-7}, but later reduced to 17 SNPs after FDR corrections and subsequently reduced to 11 SNPs after the conservative Family-Wise Error Rate (FWER) corrections. The 6 SNPs (SNP_A-2005462, SNP_A-1793595, SNP_A-2130702, SNP_A-1893548, SNP_A-2298644 and SNP_A-1809965) that passed the FDR but failed FWER correction were all found to be in high LD with each other. In this case, the alleles of the SNPs were dependent on each other and therefore FDR is a better correction than FWER. It was sufficient to analyse only one of the 6 SNPs. The two SNPs (SNP_A-1780818 and SNP_A-4271922) were found to be in high LD ($D' > 95$) with neighbouring SNPs that we had declared had no association with the risk of T2D. This may not necessarily conclude that the two SNPs are of false association. The remaining 9 SNPs were not in high LD with any of the neighbouring SNPs. This could either mean the SNPs are the true causal variants directly influencing the risk of T2D, or they are of false association. We then had a total of 12 SNPs with a potential association to be further analysed. The 12 SNPs are SNP_A-1780818, SNP_A-1811809, SNP_A-2147541, SNP_A-1859018, SNP_A-4271922, SNP_A-1833194, SNP_A-2212963, SNP_A-4280587, SNP_A-2083541, SNP_A-2155029, SNP_A-4273904 and SNP_A-2005462 (one of the 6 SNPs which was in LD). The genotypic test is preferred (most powerful) when the genetic risk model is not additive.

In the original WTCCC study, there was a cluster of SNPs in the base position of range 114.5-115.0Mb in chromosome 10 that showed strongest association signal for T2D. It was represented by the reference SNP *rs*4506565 (or SNP_A-2005462). This SNP was said to be in LD with *rs*7903146, the variant with the strongest aetiological claims. Additionally, it was proven that *rs*7903146 had caused *rs*4506565 to give strong association signals. In our study, when we used the genotypic test of association, we found 6 SNPs in the base position of range 114.5-115.0Mb. The alleles of these 6 SNPs were in high LD with each other, one of the SNPs was *rs*4506565 (or SNP_A-2005462), and coincidentally was the one found in the original study. This SNP had a moderate *p*-value of 6.618×10^{-8} (after FDR correction). The original study had used two control groups, an extensive quality control measures as well as supplementary approaches in the case-control tests of association. Nevertheless, our study may not be powerful and may be considered inconclusive. Further tests and experimental investigations are required to make conclusion on the remaining 11 SNPs that has potential association using genotypic test. Table 4 below shows the comparison of *p*-values found in this present study and the original study.

Table 4. Information on SNP_A-200546 or rs4506565 (the p-values in the present study are calculated after applying FWER correction)

	SNP	Base Position	Genotypic *p*-value
Present study	2005462	114746031	6.62e-08
Original study (WTCCC)			5.05e-12

7. Conclusion

The objective of the present study was to find SNPs that has an association with the risk of T2D for the British Caucasian population. In the present study, we assumed the common SNPs to be the cause of T2D and excluded the rare SNPs (with MAF < 0.01) before performing any tests of association. In doing so, we might have excluded SNPs that had genuine association with the risk of T2D. Furthermore, we used tests of association using the assumption that the risk model will either be general or additive. In fact, there was a possibility that the true risk model was dominant or recessive. Since the true genetic risk model for T2D is not known, we were unable to decide the best test of association to use. There were a lot of assumptions being made when conducting this study and it may raise questions on the quality of the results. In the original study, only one SNP (SNP_A-2005462) showed a strong association signal for T2D, which was proven to have indirect association to the risk of T2D. In our study, when we assumed the genetic risk model to be general and applied the genotypic test of association, we were successful in identifying SNP_A-2005462 to have potential association. Even though we took a different approach and made a small comparison to the original study, we did manage to find the important SNP. However, we identified an additional 11 SNPs under the genotypic test as having potential association. Further analyses using other case control testing may be required to justify whether there may be other direct associations or false associations. Although, this present study is very limited, it still highlights the potential of finding the SNPs that have a true association with the risk of T2D.

References

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Arun, C. S., Ngugi, N., Lovelock, L., & Taylor, R. (2003). Effectiveness of screening and preventing blindness due to diabetic retinopathy. *Diabetic Medicine, 20*, 186-190. http://dx.doi.org/10.1046/j.1464-5491.2003.t01-1-00899.x

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263-265. http://dx.doi.org/10.1093/bioinformatics/bth457

Blakesley, R. E., Mazumdar, S., Dew, M. A., Houck, P. R., Tang, G., Reynolds III, C. F., & Butters, M. A. (2009). Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology, 23*(2), 255. http://dx.doi.org/10.1037/a0012850

Bretz, F., Landgrebe, J., & Brunner, E. (2005). Multiplicity Issues in Microarray Experiments. *Methods of Information in Medicine, 44*(3), 431-437.

Bodmer, W., & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics, 40*(6), 695-701. http://dx.doi.org/10.1038/ng.f.136

Diabetes UK. (2014). Diabetes: Facts and Stats. Retrieved from http://www.diabetes.org.uk/Documents/About%20Us/Statistics/Diabetes-key-stats-guidelines-April2014.pdf

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics, 81*(3), 559-575. http://dx.doi.org/10.1086/519795

R Development Core Team. (2010). R: A language and environment for statistical computing. Viena, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Wellcome Trust Case Control Consortium (WTCCC). (2007). Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*, 661-678. http://dx.doi.org/10.1038/nature05911

World Health Organization. (1999). *Definition, Diagnosis and Classification of Diabetes mellitus and its complications.* Geneva, World Health Organization.

Ziegler, A., & Konig, I. R. (2006). *A Statistical Approach to Genetic Epidemiology. Concepts and Applications.* Weinheim, Germany: Wiley-VCH.

Zimmet, P., Alberti, K. G., & Shaw, J. (2001). Global and societal implications of the diabetes epidemic. *Nature, 414*, 782-787. http://dx.doi.org/10.1038/414782a

Note

Note 1. National Human Genome Research Institute (NHGRI). National Institutes of Health. Talking glossary of genetic terms. Retrieved fromhttp://www.genome.gov/glossary, and Genome-wide association studies. Retrieved from http://www.genome.gov/20019523.

Copyrights