A bioinformatics approach to the identification of type 2 diabetes susceptibility gene variants in Africans

Ovokeraye Hilda Oduaran

A thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy

March 2014

Supervisor: Professor Michèle Ramsay

Co-supervisor: Professor Nigel Crowther

Declaration

I, Ovokeraye Hilda Oduaran, declare that this thesis is my own, unaided work, unless otherwise specified in the text. It is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other university.

…….. ……

Ovokeraye Hilda Oduaran Date

Abstract

Type 2 diabetes (T2D) is a metabolic disease that results from complex interactions between the environment, the genetic variation and epigenetic regulation of gene expression in individuals. Beta-cell dysfunction and insulin resistance are regarded as the hallmarks of the disease as the common presentation of T2D is the inability of beta-cells to adequately respond to the insulin demands of the body. The prevalence of T2D in Africa, and particularly South Africa, is on the rise. This is very likely the result of the combination of genetic susceptibility with increasing availability and accessibility of relatively cheap, highly palatable, calorie-dense meals with no corresponding lifestyle adjustment.

This study aims to utilize available data from GWAS and gene expression arrays to identify potential variants that likely influence T2D susceptibility in African populations. Two public data repositories were mined – the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) and the National Human Genome Research Institute's (NHGRI) GWAS Catalog. The criteria for selecting the studies for inclusion were based on ten descriptive T2D-related terms taken from the GWAS catalog's pre-defined search categories. These terms were also applied to the selection of gene expression studies in GEO. These terms are: "fasting glucose-related traits", "fasting insulin-related traits", "fasting plasma glucose", "insulin resistance/response", "insulin traits", "diabetes-related insulin traits", "pro insulin levels" "Type 2 diabetes", "type 2 diabetes and 6 quantitative traits" and "type 2 diabetes and other traits". Ten Affymetrix platform-based studies in human tissues were chosen from GEO using these criteria. A Benjamini-Hochberg adjusted p-value of 0.05 was set as a cut-off for significant differentially expressed genes (7,887 genes) with 497 genes occurring in two or more studies, based on tissue- or array-type, considered candidates for downstream analysis. The GWAS catalog presented 175 "reported" genes and 218 SNPs from 51 studies matching the set T2D-related criteria.

Functional analyses done with the Database for Annotation, Visualization and Integrated Discovery (DAVID) on both the GWAS and expression studies genes lists, with similar parameters, provided enriched gene lists. The union of both lists gave a core list of 140 genes for further analyses. These genes were used to retrieve

corresponding SNPs from the 1000 Genomes data set. The choice of this database stems from the presence of whole genome sequence data, albeit relatively low coverage (4X – 6X), of individuals in several populations on different continents. The populations of interest in this study, however, were the LWK (Luhya in Webuye, Kenya), YRI (Yoruba in Ibadan, Nigeria), CEU (Utah Residents with Northern and Western European ancestry), TSI (Toscani in Italia), CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan). The choice of these populations as proxies for each of the three continents was based on data availability in this version of the 1000 genomes data. Assessing the distribution of the risk allele frequencies of the GWAS SNPs across these populations showed the risk of T2D to be highest in the African populations. Intercontinental comparisons of SNPs provided a means to identify possible variations in SNP frequencies and occurrences between populations.

Fixation indices (F_{ST}) analysis was done on pairwise combinations of each African population with a non-African population resulting in 6 pairwise comparisons. These comparisons were combined by continent with the intersecting SNPs across continents providing the African vs non-African comparison, which is the main focus of this study. F_{ST} analysis produced a list of 7 genes (Notch homolog 2 (*NOTCH2)*, Kinesin Family member 11 (*KIF11),* Nuclear Receptor Subfamily 2 (*NR2F2)*, Ribosomal Protein L35A (*RPL35A)*, Small Nuclear Ribonucleoprotein D1 Polypeptide 16kDa (*SNRPD1)*, RNA Guanylyltransferase And 5'-Phosphatase (*RNGTT),* expression analysis and RNA binding motif protein 38 (*RBM38*)) from 228 SNPs showing significant differentiation between the African continent and European/Asian continents. The integrated haplotype score (iHS) was used to identify possible signatures of selection in the LWK and YRI populations. One of these seven genes, *SNRPD1,* was also selected, based on iHS results. Its involvement has been proposed in the spliceosome pathway and the RNA splicing process. Interestingly, it has not been shown to be associated with T2D in existing literature.

The identification of specific variants in T2D susceptibility genes in Africans will not only contribute towards a knowledgebase that will be useful for developing genotyping arrays that better represent African enriched variants, but also very likely

give some insight into the missing heritability component of the pathogenesis of the disease. Such knowledge will, in the long-term, lead to better targeted therapeutics for managing T2D.

Acknowledgements

I would like to express my immeasurable gratitude to the following people and institutions:

My supervisors, Prof. Michele Ramsay and Prof. Nigel Crowther for giving me this opportunity to pursue my academic goals. Thank you for the guidance, encouragement, support, patience and the very many thought-provoking discussions that have led to the completion of this task.

The NRF, for the grant-linked bursary held by Prof. Ramsay that provided the financial support for the duration my PhD studies.

The staff of Wits Bioinformatics and the SBIMB for the incredibly conducive study environment – I couldn't have asked for a better group. Prof. Scott Hazelhurst, Dr. Ananyo Choudhury, Shaun Aron and Phelelani Mpaganse – thank you for the seemingly random ideas and assistance when I found myself stuck computationally. Freedom Mukomana, thank you for the help with formatting during the write-up process.

My lovely family – mumsie, mum, dad, Ovo, Mudi, Ejiro and Ufuoma - for all the encouragement, love and support, especially with the kids, throughout this process, I am very grateful. Michael, Abigail and Andrew, thank you for the motivation and constant reality checks that keep me grounded daily. Veray, you remain my biggest cheerleader and a blessing in my life – thank you for your patience, love, unfathomable support and your presence in my life. Your outlook on life and sheer determination inspires me daily.

My friends, from the onset and those I have been blessed to meet in the course of this journey – thank you for all the support. Pauline, Johana and Lister – thank you for the encouragement and sometimes tough love at different aspects of this project.

And finally and most importantly to God, for everything.

Poster Presentation

Faculty of Health Sciences Research Day, University of the Witwatersrand (2012)

- Oduaran OH, Choudhury A, Crowther N, Ramsay M. A bioinformatics approach to the identification of type 2 diabetes susceptibility gene variants in Africans

South African Genetics & Bioinformatics Society Conference (2012)

- Oduaran OH, Choudhury A, Crowther N, Ramsay M. A bioinformatics approach to the identification of type 2 diabetes susceptibility gene variants in Africans

Table of Contents

List of Tables

List of Figures

Abbreviations

AADM	The Africa America Diabetes Mellitus Study
AFR	African populations - LWK and YRI
ASN	Asian populations -CHB and JPT
BMI	Body Mass Index
CDC	Centers for Disease Control and Prevention
CEU	Utah residents with ancestry from northern and western Europe
CHB	Han Chinese in Beijing, China
CVD	Cardiovascular Disease
DAVID	Database for Annotation, Visualization and Integrated Discovery
DKA	Diabetic Ketoacidosis
EHH	Extended Haplotype Homozygosity
EUR	European populations - CEU and TSI
GDM	Gestational Diabetes
GEO	Gene Expression Omnibus
GO	Gene Ontology
GWAS	Genome-Wide Association Studies
H3Africa	Human Hereditary and Health in Africa consortium
HuGE	Human Genome Epidemiology
IDF	International Diabetes Federation
iES	iHS Enrichment Score
iHS	integrated Haplotype Homozygosity Score
JPT	Japanese in Tokyo, Japan
KPD	Ketosis-Prone Diabetes
LD	Linkage disequilibrium
LDL	Low Density Lipoprotein
LWK	Luhya in Webuye, Kenya
MODY	Maturity Onset Diabetes of the Young
SNP	Single Nucleotide Polymorphism
T1D	Type I Diabetes
T2D	Type 2 Diabetes
TSI	Toscani in Italia

WHO	World Health Organization
YRI	Yoruba in Ibadan, Nigeria

1. CHAPTER ONE – INTRODUCTION

1.1 Pathophysiology of Type 2 Diabetes (T2D)

Type 2 diabetes (T2D) is a complex disorder of multiple interactions between the environment and genetics, with metabolic implications. Although its aetiology is not fully understood, it is known to be characterized by the failure of the body to respond normally to insulin, and the inability of the body to compensate for this condition to maintain euglycaemia. The failure of the pancreatic beta-cells to adequately respond to an increased demand for insulin manifests clinically as elevated postprandial blood glucose and fasting levels which are some of the diagnostic criteria for T2D [1–3]. T2D is very often correlated with obesity. Despite this high correlation, obesity, a major confounder in T2D, does not necessarily result in T2D as studies have shown that ~10% of individuals with T2D are neither obese nor overweight [4]. It is also important to note that not all obese individuals become diabetic. Those who become diabetic have beta cells that are unable to maintain high insulin output to overcome the resistance, with beta cell failure then leading to diabetes. Non-obese type 2 diabetics tend to have beta-cell dysfunction but not necessarily insulin resistance, whereas obese T2D individuals are both insulin resistant and have reduced insulin secretory output [5], highlighting the problem of dysfunctional insulin response.

It has also been proposed that low birth weight owing to *in utero* nutritional deprivation influences susceptibility to obesity, T2D and other metabolic conditions later in life through the acquisition of a 'thrifty phenotype' [6]. The thrifty phenotype hypothesis suggests that fetal and infant malnutrition, and consequently poor growth result in permanent changes in glucose-insulin mechanism [6]. Some of these changes are reduced insulin secretion capacity and insulin resistance, which when combined with other risk factors like obesity, physical inactivity and ageing are critical in the development of T2D [6]. This association between poor fetal and infant nutrition and weight with obesity and T2D first highlighted by Hales and Barker [7] has been demonstrated in several studies including the Chinese Famine (1959 – 1961) Study [8] and the Dutch Hunger Famine birth cohort study [9].

1.1.1 Insulin Resistance

Insulin resistance is a physiological state where the cells in the body respond abnormally to the actions of the insulin hormone. It can result from several factors including increased levels of adipocyte-derived free fatty acids. This can inhibit insulin-dependent glucose uptake and lead to reduced glucose utilisation and oxidation in the muscle [10]. However, the main cause of insulin resistance is obesity. Other contributors to insulin resistance are adipokines. Adipokines are proteins produced and secreted by adipocytes [11]. These include interleukin-6 (*IL-6*), tumor necrosis factor alpha (*TNF-α*) and adiponectin. In obesity, the expression of *IL-6* and *TNF-α* is increased while adiponectin is under expressed. Studies have suggested that *IL-6* produced by visceral adipose tissue can directly affect hepatic metabolism as its venous drainage goes directly through the portal vein to the liver, making it possible for *IL-6* to contribute to triglyceridaemia by its stimulation of the hepatic secretion of very low density lipoprotein [12]. *TNF-α* is a pro-inflammatory cytokine that is involved in the pathophysiology of insulin resistance possibly through the mechanism of interfering with the phosphorylation of the insulin receptor substrate (*IRS-1*) [13]. Studies have also shown adiponectin to augment lipid oxidation in the skeletal muscle and to reduce hepatic glucose production in the liver [14, 15] and to increase whole body insulin insensitivity. An increased lipid oxidation implies less ectopic fat deposition that can also aggravate insulin resistance [16]. Obesity can lead to insulin resistance not only by increased adipokine levels but also by increased ectopic fat deposition in the skeletal muscle which could lead to hyperglycaemia. It is also possible that sedentary lifestyle habits and unhealthy diet could serve as initiators of a cascade of events leading to the development of insulin resistance [17].

1.1.2 Beta-cell Failure

The inability of existing pancreatic beta-cells to either increase insulin secretion or proliferate to compensate for the increased insulin need resulting from a situation of insulin resistance is regarded as a condition of beta-cell dysfunction. The exact mechanism of beta-cell failure in T2D is controversial [18], but it has been proposed that the endocrine islet's homeostatic function is poor and unable to cope with environmental or metabolic stressors such as obesity, ageing and pregnancy where levels of insulin resistance are known to increase [19]. This is because the manner of physiological adaptation of beta-cell functions to conditions like ageing and

pregnancy which is mainly achieved through the modulation of beta-cell replication is very taxing for pre-diabetic beta cells [20]. Another widely believed explanation for beta cell dysfunction is that the stress imposed on the beta-cells to increase insulin production to allow for glucose clearance from the blood wears out the cells and eventually lead to cell death which reduces the number of beta-cells producing insulin and consequently results in insufficient insulin production. Another explanation, originating from mouse models [20] is that these beta-cells do not necessarily die (loss of cells) but dedifferentiate to the original progenitor cell form [21] (loss of function) and require the correct molecular signals to be re-differentiated back to adult insulin-producing beta-cells.

Clarity on the mechanism of beta-cell dysfunction as well as insulin resistance will very likely provide more insight into the pathogenesis of the disease and possibly therapeutic solutions as these conditions are central to the development of T2D.

Figure 1.1 shows a proposed schema for the pathogenesis of T2D.

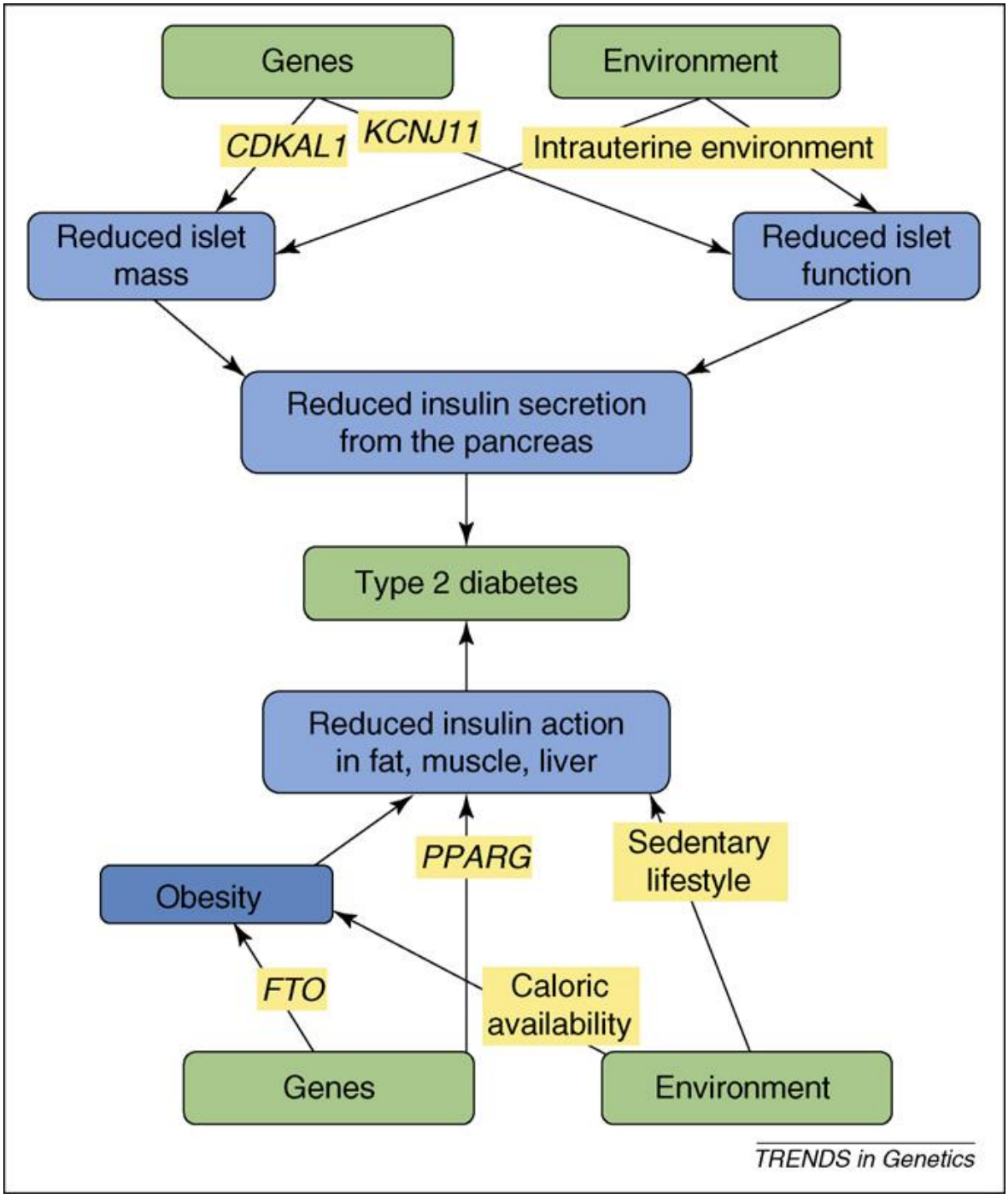


Figure 1.1 A proposed schema for the pathogenesis of T2D.

Interactions between environmental and genetic factors contribute to the processes involved in T2D with some of the genes and exposures shown in yellow. *FTO* = fat mass and obesity associated gene. *PPARG* = peroxisomal proliferator-activated receptor-g. *CDKAL1* = CDK5 regulatory subunit associated protein 1-like 1. *KCNJ11* = potassium inward rectifying channel, subfamily J, member 11. [22]

1.2 Epidemiology of the Disease

Despite the seemingly long history of T2D worldwide, it has only been recently identified as a major problem on the African continent. This can be attributed to the increasing industrialisation and acculturation [23] in Africa, with high caloric diets replacing the high fibre diets and no corresponding lifestyle adjustments. The research efforts have not been as extensive as they have been in infectious diseases like HIV/AIDS, malaria, diarrhoea, respiratory infections and tuberculosis, which according to current statistics, account for the majority of the morbidity and mortality on the continent [24]. Recent projections, however, indicate that the morbidity and mortality rate of non-communicable diseases like T2D and cardiovascular disease, a T2D comorbidity, will soon exceed those of the infectious diseases [24]. With the rapidly increasing burden of the disease, it is imperative to accelerate research efforts in the field.

1.2.1 Worldwide Prevalence and the Increasing African Burden

In a global study to estimate the prevalence of diabetes across all age-groups for the years 2000 and 2030, an increase of over 110% in the total number of people worldwide was predicted in the 30-year period [25]. The prediction for the prevalence of diabetes in that same period in sub-Saharan Africa is a projected rise from 7.2 million in the year 2000 to 18.7 million in 2030 [25]. This represents a massive 161% increase in the proportion of individuals with diabetes, which is about a 50% difference from global predictions. This study, however, acknowledged that these numbers most likely underestimate the future prevalence of diabetes as the incidence of obesity continues to rise [25]. The 6th edition of the International Diabetes Foundation (IDF) predicts a rise from 382 million people with diabetes in 2013 to 592 million people by 2035 globally (figure 1.2), 80% of whom live in low- and middle-income countries [26]. Interestingly, about 90% of all diabetic cases reported are T2D [27]. The remaining ~10% include (but is not limited to) Type I Diabetes (T1D), Gestational Diabetes (GDM), Maturity Onset Diabetes of the Young (MODY) and Ketosis-Prone Diabetes (KPD). T1D is an autoimmune disease where the insulin-producing cells in the pancreas are attacked by the defence system of the body [28]. GDM is a condition of high glucose levels in the blood that occurs during pregnancy. It is defined as any degree of glucose intolerance with onset or first recognition during pregnancy [29]. Although a temporary condition, the risk of

developing T2D later on has been shown to be higher in women who have had GDM compared to those who have not [30]. MODY, often referred to as monogenic diabetes [31], is a hereditary form of diabetes that presents itself in multiple forms. It results from mutations in an autosomal dominant gene that interferes with the production of insulin. KPD is a rather unique form of diabetes as it has been known to affect only individuals of non-Caucasian ethnicity [32]. It is regarded as a heterogeneous disease characterised by presentation with diabetic ketoacidosis (DKA) in individuals who do not necessarily fit the typical characteristics of other traditional diabetic categories as defined by the American Diabetes Association (ADA) [33]. However, T1D and T2D patients have been known to present DKA under infection or stress conditions [34]. Interestingly, in all of these forms of diabetes, the treatment course almost always ends in insulin therapy.

Figure 1.3 shows the worldwide distribution of the number of people with diabetes in 2013.The prevalence statistics of diabetes in 2013 can be seen in figure 1.4 while figure 1.5 shows the percentage of diabetes-related deaths in individuals below the age of 60 in that same year.

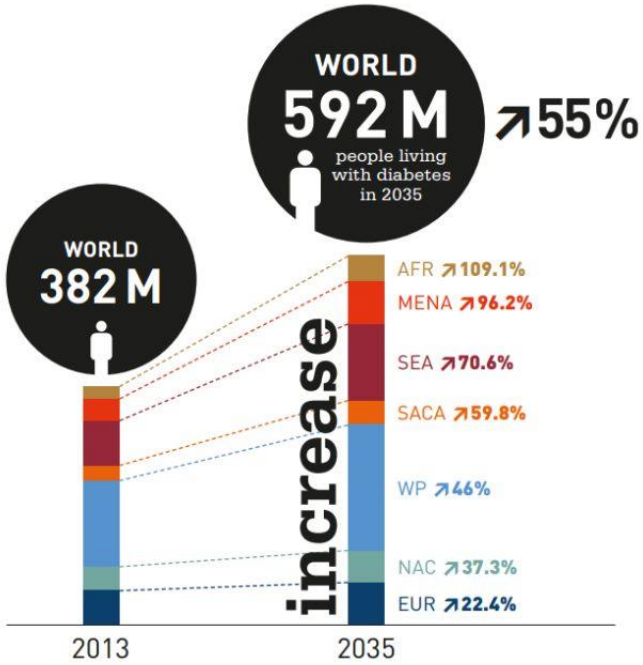


Figure 1.2 IDF global predictions for diabetes incidence increase between 2013 and 2035.

AFR = Africa, MENA=Middle East and North Africa, SEA = South East Asia, SACA = South And Central America, WP = Western Pacific, NAC = North America and Caribbean, EUR = Europe [26].

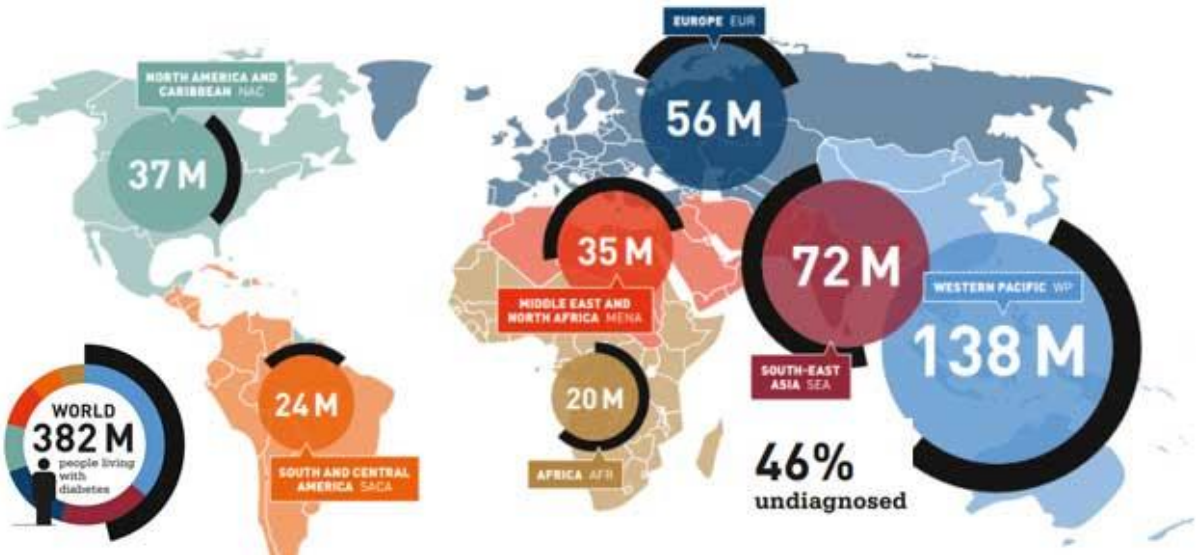


Figure 1.3 Global distribution of people (in millions) with diabetes in 2013 [26].

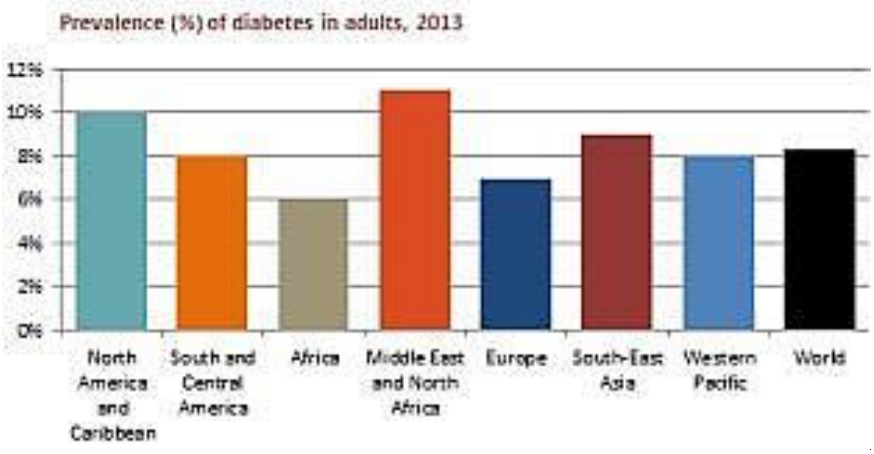


Figure 1.4 Global prevalence of diabetes (%) in 2013 [26].

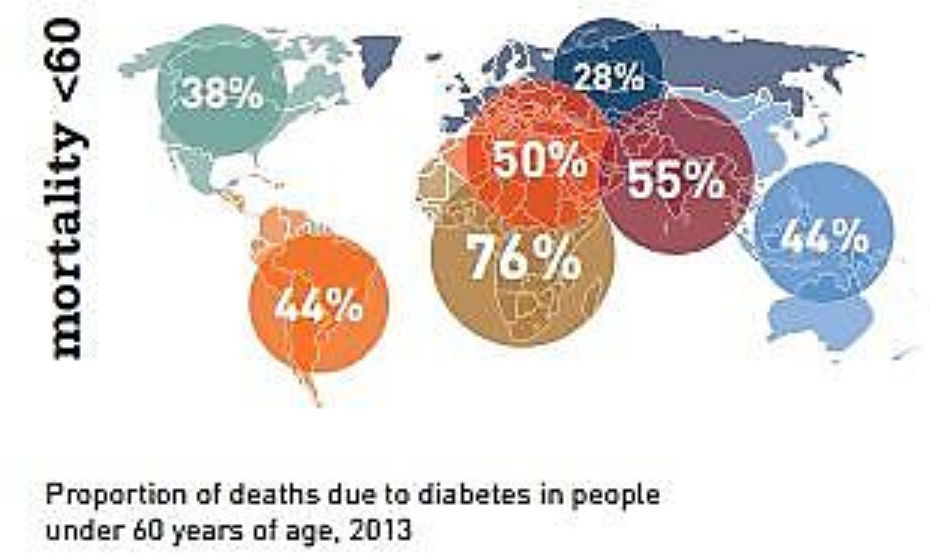


Figure 1.5 Global percentages of diabetes-related mortalities in 2013 [26].

The mortality rates shown in figure 1.5 are highest in Africa even though the continent had the lowest prevalence percentage. This again points out the need for more African-focused research efforts as the burden of the disease on the African continent is increasing greatly. Just like it has unravelled in more developed economies, the increasing prevalence of obesity on the continent is leading to a diabetes epidemic. Age-specific data have also indicated an increase in diabetes rate with age in sub-Saharan Africa [35, 36].

Obesity, nonetheless, remains a major risk factor to the development and pathogenesis of T2D. It generally occurs as a consequence of an imbalance

between energy intake and expenditure that is usually characterized by increased fat storage [37]. It results chiefly from a combination of genetics, low level of physical activity, and an environment where relatively cheap, highly palatable, calorie-rich food is widely available [38, 39]. It is therefore not surprising that the number of obese or overweight people now exceeds the underweight population worldwide [40]. However, the concept of metabolic obesity (the existence of the metabolic profile associated with obesity in non-obese subjects) which is the result of the accumulation of fat in the abdominal viscera in non- obese individuals [41], has been proposed to possibly provide an explanation for high T2D incidences in populations with seemingly normal weight individuals [42]. Asian individuals tend to develop T2D at lower body mass index (BMI) levels than Europeans [43] and the risk of developing T2D is generally higher in Asians than in Europeans [44] despite the mostly lower prevalence of overweight or obese individuals in Asian populations. Studies have shown Asians to be more likely to have a higher fat percentage or visceral adiposity than Europeans at a given BMI or waist circumference [45, 46]. Waist circumference is a measure of central adiposity. Measures of central adiposity have been shown to be better indicators of T2D risk than BMI [44].

An obesity pandemic is a prelude to increased prevalence of T2D and other non-communicable diseases, as type 2 diabetics are at a high risk of developing a range of disabling conditions like hypertension, stroke, renal failure and cardiovascular disease (CVD) [4]. It is of little surprise then that the prevalence of diabetes and CVD is on the rise in Africa as obesity levels increase across the continent.

It is important to note that in addition to the changing physical activity levels and diet, ongoing demographic changes like urbanization as well as political and economic instability contribute considerably to the burden of diabetes and other non-communicable diseases on the African continent [24].

1.2.2 Comorbidities and Complications

As has been previously indicated, the comorbid conditions and complications of T2D contribute to the burden of the disease. Some comorbidities and complications of T2D include hypertension, dyslipidemia, stroke, microvascular diseases (renal failure, neuropathy, and retinopathy) and CVD. Stroke is, more often than not, a consequence of CVD, and CVD results from complex interactions between genetic

and environmental factors together with high risk factors like diabetes and hypertension [47]. It is the most common cause of death globally [24] and, as is stated in the previous section, is increasing in prevalence in tandem with T2D.

Diabetic neuropathy results from damage to the nerves as a result of high blood pressure and high glucose levels with the most commonly affected areas being the extremities particularly the feet. Similarly, diabetic retinopathy and nephropathy (kidney failure) can be caused by consistently high blood sugar levels together with high cholesterol and high blood pressure [48, 49]. The consequence of this in nephropathy is the damage to the tiny vessels in the kidneys that serve as waste filters from the blood which implies inefficient functioning of the kidneys or in some cases, total failure. In retinopathy, the blood vessels of the retina are damaged which may potentially lead to blindness. The risk for other vision-affecting conditions like cataracts and glaucoma is also increased in T2D [48, 49]. The microvascular diseases are generally linked to hyperglycemic conditions with hyperglycemia-induced abnormalities in the hexosamine, polyol and protein kinase C pathways having been implicated in the mediation of tissue damage in T2D [50, 51]

Hypertension is a chronic condition of elevated blood pressure. About 75% of individuals with T2D also have hypertension and those with hypertension on its own show indications of insulin resistance [52]. The risk factors and complications of hypertension overlap significantly with those of T2D [53] which in some cases of CVD could result from the continuous strain imposed on the heart by elevated blood pressure.

Dyslipidemia is a condition that can result from prolonged increased insulin levels in the blood which can cause insulin resistance and consequently free fatty acid flux from the insulin resistant adipocytes [54–56]. It thus makes logical sense that in T2D, the phenotype associated with dyslipidemia is mainly attributed to insulin resistance and insulin deficiency [57, 58]. This phenotype generally presents 3 characteristic features – low HDL cholesterol concentration, increased concentration of small dense LDL cholesterol particles and high plasma triglyceride concentration. T2D is not only an independent risk factor for CVD, it also amplifies the effects of other common risk factors like hypertension and smoking. Dyslipidemia in T2D contributes to the risk of the development of CVD.

Although some of these complications could be disabling and even life-threatening, they could be stalled by good management of the disease, which includes proper glucose levels monitoring.

The complications and comorbidities associated with T2D not only complicate the management of the disease but also make it a multi-faceted challenge with an enormous burden on an already over-tasked healthcare system.

1.3 Genomic Contributions to Type 2 Diabetes

Global T2D research efforts have been able to identify a number of genetic variants with risk associations to the disease. A recent review showed variants in about 60 genes to have significant associations with the development and pathogenesis of T2D [59]. However, this list excluded genes like potassium inward rectifying channel, subfamily J, member 11, *KCNJ11* and calpain 10, *CAPN10* – genes identified in candidate gene analyses and family-based linkage studies. Some of the confirmed risk variants include rs7903146 (transcription factor 7-like 2, *TCF7L2*), rs12255372 (*TCF7L2*), rs1111875 (hematopoietically expressed homeobox, *HHEX*), rs5215 (*KCNJ11*), rs9939609 (fat mass and obesity associated gene, *FTO*) and rs1801282 (peroxisomal proliferator-activated receptor-g, *PPARG*) [60]. Studies have indicated, however, that all the risk variants together account for only about 10% of the genetic contribution to the disease [61, 62]. T2D concordance ranges between 30% and 80% as have been shown in monozygous twin studies and heritability of the disease has been indicated at 30-70% in family studies [57, 63].

The bulk of T2D research has been done in non-African populations. With the knowledge that T2D has a strong genetic component, albeit not entirely resolved, it is expected that there exists a considerable ethnic variation in the prevalence of the disease [50]. The specificity of a locus in the *HHEX* gene in conferring T2D risk in East Asians [64] demonstrates, to an extent, this ethnic-specificity. A few notable studies carried out in populations in West Africa and of West-African or African origin, in the United States, have identified gene variants and significant linkages in regions previously identified as T2D susceptibility loci in other populations [65]. The Africa America Diabetes Mellitus (AADM) Study [65], the longest running genetic epidemiological T2D study that involves Africans, included West African participants whose data have been included as a replication cohort in several genome-wide

association studies (GWAS). One such collaborations led to the observation that prostate cancer risk variants, rs4430796 and rs757210, in the transcription factor 2, *TCF2* (also known as hepatocyte nuclear factor 1 beta, *HNF1B*) gene confer protection against T2D [66]. Collaborations with deCODE Genetics elucidated the impact of recent selection in the *TCF7L2* gene in relation to T2D risk [67] and confirmed genome-wide significant associations of CDK5 regulatory subunit associated protein 1-like 1, *CDKAL1* variants in T2D susceptibility [68]. *TCF7L2* has produced one of the strongest signals and most widely reproduced associations with T2D risk in several ethnic groups [69], including Africans as the deCODE-AADM collaboration showed.

However, no GWAS has been conducted on T2D among populations on the African continent as of yet [24]. This creates a gap in our knowledge that needs to be bridged as the diverse nature of African populations could provide unprecedented insight into the missing heritability of T2D as some variants might be more relevant and therefore easier to pick up in cohorts of non-European ancestry. Also, since most of the current genome-wide study approaches are best suited to the identification of common variants, it has been speculated that rare variants could perhaps provide some more insight into the missing heritability component [59].

1.4 Study Approaches to Identifying Genomic Contributions to the Aetiology of Complex Traits

Like in many complex diseases, the global research trend in the genetics of T2D has gone through 3 main phases: family-linkage analysis and candidate-gene studies, association analysis and genome-wide association studies (GWAS) [2]. In addition, computational approaches, copy number variation (CNV) analysis, sequencing approaches, epigenetic methods as well as microarray gene expression analysis, an experimental approach based on wet lab experiments and subsequent computational analysis, have been used to prioritize candidate genes and pathways in complex diseases [70, 71]

1.4.1 Linkage/Association Analysis

Linkage analysis presents researchers with a method to map the location of disease-causing genes by the identification of genetic markers that are co-inherited with the phenotype of interest. This means that a marker that co-segregates with a gene of

interest could be used to track the gene within a family without prior knowledge of the mutation as long as it is absent in unaffected family members. It thus relies on matching of the disease with genetic markers of known location [2]. This method was effective in identifying genes in monogenic forms of the disease but not necessarily the complex form of the disease as it has had limited success as a result of weak genotype-phenotype association which is a hallmark of complex multifactorial diseases [72].

Association analysis evolved to improve on simple family linkage analysis. This fundamentally more powerful method of susceptibility gene discovery depends on robust correlations of the allele frequencies of the genetic variants to the disease [2]. Most genetic association studies examined a single variation or a set of polymorphisms near a single gene or focused on a region defined by a linkage peak determined by a family study [73]. A major challenge with this method is that the causal variant itself or a marker to which it is tightly correlated has to be directly examined for a signal to be detected. This means that attention had to be directed to specific genes or variants of interest, some of which may have been inappropriate candidates [74]. Despite the low power of these methods, genuine susceptibility variants were discovered over the course of several studies. Some of these include *CAPN10* [75, 76] and *KCNJ11* [77].

1.4.2 Candidate Gene Studies

This approach focuses on genes that have been selected for study based on the biological characteristics of the phenotype of interest and a supposed match between the presumed or known functions of the genes [2]. It is pretty much an association study that directly tests the effects of genetic variants of a potentially contributing gene [78]. Candidate gene studies, therefore, limit the number of tests for variant association to a small subset of the genome and focus hypotheses on sets of genes that are believed to be associated with the phenotype of interest based on prior knowledge [73].

1.4.3 Genome-wide Association Studies (GWAS)

Genome wide association studies (GWAS) have proven to be very successful in gene and gene variant identification owing largely to the fact that the entire genome is queried at once. It involves systematic, large-scale searches for statistically significant associations between disease and specific DNA sequence variants (e.g.

single nucleotide polymorphisms (SNPs)) [2] by querying for differences between affected and unaffected groups. GWAS has been especially successful in elucidating genetic variants that influence T2D and BMI with several GWAS confirming that a variant in *TCF7L2* gene confers risk for T2D, while a variant in the *FTO* gene confers risk for obesity/BMI [59]. An advantage of GWAS over the other methods is its unbiased approach. Prior knowledge or hypotheses of associated genes or SNPs to the phenotype of interest is not necessary as the whole genome is being interrogated at once. A caveat then is the low power to definitively identify associations with such large number of tests being performed [73]. However, large sample sizes, often more than 10000 cases and controls compensate to increase the power to detect low to moderate effects. Most of the T2D-associated variants to date have been identified via GWAS.

1.4.4 Copy Number Analysis

Copy number analysis generally refers to methods of detecting copy number variation (CNV) from analysing data resulting from a queried DNA sample. Copy number variations are structural variations that are genomic DNA modifications that result in either functional or neutral variations in the number of copies of certain sections of the DNA. They usually correspond to relatively large deletions or duplications of genomic regions on certain chromosomes. CNVs can range from 1 kilobase to several megabases and have been reported to account for about 12% of the genome making their role in disease aetiology possibly significant [79]. CNVs can contribute to disease susceptibility by influencing the gene expression level [80]. Genome-wide association analysis using a logistic regression model was used to assess disease susceptibility loci for risk of T2D in a Korean population [81]. This study identified 3 new CNV regions to be significantly associated to the disease. Also, T2D-associated CNVs have been reported in the leptin receptor, LEPR, a gene that has been implicated in obesity and diabetes, from genome-wide SNP chip data using the quantitative multiplex polymerase chain reaction of short fluorescent fragment (QMPSF) method[82]

1.4.5 Epigenetics Methods

Epigenetics generally refers to changes in gene expression that result from chromatin structure changes that do not alter the DNA sequence. These are heritable changes that can occur throughout developmental stages as well as in

response to environmental factors [83]. One mechanism of epigenetic modification is the methylation of DNA. Modifications resulting from DNA methylation or histone modification alter chromatin structure (histone modification) which can in turn alter transcription patterns as a result of reduced gene accessibility to the components of the transcription machinery [84, 85].

DNA methylation is the tagging of DNA by a methyl group, usually from dietary sources, which can activate or repress the expression of genes. Methylation tends to occur primarily at CpG sites. A study examining genotypic-epigenotypic interactions in T2D that focused on previously known genomic susceptibility regions identified a risk allele in the *FTO* locus [86]. Also, individuals prenatally exposed to famine have been shown to harbour differentially methylated regions of genes of relevance to T2D [87]. An epigenomic approach thus contributes towards providing insight into the epigenetic and environmental factors that play roles in elucidating the aetiology of T2D.

1.4.6 Sequencing and Computational/Bioinformatics Approach

The development of parallel sequencing technologies has greatly expedited the discovery of human variations on a massive scale. This has been facilitated by the ability of next generation sequencing, NGS, technologies to relatively rapidly sequence entire individual genomes (whole genome sequencing, WGS) or just the coding regions of the genomes (whole exome sequencing, WES) [88, 89] The analyses of the resulting sequencing data can be potentially bioinformatics-intensive involving several computational steps and analyses pipelines. However, WGS and WES have been very useful in the identification of de novo mutations which include single nucleotide variations as well as short insertions and deletions (indels) in various complex diseases [90–94]. Several risk variants for T2D susceptibility in individuals of different ancestries have been identified via NGS technologies. Some of these include variants in the Early Endosome Antigen 1, *EEA1*, [95].

Computational approaches to elucidating the pathogenesis of T2D have also led to the prioritization of some candidate genes like lipoprotein lipase, *LPL* and Enoyl CoA Hydratase, Short Chain, 1, *ECHS1*, both involved in fatty acid metabolism [70].. Other bioinformatics approaches include the meta-analyses and integration of data from either homogenous or heterogeneous sources. The general idea behind this

approach is to utilize the huge body of available study data with the help of bioinformatics tools to provide possible insights into the disease via the identification of risk-associated variants. Some of these sources include results from microarray expression studies, GWAS, linkage studies, interaction screens and disease similarity, gene ontologies and pathways databases [96, 97]. Meta-analyses approaches are sometimes used to augment low to moderately powered GWA studies to present statistically valid results. These methods generally involve the building of an interaction network or evidence layers based on the different data sources with genes being subsequently ranked and prioritized based on statistically-determined criteria [97]. The ability to reuse and reanalyse data in an integrative manner makes bioinformatics approaches critical in the elucidation of the pathogenesis of complex diseases.

1.4.7 Rationale for Study Approach

Genome-wide association studies have identified more T2D-susceptibility variants than any other approach prior to its inception. The idea of a GWAS on the African continent would thus be ideal in the search for possible African-specific T2D risk variants but this is not the objective of this project. It is anticipated that the use of bioinformatics tools and methods to interrogate existing non-African-focused T2D studies, while analysing the results in the context of African population genetic variation via a database with whole genome sequence information from a number of African populations, would not necessarily identify novel genes but help to identify some novel risk variants for T2D that are important African contributors to the disease. Research efforts in T2D already indicate quite clearly that susceptibility variants vary across ethnicities as has been previously mentioned with the *TCF7L2* gene. It has been shown that the risk variants in this gene, although present in the Japanese and Chinese populations, do not confer the greatest risk to date in these populations as they do in Europeans - variants in the *KCNQ1* do that [98, 99]. Also, T2D risk variants in the *HHEX* have been indicated to show specificity to East Asian populations [64]. Similarly, variants (rs1800963, rs1028583, and rs3818247) in the hepatocyte nuclear factor 4 alpha, (*HNF4A)*, region which show T2D association in the Finnish population do not show such association in the Ashkenazi Jewish population even though other SNPs (rs4810424 and rs1884614) in the *HNF4A* region show similar association levels with diabetes in both populations [100, 101].

1.5 Study Aim and Objectives

This study is based on the hypothesis that genetic variants in already identified genes that contribute to T2D susceptibility differ between Africans and non-Africans (Europeans and Asians). The objective is to use a bioinformatics approach to identify genetic variants that are likely to contribute to T2D susceptibility in African populations.

1.5.1 Aim

To mine public domain data for genes and genetic variants associated with T2D, and to combine this with differential gene expression data, in order to facilitate the identification of T2D risk elements and to make comparisons between Africans and non-Africans, using a bioinformatics approach.

1.5.2 Objectives

i. To generate a list of genes and gene variants that have been previously identified to be associated with T2D and its relevant associated traits in genome-wide association studies.

ii. To assess the risk allele frequencies from the associated GWAS variants (SNPs) across populations.

iii. To perform functional enrichment on the GWAS gene list to retrieve highly-ranked genes for further analyses.

iv. To augment the GWAS-generated gene list with a functionally-enriched list of genes that have been previously identified in T2D and its relevant associated traits in microarray-based gene expression studies.

v. To analyse the augmented gene list in the context of African population genetic diversity to identify common and unique T2D risk variants.

2 CHAPTER TWO – METHODS

An overview/workflow for this research is shown in figure 2.1 below.

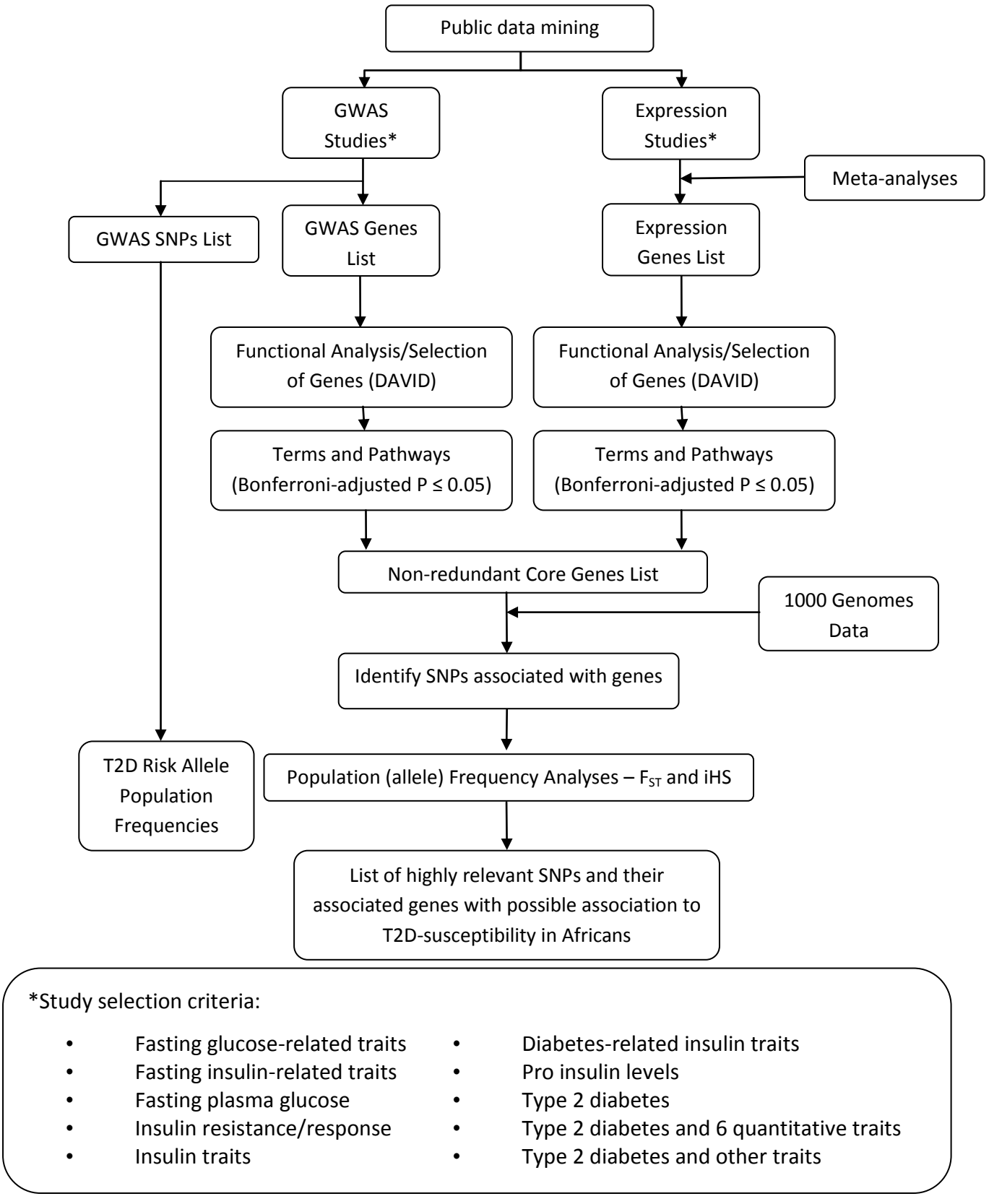


Figure 2.1. Research workflow.

2.1 Public Data Mining and Pre-processing

The retrieval of data from the public domain was done at two points – T2D-associated study data and SNP data.

2.1.1 T2D-associated Studies

Two types of publicly available T2D-associated data were used in this study: Genome Wide Association Studies (GWAS) data and gene expression studies data. The GWAS catalog, a resource from the National Human Genome Research Institute (NHGRI), was used as the primary resource for GWAS data in this study. It is a manually curated collection of published association studies identified from several sources including PubMed, NIH-distributed compilations of news and media reports, as well as from comparisons with HuGE Navigator [102], an online repository of published epidemiology literature [103, 104]. This catalogue is available for direct online query as well as for downloads for offline queries. The studies included in this catalogue attempted to assay at least 100,000 SNPs in the initial design and exclude studies in languages other than English. Information extracted from these studies by the catalog curators include the initial and replication sample sizes, strongest SNP/trait association in the study ("SNPs") as well as the risk allele if available ("Strongest SNP-Risk Allele") and the genes associated with the strongest SNP/trait association ("Reported Gene(s)"). For each SNP that was identified for inclusion in the catalog, chromosomal regions were extracted from the UCSC Genome Browser [105]. SNP association p-values, odds ratios ("OR or beta") and 95% confidence intervals ("95% CI (text)") were also reported. In the absence of a combined p-value in the study report, the p-value and effect-size from the largest sample size is reported if the initial and replication samples each show SNPs/trait associations that meet the threshold of $p < 10^{-5}$. The GWAS catalog generally does not include candidate gene-focused studies and SNP-trait associations with $P > 1.0 \times 10^{-5}$. This relatively liberal statistical threshold was chosen to accommodate GWAS scans of different sizes while using a consistent approach, and to allow for the possible examination of marginal associations [103, 104].

The National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) was used as the primary resource for microarray expression data in

this study. GEO is a Minimum Information About a Microarray Experiment (MIAME)-compliant public repository for researcher-deposited microarray, next generation sequencing and other forms of high throughput functional genomics data [106]. It stores data in well-organized hierarchies – Platform, Sample and Series records - that allow for easy accessibility. These are all submitter-provided records. Platform records contain summary descriptions of the array or sequencer. A data table that defines the array template is also included for array-based platforms. A platform record, although assigned a unique and stable GEO accession number (GPLxxx), is not specific to a single submitter and may refer to many samples submitted by multiple researchers on that specific platform. Sample records describe the biological samples, the handling and protocols it underwent and the abundance measurement of each element derived from it. A sample record is also assigned a unique GEO accession number (GSMxxx). In this case, however, the number is unique to the specific sample entity. Series records provide a description and focal point for the whole study as it links related samples together. Series records are assigned unique and stable GEO accession numbers (GSExxx). There is a higher level of organisation of GEO data – DataSet records. These are GEO series records that have been manually curated and reassembled by the GEO staff. A DataSet thus represents a collection of biologically and statistically comparable GEO samples [106]. They are also assigned GEO accession numbers (GDSxxx) and serve as the basis of GEO's suite of tools for data display and analysis.

The GWAS catalogue and GEO were queried for studies matching the following T2D-related terms: "fasting glucose-related traits", "fasting insulin-related traits", "fasting plasma glucose", "insulin resistance/response", "insulin traits", "diabetes-related insulin traits", "pro insulin levels" "Type 2 diabetes", "type 2 diabetes and 6 quantitative traits" and "type 2 diabetes and other traits". These are descriptive terms directly taken from the GWAS catalogue and applied to GEO.

- **GWAS Data**

- **Study Selection and Retrieval**

The GWAS catalogue was downloaded as a tab-delimited text file from the NHGRI website [103, 104] and queried offline for studies matching the 10 descriptive terms according to the aforementioned criteria. The text file was accessed with Microsoft's

Excel program and sorted with the in-built “sort” function. The “disease/trait” column was used as the primary sorting column for studies that matched the set criteria. These studies, together with their accompanying curated information, were selected and copied to a blank spread sheet. This spread sheet was adopted as the main document for the publicly retrieved GWAS data for this study**.**

- **Gene Expression Studies Data**

- **Genes and SNPs Lists Generation**

The genes from the “reported” genes column in the main document were selected to provide what is referred to as the “GWAS Genes list”. The “GWAS SNPs list” resulted from the “SNPs” column from the same spread sheet. For clarity in the downstream analyses, the lists will be referred to accordingly.

- **Study Selection**

GEO was queried directly online for T2D-related terms as previously indicated. This search was, however, limited to human data and studies done on the Affymetrix platform to reduce any possible bias that may result from the combination of studies done on different platforms. The choice of the Affymetrix platform stemmed from the comparably larger number of studies appearing in the search results that had been carried out on the platform at the time of this selection.

- **Data Retrieval and Analyses**

Each study was individually retrieved from GEO series records via GSE accession numbers and analysed using a number of Bioconductor packages in R. R is a computational language and an open source environment with a suite of software facilities for data manipulation, calculation and graphical display [107]. The basic functionality of R is easily extended by the availability of packages that cover a wide range of modern statistics and bioinformatics via its Bioconductor platform. Bioconductor is bioinformatics-centred and provides tools for high-throughput data analyses [108]. Bioconductor packages were frequently utilized for the analyses done in this study. For the expression data retrieval and analyses, the GEOquery [109], affy [110] and limma [111] packages were used.

First, the raw intensity data (.cel files) from each study together with their associated phenotypes (case (T2D), control, or other conditions), as specified by the originating study author, were retrieved and organized using the getGEO and getGEOSuppFiles functions in the GEOquery package. The GEOquery package serves as a bridge between the GEO repository and Bioconductor. It allows for data to be easily retrieved with very minimal to no manipulation. The retrieved raw data were read with the ReadAffy function, in the affy package, into an AffyBatch object. An AffyBatch object is an R class representation for probe level Affymetrix GeneChip data comprising mainly of multiple array intensities. This is an important data formatting step for some of the downstream analysis. The affy package contains a suite of tools for exploratory analysis of oligonucleotide arrays. Applying the rma function to the AffyBatch object, also contained in the affy package, pre-processed the data using the Robust Multichip Average (RMA) method [112]. This is a three-step method that involves background correction, quantile normalization [113] and summarization. These corrected data are then organized in a matrix of gene expression measures with the exprs function creating an expression set object (exprSet). Expression values, phenotypic and other related information are usually stored in exprSet objects. This was then used for differential gene expression analyses with the limma package.

The Linear Models for Microarray Data, limma, package is a Bioconductor implementation of the use of linear models to assess differential expression in microarray data. The general aim is to make an otherwise complex analysis relatively straightforward provided that the right parameters are designated. The limma approach was applied to this study by first specifying a ‘design matrix’ which indicates which RNA target has been hybridized to each array. This was done using the model.matrix function. This function identifies and organizes the different conditions present in the data set. A linear model is then fitted for each gene on the array, based on the design matrix, using the lmFit function, in order to systematically model the experimental data so that it can be differentiated from random variation. To make comparisons between conditions (as defined in the design matrix), a ‘contrast matrix’ (cont.matrix) was specified. Using the makeContrasts function, the contrast matrix was designed to directly compare T2D cases against controls, regardless of any other condition that may have been considered in the original

study. The contrasts.fit function allows for the fitted coefficients of the contrast matrix to be compared in several different ways. The eBayes function employs an empirical Bayes method on the fitted contrasts' coefficients to assess differential expression of probes and compute a number of statistics. Some of the resulting statistics from limma analysis include adjusted p-values for multiple testing, unadjusted p-values, log fold changes as well as F-, B- and t-statistics, which all contribute to the identification of probes that represent genes that have been differentially expressed in an experiment. The t-statistic in limma is a linear one that is easily applied to microarray data. It was developed from the hierarchical model of Lönnstedt and Speed [114, 115] and is based on a model where variances of the residuals vary from gene to gene and are assumed to be drawn from a scaled chi-square distribution [111] unlike the standard t-statistic where the assumption is that both groups are sampled from normal distributions with equal variances [116]. The moderated F-statistic in limma is an overall test of significance that combines the t-statistics of all the contrasts. It tests to see if any of the contrasts are non-zero for each gene, that is, if a gene is differentially expressed on any contrast. The B-statistic (B or lods) is the log-odds that a gene is differentially expressed [111].

The limma statistic utilized for gene selection in this study was the Benjamini-Hochberg (BH) adjusted p-value. This is a p-value that has been adjusted for multiple testing with the BH method to control the false discovery rate [117]. This simply means that, if a threshold of 0.05 is chosen, all the genes below this threshold are considered to be differentially expressed and the expected proportion of false discoveries in the set is controlled to be 5% of the set of genes. The use of adjusted p-values is important because p-values, as with other model-based methods, depend on mathematical assumptions and normality, which are not usually precisely true for microarray data [111].

The R script used for the gene expression studies analysis can be found in electronic (EA1).

- **Gene List Generation**

The differential expression analysis resulted in the identification of Affymetrix probes that needed to be annotated for clarity. The getSYMBOL function in the annotate package [118], together with the corresponding downloaded Affymetrix genome array annotation data, were used to match the differentially expressed Affymetrix probes to gene symbols. This provided a data frame for probes and their corresponding official gene symbols together with their limma statistics. A BH-adjusted p-value ≤ 0.05 was chosen as the cut-off for genes to be considered as being significantly differentially expressed in each of the selected studies. To create a core list of genes from these Affymetrix-based expression studies, the genes meeting the adjusted p-value criteria were only considered for further analysis if they were found to be significant in more than one of the selected studies based on tissue or array similarities. For example, comparisons were made between studies with samples originating from skeletal muscle tissue and not between studies with samples from the skeletal muscle and blood tissues. Likewise, studies done on the GeneChip Human Genome U133 array, for example, were directly compared with each other thereby allowing for relatively "like" comparisons. The resulting intersecting genes provided what is referred to as the "Expression Genes list".

2.1.2 SNP Data (1000 Genomes)

The 1000 Genomes data was mined to allow for the analysis of both the GWAS SNPs (risk allele frequency) and the SNPs associated with the genes that result from the functional analysis section in a population context.

To get some understanding of the genes via their corresponding markers in a population context, the 1000 Genomes project [119] data were interrogated. The 1000 Genomes project is a human genome reference project that aims to discover most of the genetic variants with frequencies of at least 1% in the studied populations by the low coverage (4X-6X) sequencing of many individuals. This includes individuals from different populations in the Americas, Asia, Africa and Europe. The populations of interest in this study from those present in the 1000 Genomes project are: LWK (Luhya in Webuye, Kenya), YRI (Yoruba in Ibadan, Nigeria), CEU (Utah Residents with Northern and Western European ancestry), TSI

(Toscani in Italia), CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan) (Table 2.1).

PLINK [120] was used in the retrieval of the relevant 1000 Genomes data from the cream-ce server in the high performance cluster at the University of the Witwatersrand, which housed files that were already converted from their native .vcf format to PLINK-ready files. PLINK is an open source whole genome association analysis toolset that efficiently computes a range of basic, large-scale analyses. It is particularly useful for its convenient handling and manipulation of large data sets like whole genome data which contains several thousands of markers genotyped for thousands of individuals. PLINK consists of five main functional domains: population stratification, identity-by-descent estimation, data management, association analysis and summary statistics. For the purposes of this research, the data management domain served the major needs of the analyses of the 1000 Genomes data set.

Table 2.1 1000 Genomes population interrogated in this study. AFR = Africa, ASN = Asia, and EUR = Europe.

Population Description	**Population Code**	**Super Population Code**	**Sample size**
Luhya in Webuye, Kenya	LWK	AFR	97
Yoruba in Ibadan, Nigeria	YRI	AFR	88
Han Chinese in Beijing, China	CHB	ASN	97
Japanese in Tokyo, Japan	JPT	ASN	89
Utah Residents with Northern and Western European ancestry	CEU	EUR	85
Toscani in Italia	TSI	EUR	98

2.2 Functional Analysis

To get some biologically meaningful insight into different gene combinations of the genes present in the both the Expression Genes list and the GWAS Genes list, they were interrogated with the Database for Annotation, Visualization and Integrated Discovery (DAVID) [121].

2.2.1 Overview of DAVID

DAVID is a bioinformatics resource kit that provides a comprehensive set of functional annotation tools which can be useful in the understanding of biological meanings of large gene lists derived from genomic studies [122]. This kit consists of a knowledgebase as well as other integrated, web-based functional annotation and analytical tool suites. These include text and pathway-mining tools such as gene name batch viewer, functional annotation chart, functional annotation clustering, and gene functional classification tool amongst others. The DAVID knowledgebase is a gene-centred database [121]. It is very comprehensive as it integrates a huge array of the major and well-known public bioinformatics resources centralized by the DAVID gene concept, a single-linkage method to amass millions of gene/protein identifiers from several public genomic resources into DAVID gene clusters [121]. This allows for an improved cross-reference capability where more than 40 publicly accessible functional annotation sources can be comprehensively integrated and centralized by DAVID gene clusters.

The gene functional classification tool contributes towards a systematic enhancement of the biological interpretation of large gene lists as it groups genes based on their functional similarities. It generates a gene to gene similarity matrix from over 75,000 terms (including pathways like the KEGG, REACTOME, PANTHER, and BIOCARTA pathways) and 14 functional annotation sources [122] with the clustering algorithms classifying highly related genes into functionally related groups. The functional annotation tool primarily provides batch annotation and gene-Gene Ontology (gene-GO) term enrichment analysis thus highlighting the most relevant GO terms associated with a gene list.

The vast annotation content coverage - over 40 annotation categories, including GO terms, protein-protein interactions, protein functional domains, disease associations, bio-pathways, sequence general features, homologies, gene functional summaries, gene tissue expressions and literatures - as well as the flexible options to display results makes it a very powerful tool within this kit for investigators to analyse genes from different biological aspects in a single space. The results display options include the functional annotation chart report which provides an annotation term-

focused view of the genes being investigated. In this report, a one-tail Fisher's Exact statistic, denoted as an EASE score, is calculated and the results shown to be statistically significant have to had passed the default DAVID thresholds. Other optional statistics available in this chart view are Fold Enrichment values as well as Benjamini, Bonferroni and FDR corrected p-values. This tool was greatly utilized in this study for gene set enrichment and reduction purposes.

The GWAS Genes list and the Expression Genes list were separately analysed by DAVID with the same parameters for prioritised gene selection.

2.2.2 Criteria for Prioritized Gene Selection

The categories and options selected in DAVID for core gene selection are listed in Table 2.2.

Table 2.2 Options selected in DAVID for analysis.

DAVID Categories	Selected Options
Disease	OMIM Disease, Genetic Association DB-Disease
Gene_Ontology	GO_TERM BP_FAT
Pathways	BBID, BIOCARTA, KEGG, PANTHER

The Online Mendelian Inheritance in Man, OMIM, Disease database [123] is a freely available and daily updated collection of human genes and genetic phenotypes that contains over 12,000 genes and information on all identified Mendelian disorders. The Genetic Association DB-Disease is a collection of human genetic association studies of complex diseases and disorders, including data extracted from published papers on GWAS and candidate gene studies [124]. These options were selected in the disease category of the DAVID tool because of the vastness of the databases in their collection of genes and associated phenotypes for both monogenic and complex traits.

Gene ontology (GO) is a powerful method of interpreting and summarizing biological functions from a given list of genes [125]. However, redundant terms do occur in the enriched GO list that usually has little, if any, additional information to the analysis. The GO Fat database was developed as part of the Annotation Tool in DAVID to

combat the challenge of redundancies in the enriched GO lists. The Gene Ontology Biological Process, GO_TERM BP_FAT, [121, 122] is thus a subset of the classification of gene products in the hierarchy of the computationally amenable encyclopaedia of gene functions and their relationships, comprising more specific terms. The Biological Biochemical Image Database, BBID, [126] is a searchable database of images of possible biological pathways, gene families, macromolecular structures, and cellular relationships. BIOCARTA [127] is a biological pathways resource that provides representations of gene-to-gene interactions in an easy-to-read diagram. It is a constantly evolving resource as it is constantly updated by new research from the scientific community. The Kyoto Encyclopaedia of Genes and Genomes, KEGG, [128] is a database resource that allows for the understanding of high-level functions and importance of the biological system, from molecular-level information. This is especially beneficial for large-scale molecular datasets generated by high-throughput experimental technologies like genome sequencing. The Protein ANalysis THrough Evolutionary Relationships (PANTHER) database [129, 130] is a system of proteins and their genes that have been classified evolutionarily by related proteins, biological processes and pathways. These selected options for the Gene_Ontology and Pathways categories in DAVID allow for possible functional relatedness of genes as well as various levels of gene-gene interactions to be explored.

These databases were interrogated by DAVID in the search for functionality and enrichment of the genes contained in the two genes lists with the Functional Annotation Chart report thresholds left as DAVID's defaults: count = 2 (minimum number of genes in the input list for the corresponding term) and EASE = 0.1 (maximum P-value/EASE score).

A Bonferroni-adjusted p-value ≤ 0.05 in DAVID's functional annotation chart was set as the threshold for significance of annotation results. The Bonferroni correction method adjusts the data for false positives or type I errors that can result from multiple testing [131]. This method, although quite conservative, was chosen in this case because of its easy and wide application to different kinds of data sets. The genes associated with the terms and pathways that met this criterion for both the

Expression and GWAS Genes lists were combined to provide a core list of genes for downstream analysis that is simply referred to as the “Gene list”.

2.3 Population (Allele) Frequency Analysis

To get some understanding of the genes via their corresponding markers in a population context, the 1000 Genomes project [119] data were interrogated via PLINK.

2.3.1 T2D Risk Allele Frequency Distribution

The GWAS SNPs list was directly analysed to assess the distribution of identified T2D risk allele frequencies across the six 1000 Genomes populations of interest. This was done by identifying the presence or absence of a risk allele from the GWAS SNPs list in each individual in these populations. The frequency of occurrence of each allele in these populations was calculated based on the number of risk alleles in each individuals belonging to the different populations. A frequency chart was then plotted to visualise and compare the risk alleles’ distribution in the African, European and Asian populations. This was done using an in-house custom Perl script (EA 2) (personal communication, Dr. Ananyo Choudhury).

2.3.2 SNPs List Generation

An already converted PLINK-ready (binary files) recent version of the .vcf files of the 1000 Genomes data (Phase1, version 3, October 2012) was accessed on the local cream-ce server at Wits University. This version containing about 36.7 million autosomal SNPs and 1.38 million short structural variants (SSVs) [132] was the 1000 Genomes data source for this study.

The coordinates of the genes resulting from the enrichment analysis by DAVID were retrieved using the Ensembl Biomart [133]. Using these coordinates, a text file containing the ranges of each gene, that is, the start and end chromosomal position together with the corresponding chromosome number, a ‘range’ file was created. This file was used in PLINK to retrieve SNPs associated with the genes from the 1000 Genomes data. In retrieving the SNPs, the ‘--geno’ and ‘--mind’ commands were used to control for missingness in the data.

```
plink --bfile 1000-all --geno 0.025 --mind 0.025 --extract Ensembl_Genes_Ed.txt --range --recode
--out V_Rec_Final_1000-all
```

The '--geno' and '--mind' thresholds were both set to 0.025. This means that individuals with more than 2.5% of missing genotype data will be excluded and only SNPs with a greater than 97.5% genotyping rate will be included in the resulting output file. This subset of the 1000 Genomes data served as the primary 1000 Genomes data of relevance in this study and was subsequently analysed for population-specific variances. A random set of SNPs of almost similar size was also retrieved from the 1000 Genomes data using the '--thin' command in PLINK.

```
plink --bfile 1000-all --geno 0.025 --mind 0.025 --thin 0.0040 --make-bed --out RandomSample1k
```

The '--thin' command is a flag that allows for the retrieval of a specified percentage of SNPs from the input data based on random sampling. This random set of SNPs was retrieved to serve as control for the specifically selected gene-associated primary data used in this study and is simply referred to as the Random SNPs list.

2.3.3 Highly Relevant SNPs Selection

In order to identify population-specific variants, pairwise fixation indices (F_{ST}) and integrated Haplotype Scores (iHS) were calculated for the different populations.

- **FST Calculations**

F_{ST} is the basic statistical measurement of differentiation between populations originally proposed by Sewall Wright [134]. It is generally a measure of the reduction in heterozygosity in populations. It is also seen as the proportion of genetic diversity resulting from allele frequency differences among populations [135]. F_{ST} provides important insights into the evolutionary processes that influence the structure of genetic variation within and among populations and is a commonly used descriptive statistic in population and evolutionary genetics [135]. F_{ST} estimates can be used in the identification of regions of the genome that have been selection targets. This statistic was calculated using an in-house Perl script that closely mirrors Wright's [134] implementation of the estimator. This script was validated against the snpStats R package's implementation of F_{ST} [136].

Attribute text files of pairwise combinations of interest were created using the 'grep' function at the command line to retrieve the specified pairwise combinations from an existing master attribute file on the local cream-ce server at Wits University. This

master attribute file accompanied the download of the original .vcf 1000 Genomes data. These files generally contain the individuals' identification and the populations to which they belong. Pairwise F_{ST} calculations were done for each of the African populations versus the populations from Europe and Asia. The resulting file was cleaned by removing the 'NAs' and negative values. F_{ST} values generally range from 0 to 1, where a zero implies that the populations being considered are freely interbreeding, that is complete panmixis [134]. A value of one indicates that the populations being considered do no share any genetic diversity and as a result any genetic variation observed can be attributed to the population structure. An F_{ST} cut-off for "very great" genetic differentiation was set at 0.25, mirroring Wright's suggested degree of divergence thresholds shown in table 2.3 [137]. This cut-off was applied to the primary dataset as well as to the randomly selected SNPs.

Table 2.3 Wright's suggestions for interpreting genetic differentiation.

F_{ST} values	**Wright's Guidelines**
0 – 0.05	Little differentiation
0.05 – 0.15	Moderate differentiation
0.15 – 0.25	Great differentiation
> 0.25	Very great differentiation

The SNPs were to be annotated to their corresponding genes to allow for the combined analysis of these SNPs from a gene perspective. This was done in R using the merge function. The merge function is part of the base package in R and allows for the joining of two data frames by a row or column common to both data sets. To accomplish this, the previously retrieved set file was reformatted to contain two columns – gene and SNP to allow for a relatively straightforward merging process. This file was loaded into R and merged with the F_{ST} pairwise result via the SNP column common to both files. The SNP count per gene was done for both the significant set and the total set of SNPs using the summary function in R.

For the SNPs meeting the 0.25 threshold, stochastic simulation of p-values as well as overrepresentation scores were computed. To calculate these values, the ratio of the number of significant SNPs per gene to the total number of SNPs in each gene was calculated. This gave the 'observed' value. The ratio of the number of significant

SNPs in the random gene set to the total number of retrieved random SNPs present in the populations compared was also calculated, giving the 'expected' ratio. Only SNPs with rsIDs were used in these calculations. The p-value was calculated as a hypergeometric distribution of these 4 values – significant SNPs per gene, total number of SNPs in each gene, significant SNPs in random gene set and total number of retrieved random SNPs present in population comparison - in Microsoft's Excel 2010. The HYPGEOMDIST function in Excel was used to perform this analysis. It calculates the probability of a given number of successes from a sample of a population given the 4 aforementioned parameters. The successes in this case refer to the number of SNPs that meet the initial cut-off of an F_{ST} value of 0.25. The overrepresentation score was calculated as the ratio of the observed value to the expected value for each gene. A gene was thus considered significantly diverged, in this study, if it had an F_{ST} value ≥ 0.25, a p-value ≤ 0.05 and an overrepresentation score of at least 2. This was done for each pairwise combination. This empirically defined cut-off was employed to allow for the selection of potentially robust F_{ST} values to control for possible errors that might be introduced as a consequence of the polymorphism of the dataset.

A combination of the pairwise SNP results by continents prior to the stochastic p-value and overrepresentation score calculation was done to produce one merged file per continent. The significantly diverged SNPs based on an F_{ST} ≥ 0.25 in the CEU_LWK, CEU_YRI, TSI_LWK and TSI_YRI pairwise analyses were combined into a non-redundant set to give the African versus European (AFR_EUR) comparison. Similarly, the combination of significant SNPs in the CHB_LWK, CHB_YRI, JPT_LWK and JPT_YRI pairwise analyses produced the African versus Asian (AFR_ASN) comparison. P-values and overrepresentation scores were then computed with these combined files. The intersecting SNPs from the AFR_EUR and AFR_ASN sets, also prior to overrepresentation scores and p-value calculations, make up the African versus non-African (AFR_nonAFR) SNPs list that was subsequently annotated to their corresponding genes using the merge function as explained above, to give the AFR_nonAFR gene list. Similar p-value and overrepresentation score calculations (as previously described) were done on this list to give the final F_{ST} AFR-nonAFR gene list.

- **Integrated Haplotype Score (iHS) Calculations**

The main purpose of these calculations was to identify possible signatures of selection in the African population. The iHS is a statistic that detects evidence of recent positive selection at a locus based on the differential levels of linkage disequilibrium surrounding a positively selected allele compared to the background allele at the same position [138]. It is an extended haplotype homozygosity (EHH)-based analysis approach that is widely used for detecting recent and strong natural selection. The basic idea driving such an analysis is that a haplotype with high frequency and high homozygosity that extends over a considerably long stretch of genome often corresponds to regions under an incomplete selective sweep. One of the landmark examples of such a selection was detected in the lactase region (LCT), where selection on lactase-persistence into adulthood is a trait that has been subject to a selective sweep in European (and some African populations) which has not completely fixed [139]. This method relies on the linkage-disequilibrium structure of local regions of the genome for identifying tracts of homozygosity within a 'core' haplotype, using the EHH as a statistic. The EHH is summed over all sites away from a core SNP, and compared between the haplotypes that carry the ancestral and the derived alleles in the SNP. The iHS therefore gives a numerical indication of the amount of extended haplotype homozygosity along the ancestral allele at a given SNP with respect to the derived allele. This statistic is normalized to have a mean of 0 and variance of 1 and standardized empirically to the distribution of observed scores over a range of SNPs with similar derived allele frequencies to make them comparable to each other [138].

A ‘set’ file was created in PLINK from the list of gene-associated SNPs generated from the 1000 Genomes data. A set file in this instance sorts, organizes and lists the SNPs according to their associated genes. This action, referred to as sub-setting in PLINK, was accomplished using the ’--write-set’ flag together with the previously described text file of ranges containing chromosome number and base positions of genes.

```
plink --file V_Rec_Final_1000-all --make-set Ensembl_Genes_Ed.txt --write-set --out
V_Rec_Final_Set_1000-all
```

The ‘iHS_calc’ script from the WHAMM package [140] was used to carry out the iHS calculations for each of the 6 populations. WHAMM is an open source analysis package that, amongst other functions, estimates patterns of homozygosity in whole

genome data sets like the 1000 Genomes [141] and HapMap [142–144]. For iHS calculations to be done using this script, the physical positions of the SNPs needed to be specified. This information can either be obtained from the LD architecture of a dataset (like the 1000 Genomes data) using tools like LDHat [145] or incorporated from existing physical maps which have been identified using thousands of individuals from different populations.

As the aim of this study was to identify the relative strength of signatures of selection in certain genomic regions compared to others rather than identifying novel signatures, physical position-based maps were chosen over LD-based maps. The Rutgers' combined linkage physical map for human genome (build GrCh37) was thus downloaded [146] and used to incorporate the required physical positions into the 1000 Genomes data. The Rutgers combined linkage physical map is a high-resolution genetic map that comprises the largest set of polymorphic markers with publicly-available genotype data. This map is well-suited as a comprehensive resource for determining genetic map information as the position of most of the included markers are corroborated by both recombination-based data and genetic sequence [146]. It incorporates SNPs as well as sequence-based positional information. However, an already incorporated file from [141] was utilized. EHH was thus integrated with respect to genetic distance in cM. The background distribution for each population was estimated by randomly sampling of 10,000 50-SNP blocks and calculating the iHS for the SNPs occurring in these blocks. The iHS scores were subsequently standardized based on the allele frequency bins derived from the background. The ancestral allele assignment was done according to the ancestral state information provided by the 1000 Genomes Consortium, which was based on a 4-way EPO alignment of human, chimpanzee, orangutan and rhesus macaque [119]. An initial cut-off for significance of SNPs was set at an absolute iHS value of 2.
iHS analysis was done on a random sample of 100,000 SNPs for each of the populations to calculate an expected ratio to serve as control for the study data iHS results.

As was done with the F_{ST} values, a stochastic simulation of p-values as well as overrepresentation scores based on observed versus expected ratios were computed for each gene. A gene was considered selected if it had a $p \leq 0.05$ and an

overrepresentation score of at least 2. Since selection is being investigated primarily in the African populations, emphasis is placed on the LWK and YRI populations.

2.3.4 Africa-identified vs. Non-Africa-identified SNPs and Genes List

The criteria for SNPs and their corresponding genes to be considered significant in this study, with the overarching aim of identifying African-specific variants associated with T2D risk and/or pathogenesis, was for the SNP to show differentiation (based on set study thresholds in the F_{ST} analysis) between African and non-African populations and selected by virtue of its presence in the iHS final gene list in the African populations.

A PhenoGram plot [147] was used to present a graphical summary of the resulting SNP numbers with different colours representing the SNPs meeting the different cut-offs and thresholds of significance at various stages of the analyses.

3 CHAPTER THREE – RESULTS

This study is divided into 3 sections – public data mining, functional analyses of the mined data and population frequency analyses to put the study in a population-specific context.

3.1 Public Data Mining

3.1.1 T2D-associated Studies

- **GWAS Data**

The GWAS catalogue was initially accessed on 1 June, 2012 and updated on 2 September, 2013. At this point, there were a total of 51 T2D-associated studies (Table 3.1) matching the set search criteria. The matching studies were published from the year 2007 to 2013. 218 SNPs were reported to be very strongly associated, at $p < 10^{-5}$, with type II diabetes according to these 51 studies. This makes up the GWAS SNPs list. The genes linked to these SNPs, as reported by the original studies, are 175. These genes make up what will be referred to as the GWAS Genes list. The difference between the number of SNPs and genes can be accounted for by the presence of multiple SNPs from the same gene being identified as strongly T2D-associated in different studies which would reduce the gene list, as well as SNPs occurring in intergenic regions. The participants in the selected studies' data sets comprised a wide range of ethnicities. These include individuals of European descent, Hispanic ancestry, African-Americans, Indian-Asians, Mexican-Americans, Japanese, Chinese, Malaysians, Filipinos and Koreans. The studies were mainly carried out on Affymetrix and Illumina platforms with sample sizes ranging from 187 to meta analyses including over 46,000 individuals.

Table 3.1 51 GWA Studies meeting the set search criteria

Date	Study
2013	Genome-wide association study in a Chinese population identifies a susceptibility locus for type 2 diabetes at 7q32 near PAX4 [148].
2013	Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India [149].
2012	A single-nucleotide polymorphism in ANK1 is associated with susceptibility to type 2

	diabetes in Japanese populations [150].
2012	A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance [61].
2012	1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data [151].
2012	A genome-wide association study identifies GRK5 and RASGRP1 as type 2 diabetes loci in Chinese Hans [152].
2012	Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases [153].
2012	Fasting glucose GWAS candidate region analysis across ethnic groups in the Multi-ethnic Study of Atherosclerosis (MESA). [154]
2012	Genome-wide association study identifies novel loci association with fasting insulin and insulin resistance in African Americans [155].
2012	A genome-wide association search for type 2 diabetes genes in African Americans [156].
2012	Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21[157].
2011	Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes [158].
2011	Genome-wide detection of allele specific copy number variation associated with insulin resistance in African Americans from the HyperGEN study [159].
2011	Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci [160].
2011	Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia [161].
2011	Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas [162].
2011	A genome-wide association study confirms previously reported loci for type 2 diabetes in Han Chinese [163].
2011	Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians [164].
2011	Genome-wide association and meta-analysis in populations from Starr County, Texas, and Mexico City identify type 2 diabetes susceptibility loci and enrichment for expression quantitative trait loci in top signals [165].
2011	Use of diverse electronic medical record systems to identify genetic risk for type 2

	diabetes within a genome-wide association study [166].
2010	Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis [167].
2010	A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B [168].
2010	Identification of new genetic risk variants for type 2 diabetes [169].
2010	New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk [170].
2010	Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes [171].
2010	A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese [172].
2009	Confirmation of multiple risk Loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population [173].
2009	Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia [174].
2009	Common genetic variation near melatonin receptor MTNR1B contributes to raised plasma glucose and increased risk of type 2 diabetes among Indian Asians and European Caucasians [175].
2009	A genome-wide association scan for acute insulin response to glucose in Hispanic-Americans: the Insulin Resistance Atherosclerosis Family Study (IRAS FS) [176].
2009	Candidate loci for insulin sensitivity and disposition index from a genome-wide association analysis of Hispanic participants in the Insulin Resistance Atherosclerosis (IRAS) Family Study [177].
2008	SNPs in *KCNQ1* are associated with susceptibility to type 2 diabetes in East Asian and European populations [98].
2008	Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data [178].
2008	A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels [179].
2008	Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels [180].
2008	Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes [181].
2008	Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus [99].
2008	Variants in MTNR1B influence fasting glucose levels [182].
2008	A variant near MTNR1B is associated with increased fasting plasma glucose levels

	and type 2 diabetes risk [183].
2007	A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes [68].
2007	Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels [184].
2007	Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls [185].
2007	A genome-wide association study identifies novel risk loci for type 2 diabetes [186].
2007	A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants [60].
2007	Genome-wide association with diabetes-related traits in the Framingham Heart Study [187].
2007	Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.[188]
2007	Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies [189].
2007	A search for variants associated with young-onset type 2 diabetes in American Indians in a 100K genotyping array [190].
2007	Identification of novel candidate genes for type 2 diabetes from a genome-wide association scan in the Old Order Amish: evidence for replication from diabetes-related quantitative traits and from independent populations [191].
2007	A 100K genome-wide association scan for diabetes and related traits in the Framingham Heart Study: replication and integration with other genome-wide datasets [192].
2007	Type 2 diabetes whole-genome association study in four populations: the DiaGen consortium [193].

To summarise, the genes from the “reported” genes column in the main GWAS catalog document were selected to provide what is referred to as the “GWAS Genes list” and the “GWAS SNPs list” resulted from the “SNPs” column of the same spread sheet. For clarity in the downstream analyses, these lists will be referred to accordingly. These genes and SNPs can be found in Appendix I.

- **Gene Expression Studies Data**

- **Data Retrieval and Analyses**

The NCBI's GEO repository was accessed in May, 2012. The number of studies meeting the stringent search criteria, on Affymetrix platforms, at this point was 10. The studies were published from the years 2003 – 2011 and comprised 5 studies with samples from the skeletal muscle, 2 from the liver, 1 from blood tissue and 2 from the pancreatic tissue. The sample donors included Asian, European, European-American, Mexican American and African-American individuals. These studies were carried out on Affymetrix platforms Hu6800, Hu133A, Hu133B, Hu95A, Hu95Av2, Hu133_X3P and Hu133Plus2. The relatively large number of studies with skeletal muscles probably stems from the notion that insulin resistance in the skeletal muscle is the earliest detectable abnormality in individuals with a high risk for T2D [194] and the fact that it is an easier tissue to access than the liver or pancreas. The microarray expression studies are listed in Table 3.2.

Table 3.2 Gene expression studies meeting the set search criteria.

Study #	GSE #	Study Title	Year	Tissue
1	21340	Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1 [194].	2003	Skeletal Muscle
2	9006	Gene expression in peripheral blood mononuclear cells from children with diabetes [195].	2007	Blood
3	22309	The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle [196].	2007	Skeletal Muscle (Vastus Lateralis)
4	12643	Transcriptional profiling of myotubes from patients with T2D: no evidence for a primary defect in oxidative phosphorylation genes [197].	2008	Skeletal Muscle (Myotubes)
5	15653	Thyroid hormone-related regulation of gene expression in human fatty liver [198].	2009	Liver
6	20966	Gene expression profiles of Beta-cell enriched tissue obtained by laser capture micro dissection from subjects with T2D [199].	2010	Pancreatic Tissue (b-cells)
7	18732	Integration of microRNA changes in vivo identifies novel molecular features of muscle insulin	2010	Skeletal Muscle (Vastus Lateralis)

		resistance in T2D [200].		
8	23343	A liver-derived secretory protein, selenoprotein P, causes insulin resistance [201].	2010	Liver (Hepatic Tissue)
9	25724	Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic beta cells [202].	2011	Pancreatic Tissue (Islets)
10	25462	Increased SRF transcriptional activity in human and mouse skeletal muscle is a signature of insulin resistance [203].	2011	Skeletal Muscle

Study 1, carried out on the Affymetrix HG-U6800 platform, investigated gene expression in skeletal muscle from metabolically characterized non-diabetic individuals, regardless of family history of T2D, and type 2 diabetic Mexican-American individuals. The study showed that the reduced expression of a number of nuclear respiratory factor-1 (NRF-1)-dependent genes that encode some of the key enzymes in mitochondrial function and oxidative metabolism play noticeable roles in T2D and insulin resistance. Although the down-regulation of NRF-1 was only observed in the diabetic cases, the expression of PPAR gamma co-activator 1-alpha and-beta (PGC1-alpha/*PPARG*C1 and PGC1-beta/PERC), co-activators of NRF-1 and PPAR gamma-dependent transcription, was decreased in both diabetic individuals and non-diabetic controls with a family history of T2D [194]. These genes, however, did not meet the significance threshold set for differential expression consideration for this study.

There were 20 raw intensity (.cel) files available for the limma analysis for study 1. The contrast matrix created compared diabetic samples to control samples with no family history. This required the analysis of only 11 of the available 20 files – T2D (5) and control (6). A similar comparison in the original study resulted in 187 genes being identified as differentially expressed with no gene showing significant differential expression at an unspecified BH-adjusted p-value threshold. However, 47 genes met the BH-adjusted $p \leq 0.05$ initial significance threshold for this study. Some of the high ranking genes based on the adjusted p-values are ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide, *ATP5B*, oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide), *OGDH* and heat shock 70kDa protein 1A *HSPA1A*.

Study 2 was carried out on two platforms. For the purposes of this work, Study 2A will refer to the analysis done on the HG-U133A platform and Study 2B will refer to the analysis done on the HG-U133B platform. The original study aimed to show that the changes in gene expression in peripheral blood mononuclear cells as a result of counter-regulatory responses to immune dysregulation, insulin deficiency, hyperglycemia as well as the dysregulation of adaptive and innate immunity, accompany type I diabetes (T1D). However, samples were retrieved from individuals with T2D, and pathways and genes common to both T1D and T2D were also investigated hence the availability of T2D samples.

There were 234 raw intensity (.cel) files available for limma analysis for study 2, with 117 raw intensity files used per platform. The original study included samples from type 1 diabetic individuals at two time points after diagnosis and at diagnosis, type 2 diabetic individuals at diagnosis and non-diabetic individuals as controls. The contrast matrix created for this study compared the samples from type 2 diabetic individuals directly to those of the control group. No T1D samples were analysed. 35 of each set of 117 files were, therefore, used for analysis – T2D (11) and control (24). As the main purpose of the original study was not primarily related to this comparison, its result was not explicitly stated in the paper. However, 18 of the 22 most highly differentially expressed genes in T1D were noted to be similarly differentially expressed inT2D at a False Discovery Rate (FDR) of 0.01 [195]. At the set threshold for initial significance in this study, 1,083 and 377 genes were found to be differentially expressed in study 2A and 2B respectively. Some of these genes are insulin-degrading enzyme, IDE, insulin-like growth factor binding protein 3, IGFBP3, and potassium inwardly-rectifying channel, subfamily J, member 1, *KCNJ1*.

Study 3, carried out on the Affymetrix HG-U95A platform, aimed to study the effects of insulin on gene expression in the skeletal muscle and provide insight into the underlying defects causing insulin resistance and the molecular basis of insulin action in skeletal muscle [196].

One hundred and ten (110) raw intensity (.cel) files were used for the limma analysis of this study. The original study probed both treated and untreated insulin-resistant

and insulin-sensitive samples as well as treated and untreated diabetic samples. The contrast matrix created for this study compared the two extremes – untreated diabetic samples and untreated insulin-sensitive samples. As a result, 35 of the 110 files were used for analysis – T2D (15) and control (20). 4,410 genes from this comparison met the initial differential expression significance cut-off. Some of these genes are insulin-like growth factor 1 receptor, IGFR1, nuclear receptor subfamily 4, group A, member 1, NR4A1, and serum response factor, SRF.

Study 4 aimed to investigate the possible co-existence of reduced mitochondrial biogenesis with impaired insulin responsiveness in the early stages of the pathogenesis of T2D as it has been shown to co-exist with insulin resistance in both high-risk and type 2 diabetic individuals. This study, done on the Affymetrix HG-U95Av2 platform, was accomplished by comparing samples from the myotubes (skeletal muscle) of obese diabetic individuals with matched obese healthy patients . No gene was found to be significantly differentially expressed in the original study after correction for multiple testing was done using thresholds of either a false discovery rate (FDR) < 0.01 or a family-wise error rate (FWER) < 0.05.

Twenty (20) raw intensity (.cel) files were available for the limma analysis of study 4. The original study involved a direct comparison between samples from diabetic individuals and a control group. The contrast matrix created for this study compared likewise. The result of the limma analysis done for the samples in this study correspond to the original study analysis as no genes were found to be differentially expressed at a BH-adjusted $p \leq 0.05$.

Study 5 aimed to identify novel transcriptional changes in the human liver that could possibly contribute to hepatic lipid accumulation, the associated insulin resistance and its complication, T2D, as well as non-alcoholic steatohepatitis [198]. In the original study, liver biopsies were taken from obese individuals, with and without T2D as well as from lean non-diabetic individuals. This study was done on the Affymetrix HG-U133A platform. For the purposes of this work, only the samples from the obese diabetic individuals were assigned to the case group while samples from the lean study participants were regarded as control in the contrast matrix. One of the analyses done in the original study involved a direct case-control comparison as

such and 91 genes were found to be differentially expressed at Q < 0.001 using an unspecified Bioconductor package [198]. Q values generally measure the proportion of false positives when a test is considered significant [204]. The specific method of this correction implemented to give this Q value, was, however, unclear. The idea was to compare two extremes as specified in the contrast matrix. But this may not have been the best comparison for this study as obesity is confounding factor in T2D and obese controls might have been better suited.

Fourteen (14) of the available 18 raw intensity (.cel) files available for the limma analysis for this study – T2D (9) and control (5) - were analysed. Using the set threshold of BH-adjusted $p < 0.05$, 31 genes were found to be significantly differentially expressed. Some of these genes include small nuclear ribonucleoprotein polypeptide E, SNRPE, catenin (cadherin-associated protein), delta 1, CTNND1 and ATPase, Ca++ transporting, type 2C, member 1, ATP2C1.

Study 6, carried out on the Affymetrix HG-U133_X3P platform, aimed to gain some insight into the abnormal secretion of insulin that occurs in the pathogenesis of T2D by investigating changes in gene expression in pancreatic b-cells [199]. This was done by the laser capture dissection of frozen sections of pancreases of diabetic and non-diabetic human cadavers. This method of dissection was applied to overcome the possible limitations that may have resulted from the use of isolated islet preparations that may affect gene expression changes and the presence of non-beta cells, duct cells and acinar cells in the islets [205].

The 20 available raw intensity (.cel) files available for the limma analysis for this study were used – T2D (10) and control (10). At an unadjusted $p < 0.01$, the original study found 1,870 genes to be differentially expressed. This was not entirely replicated in this current work as only 750 genes were found to be differentially expressed at that p-value. At the set threshold of BH-adjusted $p < 0.05$, however, only 2 genes were significant - neurofilament, light polypeptide, *NEFL*, and MyoD family inhibitor domain containing, *MDFIC*.

Study 7 aimed to elucidate the mode of action and importance of microRNAs (miRNAs) in complex human diseases. In addition to the primary miRNA analyses,

microarray expression analysis was done on muscle biopsies from the vastus lateralis (skeletal muscle) of samples from individuals with impaired glucose tolerance (IGT), normal glucose tolerance (NGT) (control) and T2D [200]. This study utilized the Affymetrix HG-U133Plus2 platform.

Ninety-one (91) of the available 120 raw intensity (.cel) files were used for the limma analysis in this study – T2D (45) and control (46). The contrast matrix for this study directly compared the T2D samples to the control. A similar comparison in the original study with both the Significance Analysis of Microarrays, SAM [206] and limma yielded no differentially expressed genes. The significance cut-off used was not specified probably because of the focus on miRNA analysis. At a BH-adjusted $p \leq 0.05$ in this study, however, 3 genes were found to be differentially expressed. These genes are SET domain, bifurcated 2, *SETDB2*, HECT domain containing E3 ubiquitin protein ligase 1, *HECTD1*, and nucleoporin like 1, *NUPL1*.

Study 8, carried out on Affymetrix HG-U133Plus2 platform, aimed to identify hepatic secretory proteins involved in insulin resistance and specifically focused on the role of selenoprotein P, *SeP*, in the control of glucose metabolism and insulin sensitivity [201].

Seventeen (17) raw intensity (.cel) files were used for the limma analysis in this study - T2D (10) and control (7). The original study probed samples from individuals with T2D and normal glucose tolerance, regarded as the control group. The contrast matrix in this study was thus created accordingly. At the set BH-adjusted $p \leq 0.05$, cut-off for initial significance, no gene was differentially expressed.

The primary aim of study 9 was to demonstrate the role of class II phosphoinositide 3-kinase (*PI3K-C2α*) in the secretion of insulin in pancreatic b-cells. The original study also probed the possible alterations in the levels of PI3K-C2α in T2D, hence the availability of T2D samples [202]. The Affymetrix HG-U133A platform was utilized for this analysis.

Thirteen (13) raw intensity (.cel) files were the input for limma analysis - T2D (6) and control (7). The original study compared samples from type 2 diabetic individuals to samples from non-diabetic individuals. The contrast matrix for this study was created accordingly – T2D vs Control. This comparison yielded 4,355 genes being

differentially expressed at the initial cut-off for significance. Some of these genes are transcription factor 7-like 2 (T-cell specific, *HMG*-box), *TCF7L2*, *HNF1* homeobox B, *HNF1B*, and hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (tri-functional protein), alpha subunit, *HADHA*.

Study 10, carried out on the Affymetrix HG-U133Plus2 platform, aimed to identify transcriptional phenotypes associated with T2D by analysing the expression of genes in the quadriceps (skeletal muscle) in samples from individuals with T2D, non-T2D individuals with a parental history of T2D and samples from non-T2D individuals without any T2D history - the control group [203]. The study specifically demonstrated the role of serum response factor, *SRF*, activity in insulin resistance.

Twenty-five (25) of the 50 available raw intensity (.cel) files were used for the limma analysis in this study - T2D (10) and control (15). The contrast matrix created for this study, like in the previous instances, directly compared the T2D samples to the control. 4 genes were differentially expressed at a BH-adjusted $p \leq 0.05$. These genes are actin-binding Rho activating protein, *ABRA*, cysteine-rich, angiogenic inducer, 61, *CYR61*, nuclear receptor subfamily 4, group A, member 1, *NR4A1*, and kyphoscoliosis peptidase, *KY*. The number of differentially expressed genes was not stated in the original study paper, *ABRA* was noted as a top-ranking that has been previously identified as an activator of *SRF* transcriptional activity [203, 207].

- **Genes List**

Using a BH-adjusted $p < 0.05$ as the cut-off for significantly differentially expressed genes, 2 of these studies were eliminated giving a total of 7,887 genes from 8 studies (EA 4). This cut-off, although stringent, produced widely varying gene numbers from each study. As a result, an additional criterion was applied to the current list of differentially expressed genes to further increase the confidence in the list of genes used for downstream analyses. This requirement for genes to be significantly differentially expressed in more than one study, based on platform or tissue type, resulted in 497 genes. This list of genes is referred to as the Expression Genes List (Appendix II).

3.2 Functional Analysis

The GWAS genes list comprising 175 genes was uploaded to DAVID. 162 of these genes mapped to DAVID's IDs. This means that only 162 of the 175 genes corresponded to DAVID's internal annotation that facilitates a comprehensive navigation of gene-associated annotation across several databases. A Bonferroni-adjusted $p \leq 0.05$ was set as the threshold for significance in this enrichment step. 16 terms/pathways met this cut-off in DAVID's functional annotation chart and resulted in a list of 46 genes – genes overlap across the different categories. These terms and pathways can be seen in Table 3.3. Running these genes through DAVID's functional annotation clusters resulted in several levels of enrichment. Figure 3.1 shows a chart of gene ontology (GOTERM_BP_FAT) terms of these 46 genes with the top 3 enrichment scores clusters.

Table 3.3 Significantly enriched terms and pathways from the GWAS genes list.

Category	Term	Gene Count	Bonferroni
OMIM_DISEASE	Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes	21	2.45×10^{-34}
GENETIC_ASSOCIATION_DB_DISEASE	diabetes, type 2	28	3.23×10^{-12}
OMIM_DISEASE	Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes	9	4.70×10^{-12}
OMIM_DISEASE	Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels	9	4.70×10^{-12}
OMIM_DISEASE	A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants	9	4.70×10^{-12}
KEGG_PATHWAY	hsa04950:Maturity onset diabetes of the young	9	1.59×10^{-08}
GENETIC_ASSOCIATION_DB_DISEASE	diabetes, type 2 triglycerides	6	1.01×10^{-05}

GENETIC_ASSOCIATION_DB_DISEASE	diabetes, gestational	8	2.41×10^{-05}
GOTERM_BP_FAT	GO:0042593~glucose homeostasis	9	2.51×10^{-05}
GOTERM_BP_FAT	GO:0033500~carbohydrate homeostasis	9	2.51×10^{-05}
OMIM_DISEASE	Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome wide association data	5	4.98×10^{-05}
GOTERM_BP_FAT	GO:0031016~pancreas development	7	0.0012569
KEGG_PATHWAY	hsa04930:Type II diabetes mellitus	6	0.0167522
GOTERM_BP_FAT	GO:0007346~regulation of mitotic cell cycle	10	0.020368
GOTERM_BP_FAT	GO:0009743~response to carbohydrate stimulus	7	0.0315192
GENETIC_ASSOCIATION_DB_DISEASE	insulin	7	0.0327099

The 497 genes of the expression genes list were uploaded to DAVID. 485 of these genes mapped to DAVID's IDs. 9 terms/pathways met the set Bonferroni-adjusted p ≤ 0.05 cut-off, resulting in 96 genes. These terms and pathways can be seen in Table 3.4. Again, the genes overlapped across the terms so that a summation of the gene count column would not result in 96 because of redundancies. Running these genes through DAVID's functional annotation clusters resulted in different levels of enrichment. Figure 3.2 shows a chart of the functional annotation clusters (gene ontology terms) of these 96 genes with the top 2 enrichment scores clusters.

Table 3.4 Significantly enriched terms and pathways from the expression genes list.

Category	**Term**	**Gene Count**	**Bonferroni**
GOTERM_BP_FAT	GO:0006396~RNA processing	52	2.17×10^{-09}
GOTERM_BP_FAT	GO:0006397~mRNA processing	38	7.76×10^{-09}
GOTERM_BP_FAT	GO:0008380~RNA splicing	35	2.14×10^{-08}
GOTERM_BP_FAT	GO:0016071~mRNA metabolic process	39	1.30×10^{-07}

GOTERM_BP_FAT	GO:0000398~nuclear mRNA splicing, via spliceosome	22	2.02×10^{-05}
GOTERM_BP_FAT	GO:0000377~RNA splicing, via trans esterification reactions with bulged adenosine as nucleophile	22	2.02×10^{-05}
GOTERM_BP_FAT	GO:0000375~RNA splicing, via trans esterification reactions	22	2.02×10^{-05}
KEGG_PATHWAY	hsa03040:Spliceosome	18	2.13×10^{-04}
GOTERM_BP_FAT	GO:0010605~negative regulation of macromolecule metabolic process	46	0.015568

About 4 percent of the combined total number of genes (663 non-redundant genes) was excluded from the annotation process and subsequent analyses as a result of the absence of DAVID IDs. However, the combination of the resulting genes from this enrichment analysis with DAVID produced a non-redundant list of 140 genes with an overlap of 2 genes between the GWAS and expression results. The list of 140 genes is henceforth referred to as the Gene List which is the core gene set for downstream analysis in this study. This list can be found in Appendix III.

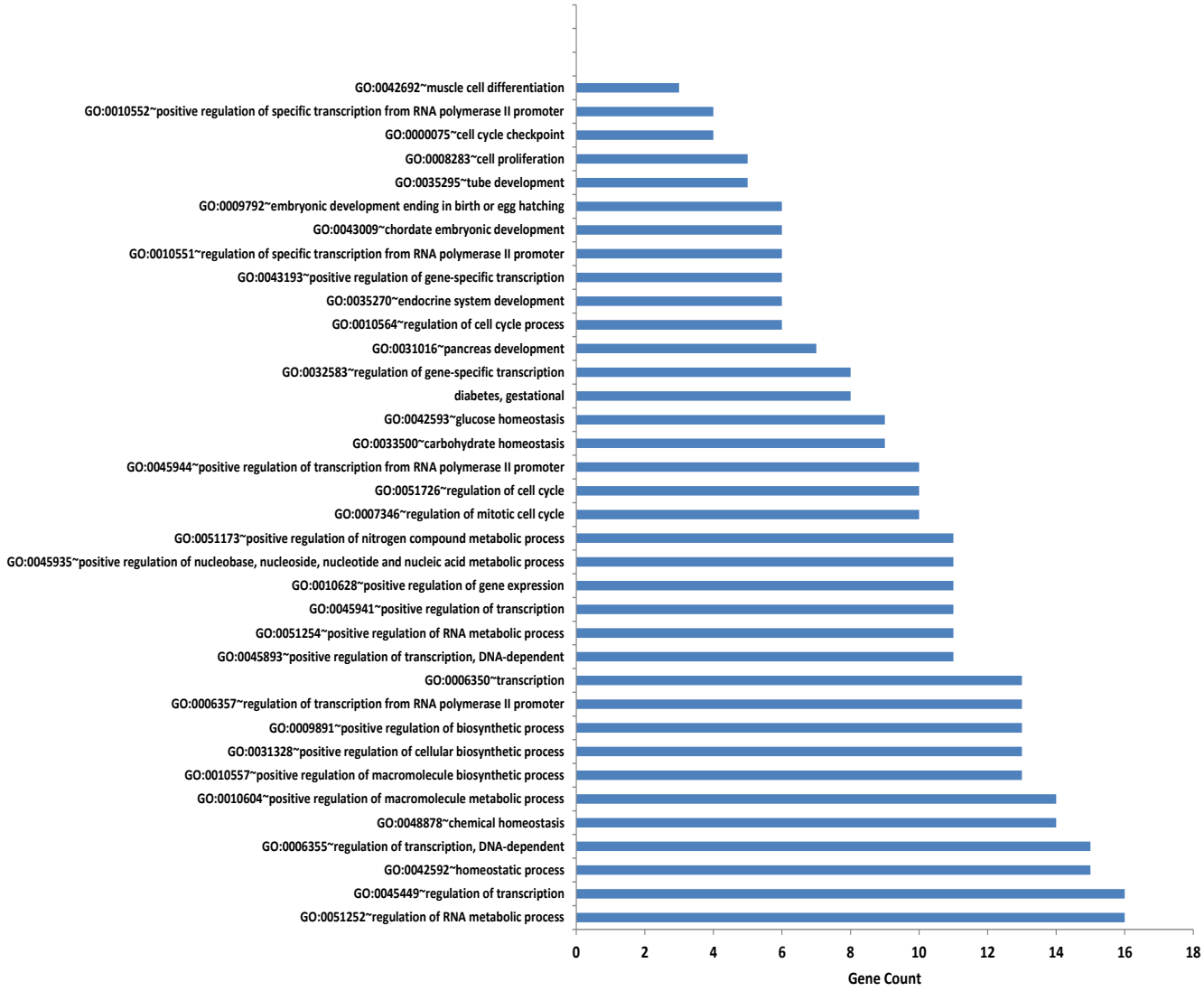


Figure 3.1 Functional annotation terms based on genes from the top 3 GWAS genes enriched clusters in DAVID.

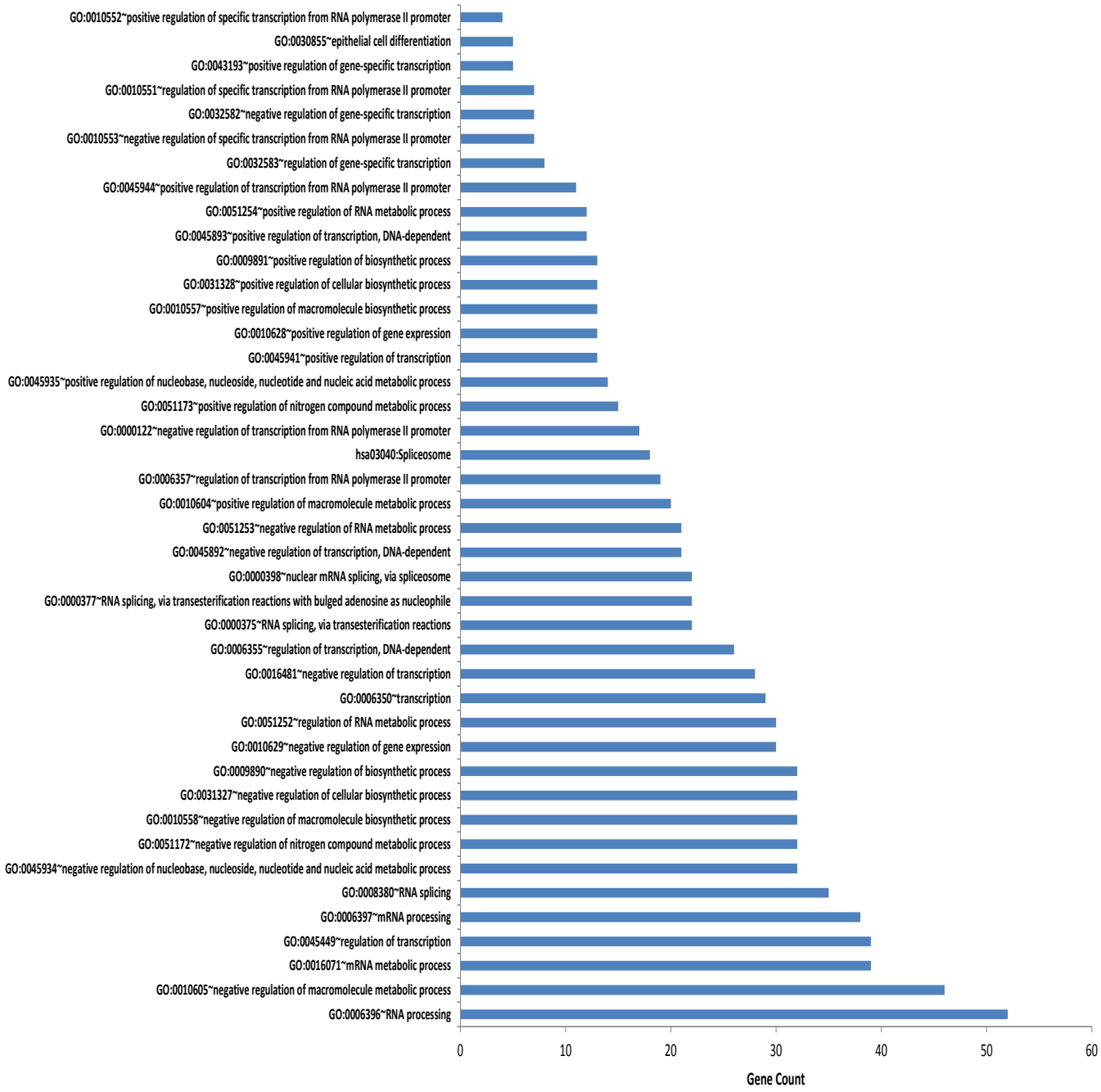


Figure 3.2 Functional annotation terms based on genes from the top 2 microarray expression genes enriched clusters in DAVID.

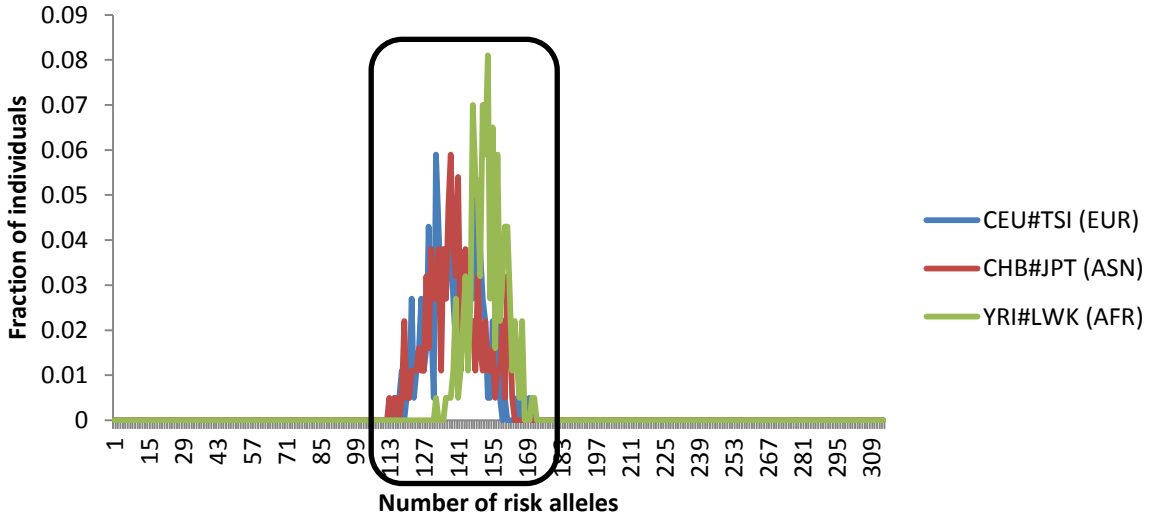
3.3 Population (Allele) Frequency Analysis

3.3.1 Risk Allele Frequency Analysis

Of the 218 SNPs identified in the GWAS studies, the risk alleles for 58 of them were not indicated. It is common practice to assign the risk allele as the minor allele in the

population being studied. Even though this is very often an accurate assumption, it is not always the case. I therefore decided to err on the side of caution and only include the SNPs where the risk allele was clearly stated and excluded those with missing risk alleles in the GWAS catalog. The risk allele frequency analysis was thus carried out with 160 SNPs. 157 of these SNPs were identified by the 1000 Genomes data meaning that the maximum number of risk alleles possible per individual ranged from 0 to 314 (157 SNPs X 2 alleles) as an individual could have either 0, 1 or 2 risk alleles at each SNP locus. This produced the resulting frequency distribution plot in Figure 3.3. The allele distribution observed here corroborates previously reported findings on 12 T2D risk alleles in a similar intercontinental frequency plot that the risk alleles for T2D are clearly present in higher frequencies in sub-Saharan Africa than in Asia and Europe [208]. A T-test was done to substantiate the observed risk allele differences where the null hypothesis was rejected for the Asian/African and European/African comparisons but not for the Asian/European comparison. This is shown in table 3.5.

A



B

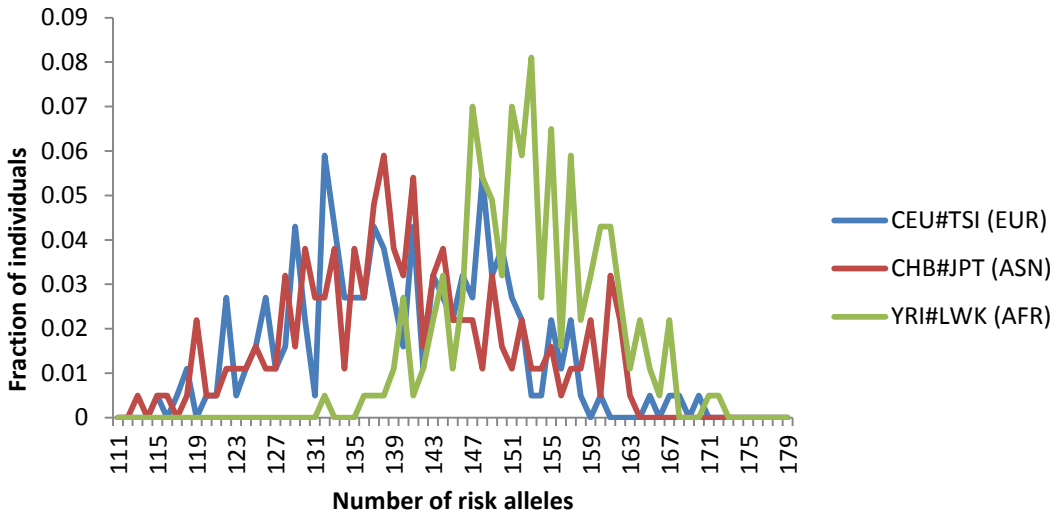


Figure 3.3 GWAS SNPs list risk allele frequency distribution. **A** shows the distribution of all 157 SNPs and **B** zooms in on the region (inset in **A**) of obvious intercontinental differences.

Table 3.5 Risk allele frequency distribution T-test results. Comparisons with p-values less than 0.0001 are in bold fonts.

Pairwise Comps 1	Avg. No. of Risk Allele Occurrence	Risk Alleles Used	Pairwise Comps 2	Avg. No. of Risk Allele Occurrence	T-Score	P-value	Null Hypothesis
CEU#TSI	140.22	157	**LWK#YRI**	153.28	13.56	**$<10^{-12}$**	Rejected
CHB#JPT	140.13	157	**LWK#YRI**	153.28	13.27	**1.62×10^{-11}**	Rejected
CEU#TSI	140.22	157	CHB#JPT	140.13	0.08	0.93832	Not rejected

3.3.2 SNP List Generation

The Ensembl database (release 73) was interrogated via its biomart package implementation, 'biomaRt' [209] in R for the retrieval of the co-ordinates of the genes contained in the gene list. The attributes retrieved, of the 140 genes resulting from the GWAS analysis and expression array studies following analysis in DAVID, include the chromosome name, start and end positions of the genes on the chromosome as well as the strand information. The result was saved as a tab-delimited text file which was subsequently converted to a space-delimited text file that was a preferred input data format for PLINK. The retrieved file showed 3 of the 140 genes to be located on sex chromosomes – 2 (*NONO* and *RBM3*) on the X

chromosome and 1 on the Y chromosome (*RBMY1A1*) which were excluded from further analyses.

The attributes needed to retrieve the SNPs associated with the genes list are the chromosome name and the gene's start and end position on the chromosome. The space delimited text file was thus adjusted accordingly to create a file with these positional ranges. This file was used to query the 1000 Genomes data on the local cream-ce server at Wits University. The missingness threshold of 0.025 that was set reduced the incorporation of incomplete genotype data in the resulting SNPs list.

The query produced a list of 154,144 autosomal SNPs that were used in downstream analysis. These SNPs served as the main 1000 Genomes data set for this study. Knowing the total number of SNPs contained in the entire 1000 Genomes data made it possible for an estimate to be made of the percentage of this total that would give a relatively comparable, albeit not the exact number, of random SNPs to serve as control for the selected gene list SNPs. A random set of SNPs was included in this study to serves as a control data set and eliminate any bias that might be introduced into the results by intra-data analysis. Hence, the results from the random sample combinations were utilized in the p-value calculations that were critical to the assessment of SNP/gene significance in this study.

The '--thin' command was set to randomly retrieve 0.40% of the total SNP count of the 1000 Genomes. This flag was used together with the same missingness parameters in the main data set. This resulted in a list of 151,713 random SNPs that is subsequently referred to as the Random SNPs list.

- **SNP List Analysis**

To facilitate the population analysis aspect of this study, a number of files needed to be created. A 'set' file containing the SNPs in each gene, ordered by base position, was created using the same file with chromosome position ranges previously used in the retrieval of SNPs corresponding to the 137 genes. This was accomplished with the '--make-set' and '--write-set' commands in PLINK. An 'attribute' text file, containing the identification for each individual together with the population group to which they belong, for each pairwise population combination, [MyTSI_YRI] for

example, was created from an existing master attribute file on the cream-ce server at Wits University. The attribute files were used to create population pairwise versions of the 1000 Genomes data that were used for downstream analysis. Table 2.1 shows the numbers of individuals present in each of the 1000 Genomes populations interrogated in this study.

- **F_{ST}**

The in-house Perl script utilized the population group data from the attribute files and the genotype information from the corresponding 1000 Genomes data to calculate pairwise F_{ST}s. Table 3.5 shows a summary of the F_{ST} results from the pairwise combinations – CEU_LWK, CEU_YRI, TSI_LWK, TSI_YRI, CHB_LWK, CHB_YRI, JPT_LWK and JPT_YRI – of the SNPs that meet the initial cut-off of 0.25 for significant genetic differentiation. A similar pattern of differentiation was also observed with the random SNPs list (Table 3.6 bottom table).

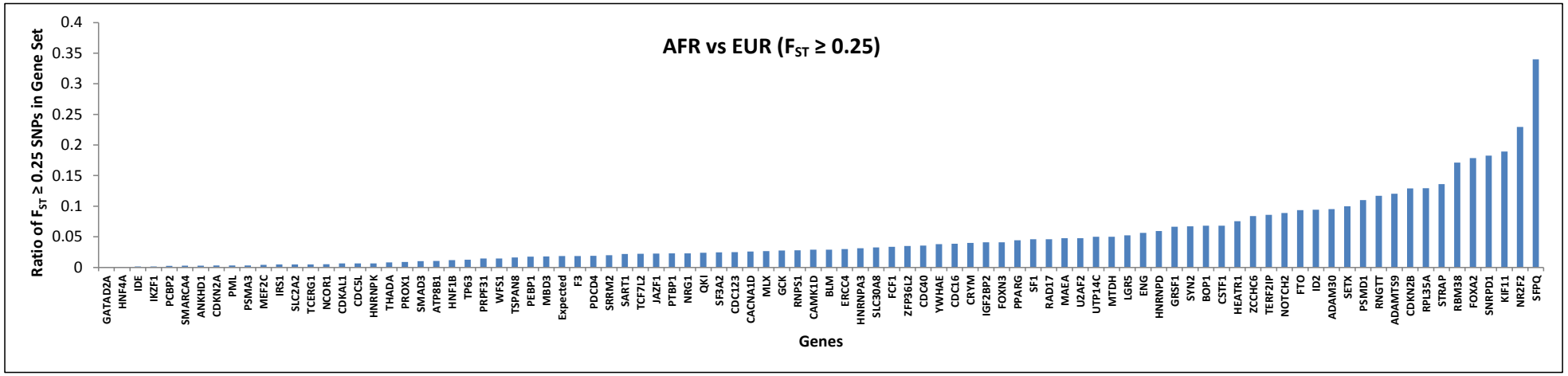
Table 3.6 Pairwise F_{ST} analysis results. The top table shows the results from the study data and the bottom table shows the random set results

Study Data	**CEU_LWK**	**CEU_YRI**	**TSI_LWK**	**TSI_YRI**	**CHB_LWK**	**CHB_YRI**	**JPT_LWK**	**JPT_YRI**
Total SNPs	154144	154144	154144	154144	154144	154144	154144	154144
F_{ST}≥0.25								
SNPs	1179	1541	1102	1367	2194	2524	2201	2581
Genes	77	83	73	76	83	83	90	91

Random Set	**CEU_LWK**	**CEU_YRI**	**TSI_LWK**	**TSI_YRI**	**CHB_LWK**	**CHB_YRI**	**JPT_LWK**	**JPT_YRI**
Total SNPs	151713	151713	151713	151713	151713	151713	151713	151713
F_{ST}≥0.25								
SNPs	1291	1577	1158	1444	1916	2172	1961	2228

Only SNPs with rsIDs were considered in SNP counts in subsequent analysis. The number of genes meeting the stochastic $p \leq 0.05$ and an overrepresentation score of at least 2, for each pairwise comparison are in brackets: CEU_LWK (13), CEU_YRI (9), TSI_LWK (10), TSI_YRI (6), CHB_LWK (7) and JPT_LWK (5). A combination of the African/European and the African/Asian pairwise SNP results produced the AFR_EUR (39) and AFR_ASN (45) comparisons respectively. 824 SNPs representing 63 genes were common to both intercontinental SNPs lists at the initial $F_{ST} \geq 0.25$ cut-off, giving the AFR_nonAFR SNPs list for that threshold. The final F_{ST} results for the AFR_nonAFR comparison, however, indicate that 7 genes from 228 SNPs are significantly differentiated between Africans and non-Africans. 2 (Notch homolog 2, *NOTCH2*, Kinesin Family member 11, *KIF11*) of these genes originate from the GWAS analysis, 4 (Nuclear Receptor Subfamily 2, *NR2F2*, Ribosomal Protein L35A, *RPL35A*, Small Nuclear Ribonucleoprotein D1 Polypeptide 16kDa, *SNRPD1*, and RNA Guanylyltransferase And 5'-Phosphatase, *RNGTT*) from the gene expression analysis and 1 (RNA binding motif protein 38, *RBM38*) from both.

Figure 3.4 shows genes with SNPs that met the $F_{ST} \geq 0.25$ cut-off. A list of the AFR_nonAFR comparison genes and their corresponding SNPs can be found in Appendix IV.



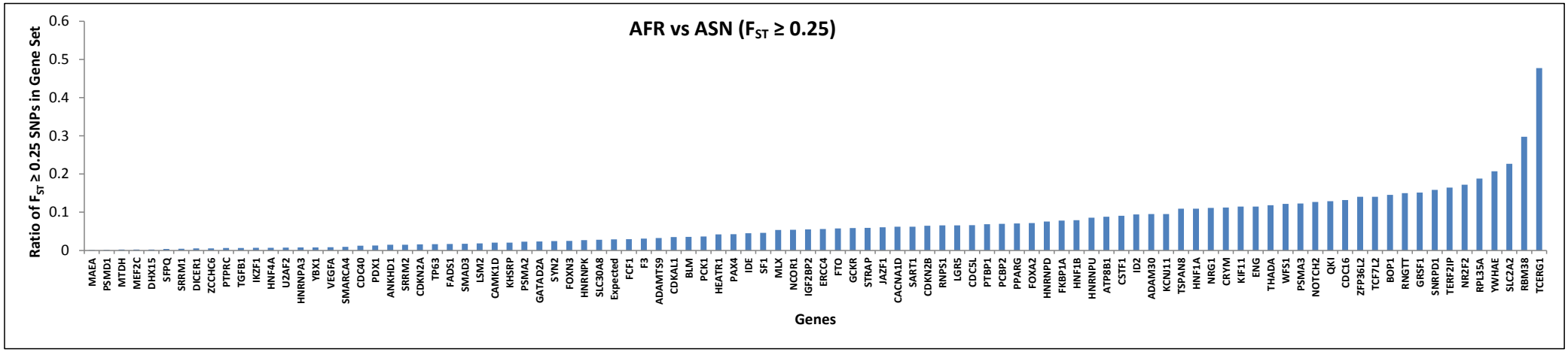


Figure 3.4 Intercontinental F_{ST} comparisons. The blue bars indicate the ratio of SNPs that met the set cut-off to the total number of SNPs present per gene.

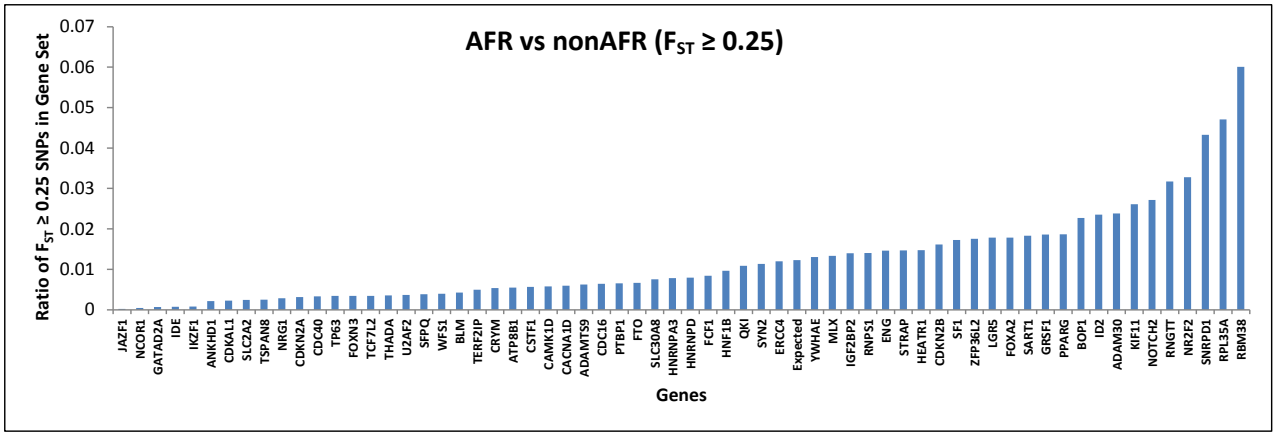


Figure 3.5 Combined intercontinental F_{ST} comparisons. The blue bars indicate the ratio of SNPs that met the set cut-off to the total number of SNPs present per gene.

- **Integrated Haplotype Score (iHS) Analysis**

The main purpose of the iHS analysis in this study is to identify possible signatures of selection in the African population. The results from the custom iHS_calc script from the WHAMM package usually comprise positive and negative scores. Absolute values of iHS were used as the designation of positive or negative scores is usually determined by the lengths of haplotypes on the ancestral and derived allele backgrounds relative to each other. A positive score means that the haplotypes on the ancestral allele background are longer compared to the derived allele while a negative score implies that the reverse is the case. Under this model, selected derived alleles are expected to contain excessive LD relative to the background. However, positive iHS scores greater than 2 are also considered as candidates for selection. This is because the ancestral allele may be hitchhiking along with the selected allele, or is itself a target for selection. Therefore selection could be in either direction. It is important to note that other measures for detecting signatures of selection could have been utilized. However, the choice of iHS for this study stemmed from the need to identify strong selection signals. Furthermore it is a frequently used tool in population genetic studies and is therefore regarded as a reliable instrument.

Of the 100,000 SNPs used in the random iHS run, the number of SNPs with |iHS| > 2 for each of the populations is indicated in brackets – CHB (5,594), JPT (5,112), TSI (6,164), CEU (6,164), LWK (6,140) and YRI (5,928).

The initial iHS result from the YRI population comprised 59,774 SNPs representing 137 genes. 1,895 SNPs representing 106 genes met the |iHS| > 2 cut-off. Of these genes, 2 appeared to be selected based on the calculated p-value and overrepresentation score thresholds. These genes are GRSF1 and *SNRPD1*. for the LWK population contained 59,774 SNPs representing 137 genes. The initial iHS result from the YRI population comprised 55,899 SNPs representing 133 genes. 1,681 SNPs (107 genes) met the |iHS| > 2 cut-off with only 2 of these genes meeting the set p-value and overrepresentation score thresholds. The genes are *SNRPD1* and ADAM30. A summary of the iHS results for the six study populations can be seen in Table 3.7.

Table 3.7 iHS analysis results for the six study populations.

	LWK	**YRI**	**CEU**	**TSI**	**CHB**	**JPT**
All SNPs (Genes)	59,774 (133)	55,899 (133)	36,282 (134)	55,899 (133)	34,173 (131)	34,690 (133)
SNPs with |iHS| > 2 (Genes)	1,895 (106)	1,681 (107)	892 (83)	1,812 (107)	2005 (111)	878 (85)
Selected Genes ($p < 0.05$ Overrepresentation score > 2)	2 (*GRSF1, SNRPD1*)	2 (*ADAM30, SNRPD1*)	1 (*SFPQ*)	1 (*SNRPD1*)	8 (*IRS1, NCOR1, BOP1, TCERG1, FOXA2, RPL11, CRYM, PDCD4*)	3 (*BOP1 FOXA2, FADS1*)

This analysis shows one T2D gene, *SNRPD1*, which was significantly differentiated from the non-African populations based on F_{ST} results to be selected in African populations even though it was also selected in one of the European populations, TSI. The SNPs contained in the LWK *SNRPD1* (15), however, are not exactly the same as those found in the YRI and TSI populations (18 SNPs). Nine *SNRPD1* SNPs overlap across these three populations, two (rs2959527 and rs34202260) of

which was also had high F_{ST} values. Six SNPs in this gene appear to be unique to the LWK population, 1 of which appears to be highly differentiated in AFR_non-AFR comparisons and is in high LD with three other selected LWK *SNRPD1* SNPs.

Table 3.8 shows a list of the iHS-selected *SNRPD1* SNPs in the YRI and LWK populations. The nine *SNRPD1* SNPs that appear to show significant population differentiation between Africans and non-Africans are rs2847117, rs2847139, rs2850556, rs2850558, rs2850568, rs2959525, rs2959527, rs3017641 and rs34202260.

The PhenoGram plot [147] in figure 3.6 below shows the distribution of the 824 SNPs representing 64 genes that showed significant genetic differentiation between the African and non-African populations based on the initial $F_{ST} \geq 0.25$ threshold. The colour codes indicate the SNPs that met the cut-offs of subsequent calculations. For example, red shows the SNPs that not only have F_{ST} values greater than 0.25 but also have a stochastic $p < 0.05$ and an overrepresentation score above 2.

Table 3.8 *SNRPD1* SNPs. The top and bottom tables show the list of SNPs selected in the YRI and LWK populations respectively. *SNRPD1* is located on chromosome 18. POS = Base Position, DAF = Derived Allele Frequency and std_iHS = Standardized Integrated Haplotype Score.

SNP	POS	DAF	std_iHS
rs138550124	19194968	0.0114	2.256
rs138737742	19195302	0.0114	2.171
rs142407732	19205213	0.0227	2.037
rs142827866	19199483	0.1023	2.082
rs182398649	19205277	0.108	2.96
rs186258230	19200019	0.017	2.229
rs187422483	19195514	0.0284	2.272
rs192860228	19202139	0.0511	2.494
rs193270098	19205527	0.0114	-3.618
rs2850560	19198301	0.1875	2.723
rs28675778	19201316	0.1193	2.759
rs2959527	19204607	0.1023	2.971

rs3017643	19205141	0.1818	2.003
rs73960450	19203507	0.1193	2.345
rs74546622	19197425	0.1136	2.122
rs78726174	19198070	0.1136	2.071
rs9947856	19195083	0.1193	2.143

SNP	**POS**	**DAF**	**std_iHS**
rs142827866	19199483	0.1392	2.492
rs149186410	19197804	0.0258	3.762
rs182398649	19205277	0.1701	2.775
rs185881584	19204969	0.0103	2.081
rs2847139	19198959	0.067	2.662
rs2850560	19198301	0.2165	2.271
rs28675778	19201316	0.0825	3.066
rs2959526	19204474	0.1701	2.829
rs2959527	19204607	0.1082	3.144
rs3017643	19205141	0.1701	2.812
rs34202260	19199672	0.1598	3.561
rs73960450	19203507	0.0825	2.743
rs9947856	19195083	0.0825	2.06
rs9950960	19201795	0.0412	2.038
rs9964889	19201845	0.0412	2.038

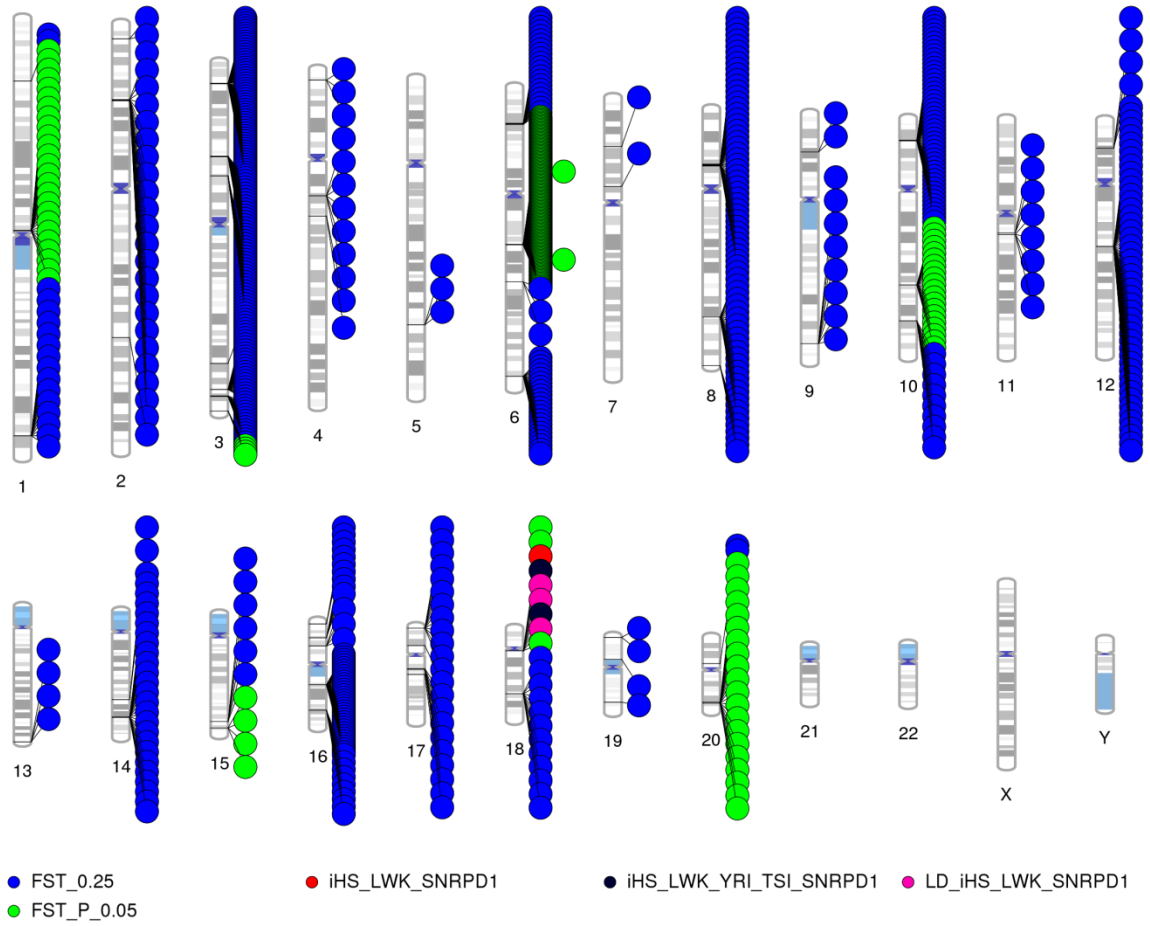


Figure 3.6 Genomic distribution of 824 SNPs representing 64 genes differentiated between African and non-Africans. All 824 SNPs met the $F_{ST} \geq 0.25$. Blue is the default colour of all the SNPs present. Light green circles indicate SNPs that also met the $p \leq 0.05$ and overrepresentation score > 2 threshold (228 SNPs representing 7 genes). The red circle shows the SNP that was selected only in the LWK population and happens to be in high LD with the 3 SNPs indicated by pink circles. The dark blue circle shows the two SNPs that appeared to be selected in the TSI, YRI and LWK populations.

From the different analyses to identify T2D genes of high significance in the African populations, *SNRPD1* would be the highest ranking because of its proposed selection (iHS) and population differentiation (F_{ST}) in this study. However, the six other highly differentiated genes (*NOTCH2*, *NR2F2*, *RBM38*, *RPL35A*, *SART1*, *and RNGTT*) very likely also give some insight into the pathogenesis of T2D in Africa. An updated research workflow showing the resulting gene numbers at each step is shown below.

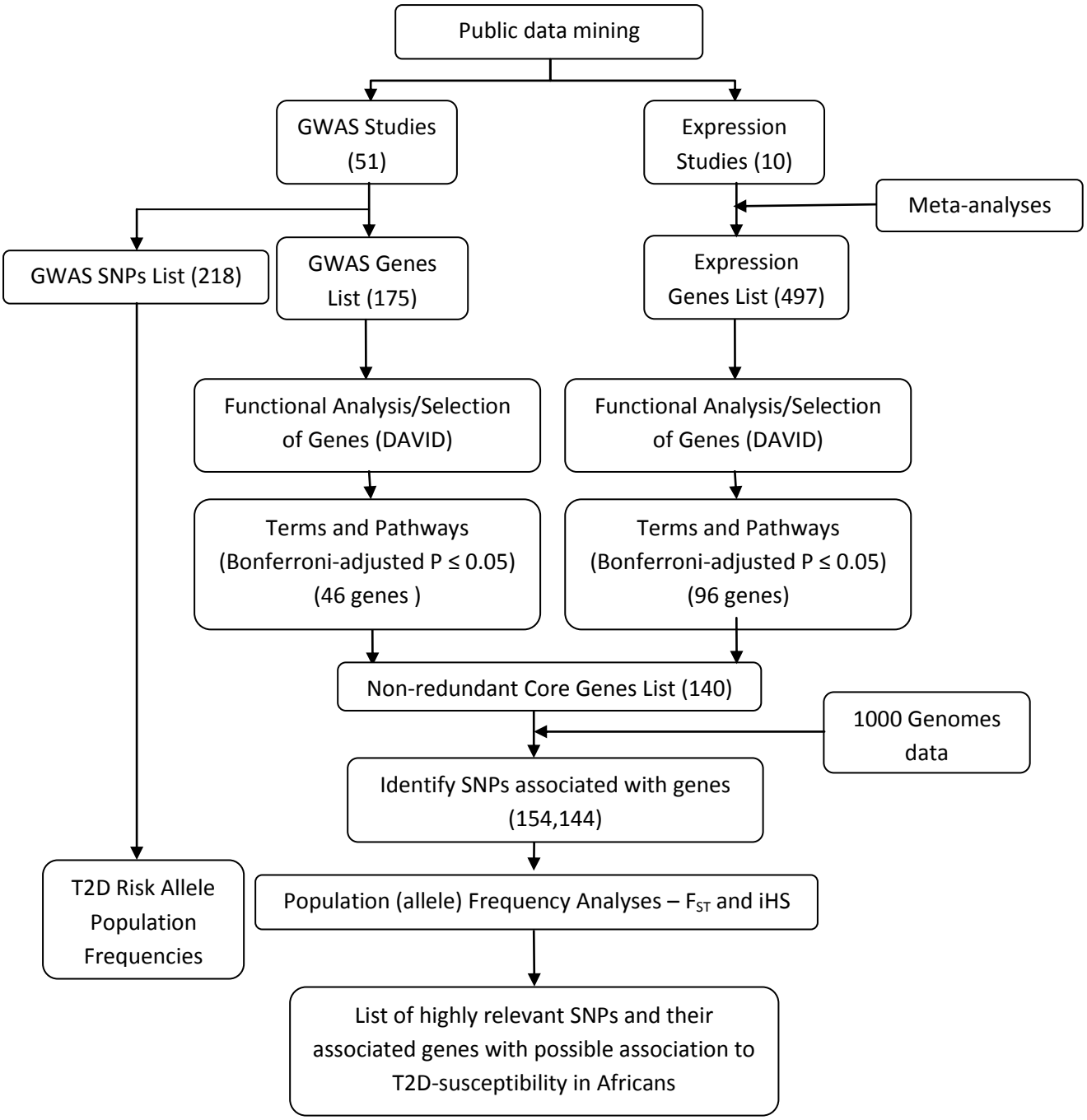


Figure 3.7 Research workflow with corresponding analyses of outcomes**.**

4 CHAPTER FOUR – DISCUSSION AND CONCLUSION

4.1 Discussion

The focus of this study is to gain insight into the pathogenesis of T2D from an African perspective by analysing existing non-African data in the context of African genetic diversity using the 1000 Genomes database.

4.1.1 Public Data Mining, Retrieval and Analyses

The retrieval of studies from the NHGRI's GWAS Catalog and NCBI's Gene Expression Omnibus were mainly to broaden the scope of the analysis to include previous efforts from both GWAS and non-GWAS sources. There are more T2D-related GWAS in databases than there are of T2D-related microarray expression studies as the search results in this study indicated – 51 GWAS compared to 10 microarray studies. This perhaps highlights the recent research trend in the genetics of complex diseases [2]. The total number of genes retrieved from both methods, however, is not comparable to the study numbers. This can, perhaps, be attributed to the compilation method used by the curators of the GWAS catalog. As stated in the methods section, the SNP information extracted from studies that are added to the catalog are those of SNPs with the strongest trait association [103]. This means that only a few SNPs and their corresponding genes from the entire body of SNPs in a study are included in the GWAS catalog, with one gene usually being associated with more than one SNP. This very likely limited the number of genes included in this analysis and the GWAS catalog report was used without further interrogation of the individual studies. The studies from the expression studies database, GEO, required further analyses and had no pre-set limitations on significance cut-off values. One of the objectives of this study was to create a relatively straightforward analysis pipeline that could utilize readily available data in curated databases without manual labour-intensive individual study data curation. It is also important to note that expression studies are gene-focused whereas GWAS are usually SNP-focused.

157 of the 218 T2D-related SNPs originated from the GWAS produced very interesting results in an assessment of risk allele frequencies across the African, Asian and European continents. It is possible that the retrieval and inclusion of the missing alleles absent from the GWAS catalog might have made the resulting

distribution more robust but the purpose of using only robust data would be defeated. Besides, the GWAS catalog is a manually curated database and there may have been reasons for such exclusions even when the original associated papers specified the risk alleles in question. However, purpose of the risk allele frequency calculation across the queried populations to assess the distribution of identified T2D risk allele frequencies across the six 1000 Genomes populations of interest was accomplished. A similar distribution was done in a 2012 PloS Genetics paper with only 12 risk alleles [208].

The analysis showed that the occurrence of previously identified T2D risk alleles, mostly identified in non-African populations, is highest in Africa, with little difference seen between Asia and Europe. This is consistent with the phenomenon that associates T2D risk alleles with human migration patterns - decrease in frequency as migration proceeds from Africa to Europe and Asia [208, 210]. This decrease in frequencies could perhaps be attributed to differences in agricultural development [211], historical events such as movement out of sub-Saharan Africa to areas with different climatic conditions [212] and even the thrifty genotype phenomenon [213]. The central premise of the thrifty gene hypothesis is that humans evolved from our hunter-gatherer predecessors that had to be efficient in food storage and utilization as they had to go through cycles of feast and famine. Natural selection appears to have favoured the ancient humans that had bodies that were better adapted to fuel storage or utilization as they were more likely to survive the famine portion of the cycle [214]. In other words, during famine, a predisposition to insulin resistance may have conferred protection or advantage to individuals by favouring glucose use in organs that operate through an insulin-independent mechanism like the brain, while reducing muscle utilization of glucose [208, 213]. Based on recent T2D prevalence statistics, this pattern of risk allele distribution is not exactly intuitive as one would expect non-African populations, who have a higher prevalence of T2D than African populations, to present higher frequencies of risk alleles since they have existed in a more resourced environment for a long period of time. However, environmental factors are very important in the aetiology of diabetes and the high prevalence of obesity in non-African populations in conjunction with gene-based predisposing factors may explain the higher prevalence in the developed world. The increasing availability of cheap, highly palatable, calorie-rich food combined with relatively

limited physical activity in recent times, appears to have created a situation where these previously advantageous thrifty genes now make us vulnerable to metabolic diseases like obesity and diabetes [213]. Thus, the incidence of T2D in Africa may increase dramatically as this obesogenic environment becomes more prominent within African populations harbouring high frequencies of diabetes-associated risk alleles.

This difference in risk allele occurrence and the current T2D incidence rates observed between the continents perhaps highlights the influence of the role of environmental differences which affect epigenetic risk factors in the occurrence of T2D as has been alluded to in previous studies [215].

The analyses of the expression studies data with limma produced largely varying results from no genes being significantly differentially expressed in some studies to over 4,000 genes identified in a study as being differentially expressed at a p-value corrected to account for multiple testing, a Benjamini-Hochberg $p \leq 0.05$. The tissue types, sample collection methods, platforms and even study-specific hypotheses differences very likely played a role in this widely varying significantly expressed gene numbers per study despite using the same basic analysis pipeline. It is probably then, not ideal to utilize pipelines for such secondary microarray data analysis despite the inclination to do so in this age of automation, as the questions of the original study are usually geared towards relatively different aims and objectives. For this study it was therefore necessary to generate a core list of genes – 497 genes – that took both the different tissues and Affymetrix platforms into account. Interestingly, only 8 of these genes were also on the GWAS gene list – *SRR, AP3B1, ARF5, CDC123, FOXN3, GRB10, IDE* and *RBM38*, with only 2 (*IDE* and *RBM38*) of these genes remaining in the post-DAVID analysis gene list. This represents about 2% and 4% of the expression and GWAS gene lists respectively, in both the pre- and post-DAVID analysis. The small number of overlapping genes between the two methods suggests that more genes were incorporated in the analysis as there were not many redundant genes. This is probably because GWAS identifies SNPs based on the strength of its association to the phenotype of interest whereas gene expression studies on the levels of differential expression of genes between cases (diabetics) and controls (non-diabetics). It is therefore believed that the combination of data from both methods provided a more comprehensive gene set for this study

than a single method would have. This combined approach has been previously demonstrated in a study that incorporated prostate cancer data sets from microarray array analysis and GWAS [216] where there was an overlap of over 80% of the genes identified with the two methods. An overlap of this nature would ideally serve as a validation of sorts for genes identified from the two experimental methods but the number of available study data in T2D have not quite reached the extent of some of the more extensively studied diseases like cancer.

4.1.2 Functional Analysis

Functionally analysing the GWAS gene lists produced 2 significant KEGG pathways – type 2 diabetes and maturity onset diabetes of the young. A GO term that stood out from this list was that of GO: 0031016~ pancreas development. Seven (7) of the GWAS genes were associated with this term. The pancreas plays a major role in the pathogenesis of T2D. The dysregulation of β-cell function in a situation where there is an increased insulin demand created by insulin resistance in peripheral tissues can attenuate the compensation mechanism to maintain insulin production levels and have consequences in the development and pathogenesis of T2D [2, 3].

One (1) KEGG pathway – spliceosome – was significant from the analysis of the expression gene list. The spliceosome pathway is involved in the assembly of the spliceosomal complex that facilitates the trans esterification reactions that lead to alternatively spliced mRNA transcripts [217].

A closer look at the resulting 46 GWAS and 96 Expression genes contributing to the core gene list via functional annotation clusters of the gene ontology terms revealed that the highest gene counts were in the regulation of macromolecule and RNA biosynthetic and metabolic processes on both lists. This was not entirely unexpected as T2D is characterized as a metabolic disorder [1, 18] but it was interesting to see that despite the little overlap of genes between the study methods, there was an overlap of 24 GO terms with the highest ranking terms on both lists associated with metabolic processes. Some of the metabolism-associated GO terms are regulation of RNA metabolic process (GO:0051252) from the GWAS analysis and mRNA metabolic process (GO:0016071) from the microarray expression analysis. Defective RNA metabolism has been previously implicated in the aberrant expression of the insulin receptor gene which could have serious consequences for T2D via insulin

resistance [218]. These terms created confidence in the data going forward with subsequent analyses as the process and method of gene selection was also crucial to the study outcome.

It is, however, important to note that the potential of circular argument does exist in the combined use of GWAS/microarray data and functional analysis as databases are generally interconnected. As a result, there is a possibility of functions being assigned to genes from the GWAS or microarray databases from which the genes were identified. Caution thus needs to be taken in the interpretation of the data. Also, since gene function is inferred from a gene similarity matrix that involves pathways and gene-GO-term enrichment, amongst other measures as previously stated [121, 122], it is important to take into consideration the dynamic nature of these annotations as new information becomes available and that not all of these annotations are experimentally validated. Whilst experimentally validated annotations are ideal, functional annotations inferred from pathway memberships, sequence similarity scores, co-occurrence probability, etc., as utilized to some extent by DAVID, tend to provide a relatively good insight into gene functions. A schematic overview of some of the significant functional annotation terms associated with GWAS and microarray expression studies and a possible relationship/role in the pathogenesis of T2D is shown in figure 4.1 below.

It is not believed that he exclusion of the 25 genes not identified by DAVID from a combined total of 663 non-redundant genes from the analysis in this instance, had a substantial impact on the discovery of novel T2D candidates or the research outcome as a whole.

4.1.3 Population (Allele) Frequency Analysis

The 1000 Genomes project employed a sequence-based approach in its identification of many novel SNPs across different populations, and this is one of the major achievements of the project [119]. This is in contrast to genotyping-based methods used in the Human Genome Diversity Project (HGDP) and HapMap projects [142–144, 219, 220]. The 1000 Genomes project provides an unbiased estimate of human genetic variation in different populations across the globe [119].

In analysing the SNPs in a population context, an estimator of genetic differentiation was utilized. The fixation index, F_{ST}, is amongst the most commonly used estimators

of genetic differentiation between populations [221]. The Perl script that was used to implement this statistical method allows for the direct input of PLINK data formats, thereby minimizing any pre-analysis manipulation of the study data. The number of genes that showed inter-continental differentiation after the two-step cut-off criteria was considerably small. However, all but 1 (*SNRPD1*) have been associated to different extents with the pathogenesis of T2D. The *NOTCH2* gene encodes a member of the Notch family. The resulting protein is cleaved in the trans-Golgi network and serves as a receptor for membrane-bound ligands. Notch family members tend to influence a variety of developmental processes by controlling cell fate decisions - it has been shown to play a role in renal, hepatic and vascular development as well as in the implementation of differentiation, proliferation and apoptotic programs [222]. *NOTCH2* is very likely associated with beta cell function in the pancreas as it plays a critical role in the pancreatic development of the fetus [223, 224].

The *KIF11* gene encodes a motor family protein. Members of this protein family are known to have gene products that are involved in centrosome separation, chromosome positioning and establishing a bipolar spindle during mitosis [225]. *KIF11* resides in an established T2D-susceptibility locus on chromosome 10q23.33. This locus also houses the insulin-degrading enzyme, *IDE* and *HHEX* genes. The *HHEX-IDE-KIF11* locus was first associated with T2D risk in individuals of European ancestry but has also been replicated in Chinese and Japanese populations [64, 226, 227]. Studies have shown this locus to be associated with increased BMI in childhood [228], a risk factor for obesity and T2D, with mRNA and protein levels of the *KIF11* gene dropping drastically at the start of adipogenesis [229, 230].

The *NR2F2* gene encodes a member of the steroid hormone superfamily of nuclear receptors [231]. Also referred to as *COUP-TFII*, *NR2F2* has a broad tissue expression profile, including the pancreas. It has been shown to be both a target and regulator of the Wnt/beta-catenin signalling system [232, 233], a pathway that is involved in insulin sensitivity and adipocyte differentiation [234]. This pathway has been shown together with the hedgehog signalling pathway to repress adipogenesis in mammalian systems [235]. *NR2F2* is known to act downstream of hedgehog signalling and is critical to the full expression of the anti-adipogenic effect driven by the sonic hedgehog anti-adipogenic factor [236]. It can perhaps then be assumed

that a polymorphism in the *NR2F2* gene could possibly disrupt its dominant role in the repression of adipocyte differentiation [236], which could contribute to the development of T2D.

The *RNGTT* gene codes for a bi-functional mRNA–capping enzyme that gets recruited to the transcription complex by the phosphorylation of RNA polymerase II [231]. It is involved with the mRNA surveillance pathway which serves as a quality control mechanism that detects and degrades anomalous mRNAs [128].

The *RPL35A* gene encodes a ribosomal protein that is required for the proliferation and viability of hematopoietic cells. It is involved in the reactome pathway for the metabolism of proteins. This pathway involves the regulation of insulin-like growth factor (IGF) transport and the uptake of the insulin-like growth factor binding proteins (IGFBPs) some of which are expressed in the liver [237]. Insulin and IGF have overlapping functions and control several aspects of growth, survival and metabolism in a wide range of mammalian tissues [238]. A dysfunction in the reactome pathway could possibly produce disrupted insulin-like effects that might be consequential in the development of T2D as insulin and *IGF-1* both stimulate the Wnt/beta-catenin pathway [239].

The *RBM38* gene encodes an RNA-binding protein that appears to be expressed in a number of tissues including the liver, spleen, lung, brain and pancreas. It is involved in the regulation of RNA splicing, negative regulation of cell proliferation, mRNA processing and DNA damage response [231]. This protein binds specifically to the 3'-UTR of cyclin-dependent kinase inhibitor 1A, *CDKN1A* transcripts, thereby maintaining the stability of these transcripts [240]. *CDKN1A* has been shown to have altered DNA methylation profile and differential expression in human T2D islets that contribute to disrupted insulin and glucagon secretion highlighting the role of epigenetics in the pathogenesis of T2D [241]. A SNP near the *RBM38* gene has been previously identified to be associated with T2D in Punjabis with Pakistani ancestry [149].

Interestingly, the variants showing great genetic differentiation in this study in genes like *RBM38*, *KIF11* and *NOTCH2* generally differ from the specific variants that have been previously identified to show T2D association in the literature. For example, rs328506, rs7923837 and rs1111875, and rs10923931 from *RBM38*, *KIF11* and

NOTCH2 respectively were not among the list of highly differentiated SNPs in the African populations analysed even though they have been observed to show strong association to T2D in existing literature [149, 181, 242].. The variants identified in this study will perhaps give us some insight into the African genetic variation and its contribution to the nature of the disease. It is, however, important to note that for this study, it is assumed that the oldest human populations reside on the African continent and as a result have greater genetic variation than populations that have migrated out of Africa and have been through several bottle necks. In addition Khoisan admixture among many black African populations has further contributed to the genetic variation on the continent [243, 244].

Since the essence of this study was to identify not just variants that differ between Africa and other continents but variants that are also selected in Africa that could either confer a risk to or protect against T2D, it was important to see which of the differentiated genes had signatures of selection in the African populations. Signatures of selection were also sought for these genes in the European and Asian populations to ensure that patterns of selection seen in the African populations were indeed unique to, or different in, Africa and not duplicated in the non-African populations.

The *SNRPD1* gene encodes a small nuclear ribonucleoprotein (*SNRNP*) that can act as a charged protein scaffold to strengthen *SNRNP-SNRNP* interactions through nonspecific electrostatic contacts with RNA or just assist in the assembly of SNRNP [231, 245]. Its involvement has been proposed in the spliceosome pathway and the RNA splicing process [246].It is rather interesting to find an identical set of SNPs from the *SNRPD1* gene showing signatures of selection in both the TSI and YRI populations whereas only a portion of these SNPs were selected in the LWK population. This is perhaps the result of gene flow as a consequence of one of the oldest population admixtures in sub-Saharan Africa in the central West African populations about 4,000 – 12,000 years ago. This migration into Africa from Eurasia occurred during a period when the Sahara desert was thought to have become green due to humid conditions [247]. This mixture probably resulted in independent adaptations in both populations as the direction of selection of the SNPs, based on positive and negative iHS scores, are the same. It is also important to note that in the investigation of differences between African and non-African genome, the extent of

admixture between Africans and Neanderthals and Denovisans is less than that observed in non-African genomes. It has been proposed that some selection has occurred for the Neanderthal contributions, for example in genomic regions that contain genes involved in lipid catabolism [248].

In terms of genetic distribution amongst populations in this analysis, only 1 of the 8 differentiated *SNRPD1* SNPs, rs2959527, was common to the 3 populations. This SNP was not in LD with any other SNP on the list of significantly differentiated SNPs or selected SNPs in the African populations. Of greater interest, however, is the SNP, rs2847139 that was selected only in an African population, LWK, and significantly differentiated from the non-African populations. The high LD observed between this SNP and 3 (rs2850568, rs2959525, rs3017641) other highly differentiated SNPs suggests a possible haplotype block in the *SNRPD1* gene being considerably implicated in the pathogenesis of T2D. Although the *SNRPD1* gene has not been previously associated with T2D in existing literature, the presence of *SNRPD1* SNPs in the iHS results of both African populations as well as F_{ST} analysis results in this study perhaps present it as a novel candidate for further investigation with respect to T2D. With the knowledge that *SNRPD1* features prominently in the spliceosome pathway [246], which happens to rank highly in the functional annotation results of the expression gene set, it is possible that alternative splicing of transcripts are abnormally regulated as a result of a polymorphism in the gene leading to an increased risk of T2D development via an as yet unknown mechanism. It has been previously shown that the recognition of splicing sites is dependent on the protein composition of the spliceosome [249]. The spliceosome architecture could thus be altered by the dysregulated expression of a coding gene and consequently its splicing process which may have implications on different biochemical levels. A proinsulin gene splice variant has been previously demonstrated to increase the translation efficiency in human pancreatic islets [250]. Also, a number of obesity-related genes which may cause a predisposition to T2D such as Lipin 1(*LPIN1*), low density lipoprotein receptor (*LDLR*) and insulin receptor (*INSR*) are regulated by alternative splicing mechanisms [251–254]. Splicing factor, arginine/serine-rich 10 (transformer 2 homolog, Drosophila), *SFRS10,* known as transformer-2 protein homolog beta, *TRA2B*, in humans, is an RNA-splicing protein that has been shown to be down regulated in the muscle and liver of obese

individuals, This protein was demonstrated to help in the regulation of *LPIN1*, a key regulator of lipid metabolism [255]. It has also been proposed that a defective alternative splicing mechanism, can lead to a truncated less functional isoform in the *TCF7L2* gene, the most strongly associated T2D gene to date [256]. This isoform is associated with elevated levels of serum free fatty acids and plasma glucose as well as with adipose tissue insulin resistance with consequences for T2D [251]. The overexpression of different lengths of *TCF7L2* mRNA variants have also been demonstrated to have either protective effects on beta-cell function and survival or induced impaired insulin secretion and apoptosis in human islets [257, 258].

It is possible that the role of *SNRPD1* in the spliceosome conferred it with a crucial regulatory role that has made it a candidate for selection. However, as its targets are yet to be characterized, it is difficult to make such speculations. Further studies to elucidate the regulatory pathways and *SNRPD1*-associated genes will very likely provide more insight and possibly uncover its functional significance. The emphasis of this section has been on some of the SNPs in the *SNRPD1* gene and a possible mechanism of action via the spliceosome complex in suggesting its role in the pathogenesis of T2D in Africa. It is, however, necessary to guard against over interpretation of the results as it is difficult to speculate on whether the selection observed in this gene is as a result of a consequence of the disease or of predisposing factors that increase the risk of the development of T2D. Also, the inclusion of the intergenic SNPs in the downstream analyses would have arguably enriched the selection process, however, this was primarily a gene-centric approach. The aim of the study was to identify potentially novel population-specific gene-associated variants in the queried African populations. The definition of novelty of SNPs for the scope of this research is the presence of SNPs in the queried African populations in the 1000 Genomes data set (Phase1, version 3, October 2012 upload) that have not been previously identified in other populations. The majority of the SNPs in this upload had been assigned rsIDs such that the effect of the exclusion of SNPs without rsIDs would be minimal as only a small number of SNPs were excluded. However, it is also possible that the excluded SNPs may actually be potentially interesting candidate SNPs for T2D in Africans

The computational approach employed in this study primarily reanalysed microarray raw data and incorporated the results of GWAS in the prioritization of T2D-relevant

genes and subsequently SNPs. Like in many other meta-analyses-focused studies, pathway memberships and gene ontologies were considered in the gene prioritization steps. However, the integration of the 1000 Genomes dataset in an attempt to retrieve possibly novel T2D-relavant SNPs in African populations had not been previously done. This study therefore presents the possibility of identifying potentially relevant disease-associated SNPs from existing data by utilizing curated databases to computationally make inferences from populations that were not involved in the original studies.

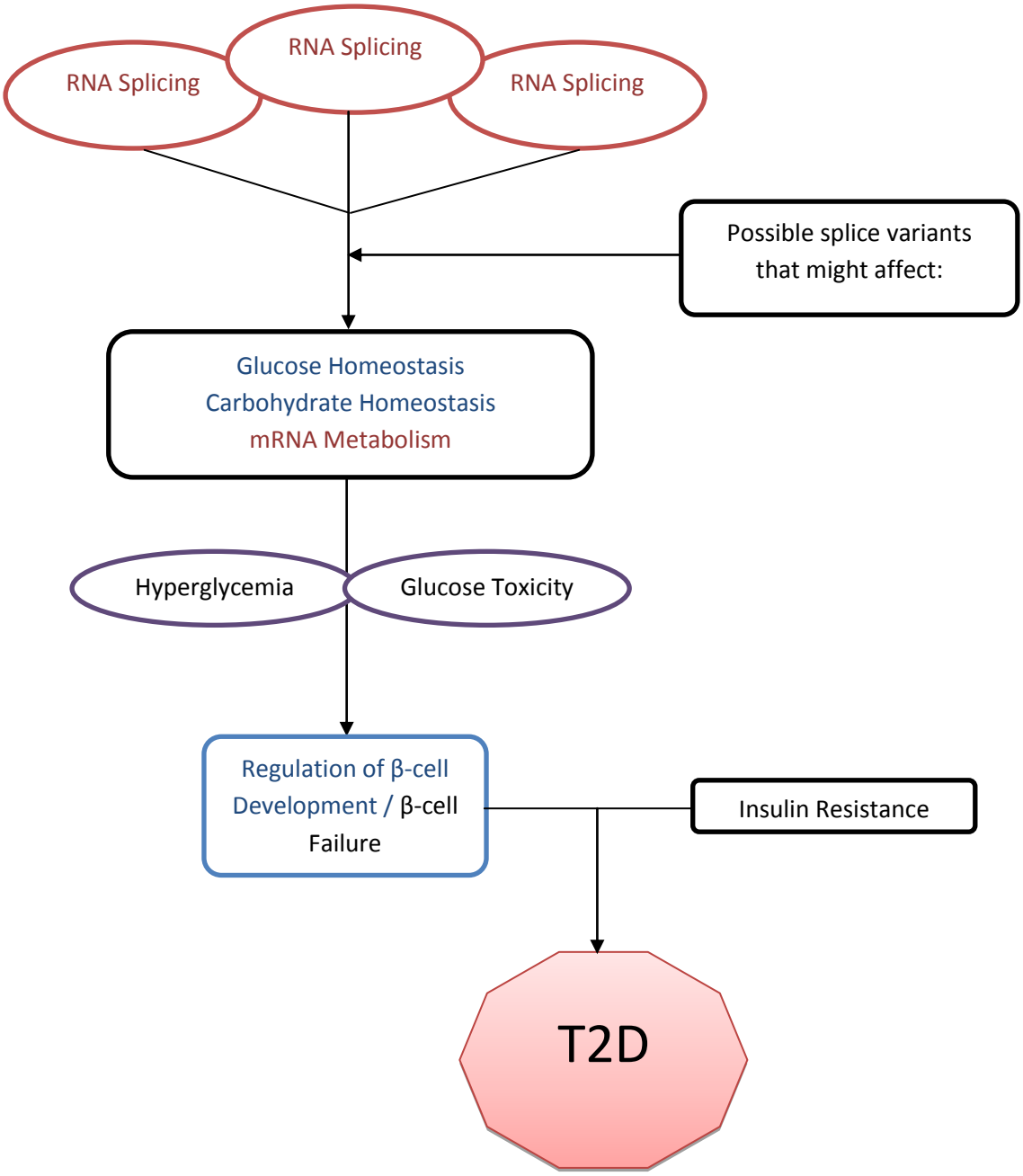


Figure 4.1 A schema of a possible relationship between the functional annotation terms from GWAS and microarray studies and the pathogenesis of T2D. Blue and red colours represent terms associated with GWAS and microarray studies respectively.

4.2 Conclusion

The multi-factorial nature of type 2 diabetes complicates the elucidation of the aetiology of the disease. Beta-cell dysfunction and insulin resistance remain the two relative constant features in the development of T2D; however, there are multiple avenues and processes that could lead to these conditions. Different global environments can influence the epigenome, which regulates gene expression patterns, and contribute to varying extents to the development of the disease. For example, in T2D, diet and physical inactivity play a major role. It thus makes logical sense that the component of the disease where some measure of insight can be gained is in its genetics. Studies have shown that humans are about 99.9% genetically similar [259]. The ~0.1% variation, however, has proven to be quite consequential with respect to disease aetiology and appropriate treatment courses.

This study attempted to take advantage of the similar clinical presentation of T2D across populations to hone in on the possible differences in the African nature of the disease as research has been extensive in non-African populations and T2D molecular research on the continent has been relatively minimal. The genetic differentiation observed, based on the F_{ST} analysis, of different sets of variants from those identified in other populations in genes that have been previously associated with T2D perhaps suggests that there is some specificity in T2D-susceptibility on the African continent. The selection and genetic differentiation of the *SNRPD1* gene variants, which do not appear to have been previously associated with T2D in the literature, perhaps presents it as an interesting candidate in the pathogenesis of T2D, worth studying more intricately.

It is important to note that the analysis approach and general workflow applied in this project that prioritized streamlining highly differentiated loci to those selected in the queried African populations may have eliminated the chance of identifying other potential variants of interest in the pathogenesis of T2D, that are common across multiple continental populations. However, the main aim of this work was to identify those T2D-associated variants that show some level of specificity to the interrogated African populations to give us some possible insights into the pathogenesis of the disease on the continent.

Considering the prevalence projections for type 2 diabetes on the African continent and the current state of research efforts in the field, it would appear as though the future is rather bleak. It is anticipated, however, that proposed large scale research efforts on the continent like the Human Hereditary and Health in Africa consortium, H3Africa, will provide some insight not only to the African nature of the disease but also to identifying more of the missing heritability component in the pathogenesis of the disease. Elucidating some of the “African nature” of T2D would in the short term provide more appropriate continental information for chip design and in the long term, more targeted therapies to combat and possibly prevent the progression of the disease.

4.2.1 Study Limitations

- **GWAS**

It is highly unlikely that the representation of the genome from most of the current GWA studies accurately capture the diversity present in African populations. It is, therefore, possible that potentially T2D-relevant variants may have been missed.

- **Gene expression studies**

It is known that the non-homogenous nature of the tissue samples (except in study 6 where a laser capture was done) will affect the expression profiles of the genes. Also, the use of a similar analyses protocol on experiments with different hypotheses and questions, with the overarching aim of a ‘one-touch’ analysis pipeline, may have been too stringent in some of the studies. The sample sizes may have had an effect on the limma analysis results as some of the studies had rather small sample sizes. It will be worthwhile for the criteria for expression data deposition to be reviewed as some of the available data are rather inconsistent and the sample sizes rather small.

It is also possible that differences in gene expression observed may not be related to polymorphisms within that gene but to polymorphisms in other genes that control the transcription of the index gene.

- **Study hypothesis**

It is possible that the use of non-African samples to retrieve African-specific information on T2D may have limited the outcome of this study as some genes might have been missed. The aim of the study was not to necessarily identify novel genes, but rather novel African-specific variants in existing genes. However, it is believed

that the clinical similarity in the presentation of the disease [260] gives some measure of validity to the study rationale.

4.2.2 Future Directions

The logical next step will be to test the results of this selection and analysis approach in a high risk group in Africa by genotyping the SNPs of interest from not only the *SNRPD1* gene but also those from the 6 other highly differentiated genes. It would also be interesting to investigate the downstream targets of these genes in the tissues most affected by T2D.

With the changing landscape of genomic science, it will be interesting to see what whole genome sequencing of type 2 diabetic individuals in Africa would highlight. Also, most of the GWAS performed to date have a bias in their experimental design as they generally detect susceptibility effects attributable to common SNPs. Thus, a more in-depth analysis of rare variants as well as copy number variants might provide some clues to the pathogenesis of T2D.

REFERENCES

1. Kuzuya T, Nakagawa S, Satoh J, Kanazawa Y, Iwamoto Y, Kobayashi M, Nanjo K, Sasaki A, Seino Y, Ito C, Shima K, Nonaka K, Kadowaki T: **Report of the Committee on the classification and diagnostic criteria of diabetes mellitus**. *Diabetes Res Clin Pract* 2002, **55**:65–85.

2. McCarthy MI: **Genomics, Type 2 Diabetes, and Obesity**. *N Engl J Med* 2010, **363**:2339–2350.

3. Keller MP, Choi Y, Wang P, Belt Davis D, Rabaglia ME, Oler AT, Stapleton DS, Argmann C, Schueler KL, Edwards S, Steinberg HA, Chaibub Neto E, Kleinhanz R, Turner S, Hellerstein MK, Schadt EE, Yandell BS, Kendziorski C, Attie AD: **A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility**. *Genome Res* 2008, **18** (5):706–716.

4. Kumanyika S, Jeffery RW, Morabia A, Ritenbaugh C AV: **Obesity prevention: the case for action**. *Int J Obes* 2002, **26**:425–436.

5. Boden G: **Role of Fatty Acids in the Pathogenesis of Insulin Resistance and NIDDM**. *Diabetes* 1997, **46** (1):3–10.

6. Hales CN, Barker DJP: **The thrifty phenotype hypothesis: Type 2 diabetes** . *Br Med Bull* 2001, **60** (1):5–20.

7. Hales CN, Barker DJP: **Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis**. *Diabetologia* 1992, **35**:595–601.

8. Li Y, He Y, Qi L, Jaddoe VW, Feskens EJM, Yang X, Ma G, Hu FB: **Exposure to the Chinese Famine in Early Life and the Risk of Hyperglycemia and Type 2 Diabetes in Adulthood**. *Diabetes* 2010, **59** (10):2400–2406.

9. Ravelli ACJ, van der Meulen JHP: **Prenatal exposure to famine and health in later life**. *The Lancet* 1998, **351**:1362.

10. Boden G, She P, Mozzoli M, Cheung P, Gumireddy K, Reddy P, Xiang X, Luo Z, Ruderman N: **Free Fatty Acids Produce Insulin Resistance and Activate the Proinflammatory Nuclear Factor-κB Pathway in Rat Liver**. *Diabetes* 2005, **54** (12):3458–3465.

11. Trayhurn P, Wood IS: **Adipokines: inflammation and the pleiotropic role of white adipose tissue**. *Br J Nutr* 2004, **92**:347–355.

12. Nonogaki K, Fuller GM, Fuentes NL, Moser AH, Staprans I, Grunfeld C, Feingold KR: **Interleukin-6 stimulates hepatic triglyceride secretion in rats.** *Endocrinology* 1995, **136**:2143–2149.

13. Antuna-Puente B, Feve B, Fellahi S, Bastard J-P: **Adipokines: the missing link between insulin resistance and obesity.** *Diabetes Metab* 2008, **34**:2–11.

14. Berg AH, Combs TP, Du X, Brownlee M, Scherer PE: **The adipocyte-secreted protein Acrp30 enhances hepatic insulin action**. *Nat Med* 2001, **7**:947–953.

15. Fruebis J, Tsao T-S, Javorschi S, Ebbets-Reed D, Erickson MRS, Yen FT, Bihain BE, Lodish HF: **Proteolytic cleavage product of 30-kDa adipocyte complement-related protein increases fatty acid oxidation in muscle and causes weight loss in mice**. *Proc Natl Acad Sci* 2001, **98** (4):2005–2010.

16. Ravussin E: **Adiponectin enhances insulin action by decreasing ectopic fat deposition**. *Pharmacogenomics J* 2002, **2**:4–7.

17. Turcotte LP, Fisher JS: **Skeletal Muscle Insulin Resistance: Roles of Fatty Acid Metabolism and Exercise**. *Phys Ther* 2008, **88** (11):1279–1296.

18. Talchai C, Lin H V, Kitamura T, Accili D: **Genetic and biochemical pathways of β-cell failure in type 2 diabetes**. *Diabetes, Obes Metab* 2009, **11**:38–45.

19. Accili D, Ahrén B, Boitard C, Cerasi E, Henquin J-C, Seino S: **What ails the β-cell?** *Diabetes, Obes Metab* 2010, **12**:1–3.

20. Talchai C, Xuan S, Lin HV, Sussel L, Accili D: **Pancreatic β Cell Dedifferentiation as a Mechanism of Diabetic β Cell Failure**. *Cell* 2014, **150**:1223–1234.

21. Weinberg N, Ouziel-Yahalom L, Knoller S, Efrat S, Dor Y: **Lineage Tracing Evidence for In Vitro Dedifferentiation but Rare Proliferation of Mouse Pancreatic β-Cells**. *Diabetes* 2007, **56** (5):1299–1304.

22. Prokopenko I, McCarthy MI, Lindgren CM: **Type 2 diabetes: new genes, new understanding.** *Trends Genet* 2008, **24**:613–21.

23. Rossi-Espagnet A, Goldstein GB TI: **Urbanization and health in developing countries: a challenge for health for all**. *World Heal Stat Q* 1991, **44**:185–244.

24. Tekola-Ayele F, Adeyemo AA, Rotimi CN: **Genetic epidemiology of type 2 diabetes and cardiovascular diseases in Africa.** *Prog Cardiovasc Dis* 2013, **56**:251–60.

25. Wild S, Roglic G, Green A, Sicree R, King H: **Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030** . *Diabetes Care* 2004, **27** (5):1047–1053.

26. **IDF Diabetes Atlas** [http://www.idf.org/diabetesatlas]

27. **Fact Sheet** [www.who.int]

28. Bluestone JA, Herold K, Eisenbarth G: **Genetics, pathogenesis and clinical interventions in type[thinsp]1 diabetes**. *Nature* 2010, **464**:1293–1300.

29. American Diabetes Association [ADA]: **Diagnosis and Classification of Diabetes Mellitus**. *Diabetes Care* 2008, **31**(Supplement 1):S55–S60.

30. Metzger BE CD: **Summary and recommendations of the Fourth International Workshop-Conference on Gestational Diabetes Mellitus. The Organizing Committee**. *Diabetes Care* 1998, **21**(Suppl 2):B161–7.

31. Edghill EL, Bingham C, Slingerland AS, Minton JAL, Noordam C, Ellard S, Hattersley AT: **Hepatocyte nuclear factor-1 beta mutations cause neonatal diabetes and intrauterine growth retardation: support for a critical role of HNF-1β in human pancreatic development**. *Diabet Med* 2006, **23**:1301–1306.

32. Patel SG, Hsu JW, Jahoor F, Coraza I, Bain JR, Stevens RD, Iyer D, Nalini R, Ozer K, Hampe CS, Newgard CB, Balasubramanyam A: **Pathogenesis of A−β+ Ketosis-Prone Diabetes**. *Diabetes* 2013, **62** (3):912–922.

33. Mauvais-Jarvis F, Sobngwi E, Porcher R, Riveline J-P, Kevorkian J-P, Vaisse C, Charpentier G, Guillausseau P-J, Vexiau P, Gautier J-F: **Ketosis-Prone Type 2 Diabetes in Patients of Sub-Saharan African Origin: Clinical Pathophysiology and Natural History of β-Cell Dysfunction and Insulin Resistance** . *Diabetes* 2004, **53** (3):645–653.

34. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng H, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, Shan Y, Li S, Yang Q, Asan, Ni P, Tian G, Xu J, Liu X, Jiang T, Wu R, et al.: **Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude**. *Science (80-)* 2010, **329**:75–78.

35. Motala AA, Pirie FJ, Gouws E, Amod A, Omar MAK: **High incidence of Type 2 diabetes mellitus in South African Indians: a 10-year follow-up study**. *Diabet Med* 2003, **20**:23–30.

36. Ahrén B, Corrigan CB: **Prevalence of diabetes mellitus in north-western Tanzania**. *Diabetologia* 1984, **26**:333–336.

37. Tuei VC, Maiyoh GK, Ha C-E: **Type 2 diabetes mellitus and obesity in sub-Saharan Africa**. *Diabetes Metab Res Rev* 2010, **26**:433–445.

38. Prentice AM, Hennig BJ, Fulford AJ: **Evolutionary origins of the obesity epidemic: natural selection of thrifty genes or genetic drift following predation release[quest]**. *Int J Obes* 2008, **32**:1607–1610.

39. Harris JL, Pomeranz JL, Lobstein T, Brownell KD: **A Crisis in the Marketplace: How Food Marketing Contributes to Childhood Obesity and What Can Be Done**. *Annu Rev Public Health* 2009, **30**:211–225.

40. Mendez MA, Monteiro CA, Popkin BM: **Overweight exceeds underweight among women in most developing countries**. *Am J Clin Nutr* 2005, **81** (3):714–721.

41. Hamdy O, Porramatikul S A-OE: **Metabolic Obesity: The Paradox Between Visceral and Subcutaneous Fat**. *Curr Diabetes Rev* 2003, **2**:367–73.

42. Meigs JB, Wilson PWF, Fox CS, Vasan RS, Nathan DM, Sullivan LM, D'Agostino RB: **Body Mass Index, Metabolic Syndrome, and Risk of Type 2 Diabetes or Cardiovascular Disease**. *J Clin Endocrinol Metab* 2006, **91**:2906–2912.

43. Yoon K-H, Lee J-H, Kim J-W, Cho JH, Choi Y-H, Ko S-H, Zimmet P, Son H-Y: **Epidemic obesity and type 2 diabetes in Asia.** *Lancet* 2006, **368**:1681–8.

44. Huxley R, James WPT, Barzi F, Patel J V, Lear SA, Suriyawongpaisal P, Janus E, Caterson I, Zimmet P, Prabhakaran D, Reddy S, Woodward M, Collaboration the O in A: **Ethnic comparisons of the cross-sectional relationships between measures of body size with diabetes and hypertension**. *Obes Rev* 2008, **9**:53–61.

45. Lear SA, Humphries KH, Kohli S, Chockalingam A, Frohlich JJ, Birmingham CL: **Visceral adipose tissue accumulation differs according to ethnic background: results of the Multicultural Community Health Assessment Trial (M-CHAT)**. *Am J Clin Nutr* 2007, **86** (2):353–359.

46. Deurenberg P, Deurenberg-Yap M, Guricci S: **Asians are different from Caucasians and from each other in their body mass index/body fat per cent relationship**. *Obes Rev* 2002, **3**:141–146.

47. Lo M, Mitsnefes M: **Adiponectin, cardiovascular disease, chronic kidney disease: emerging data on complex interactions**. *Pediatr Nephrol* 2012, **27**:521–527.

48. Fowler MJ: **Microvascular and Macrovascular Complications of Diabetes**. *Clin Diabetes* 2008, **26** (2):77–82.

49. Creager MA, Lüscher TF, of prepared with the assistance, Cosentino F, Beckman JA: **Diabetes and Vascular Disease: Pathophysiology, Clinical Consequences, and Medical Therapy: Part I** . *Circ* 2003, **108** (12):1527–1532.

50. Pettitt DJ, Saad MF, Bennett PH, Nelson RG, Knowler WC: **Familial predisposition to renal disease in two generations of Pima Indians with Type 2 (non-insulin-dependent) diabetes mellitus**. *Diabetologia* 1990, **33**:438–443.

51. Greene DA, Lattimer SA, Sima AAF: **Sorbitol, Phosphoinositides, and Sodium-Potassium-ATPase in the Pathogenesis of Diabetic Complications**. *N Engl J Med* 1987, **316**:599–606.

52. Centers for Disease Control and Prevention [CDC]: **Diabetes Fact Sheet**. 2007.

53. Grundy SM, Benjamin IJ, Burke GL, Chait A, Eckel RH, Howard B V, Mitch W, Smith SC, Sowers JR: **Diabetes and Cardiovascular Disease: A Statement for Healthcare Professionals From the American Heart Association** . *Circ* 1999, **100** (10):1134–1146.

54. Krauss RM, Siri PW: **Dyslipidemia in type 2 diabetes.** *Med Clin North Am* 2004, **88**:897–909, x.

55. Del Pilar Solano M, Goldberg RB: **Management of diabetic dyslipidemia.** *Endocrinol Metab Clin North Am* 2005, **34**:1–25, v.

56. Chahil TJ, Ginsberg HN: **Diabetic dyslipidemia.** *Endocrinol Metab Clin North Am* 2006, **35**:491–510, vii–viii.

57. Almgren P, Lehtovirta M, Isomaa B, Sarelin L, Taskinen MR, Lyssenko V, Tuomi T, Groop L: **Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study**. *Diabetologia* 2011, **54**:2811–2819.

58. Taskinen M-R: **Diabetic dyslipidemia**. *Atheroscler Suppl* 2002, **3**:47–51.

59. KJ Basile MJ: **Genetic Susceptibility to Type 2 Diabetes and Obesity: Follow-Up of Findings from Genome-Wide Association Studies**. *Int J Endocrinol* 2014.

60. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding C-J, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li X-Y, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, et al.: **A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants**. *Sci* 2007, **316** (5829):1341–1345.

61. Manning AK, Hivert M-F, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, Rybin D, Liu C-T, Bielak LF, Prokopenko I, Amin N, Barnes D, Cadby G, Hottenga J-J, Ingelsson E, Jackson AU, Johnson T, Kanoni S, Ladenvall C, Lagou V, Lahti J, Lecoeur C, Liu Y, Martinez-Larrad MT, Montasser ME, Navarro P, Perry JRB, Rasmussen-Torvik LJ, Salo P, Sattar N, et al.: **A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance**. *Nat Genet* 2012, **44**:659–669.

62. Saxena R, Elbers CC, Guo Y, Peter I, Gaunt TR, Mega JL, Lanktree MB, Tare A, Castillo BA, Li YR, Johnson T, Bruinenberg M, Gilbert-Diamond D, Rajagopalan R, Voight BF, Balasubramanyam A, Barnard J, Bauer F, Baumert J, Bhangale T, Böhm BO, Braund PS, Burton PR, Chandrupatla HR, Clarke R, Cooper-DeHoff RM, Crook ED, Davey-Smith G, Day IN, de Boer A, et al.: **Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci.** *Am J Hum Genet* 2012, **90**:410–25.

63. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H: **Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study.** *Diabetologia* 1999, **42**:139–145.

64. Wu Y, Li H, Loos RJF, Yu Z, Ye X, Chen L, Pan A, Hu FB, Lin X: **Common variants in CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8 and HHEX/IDE genes are associated with type 2 diabetes and impaired fasting glucose in a Chinese Han population**. *Diabetes* 2008.

65. Rotimi CN, Chen G, Adeyemo AA, Furbert-Harris P, Guass D, Zhou J, Berg K, Adegoke O, Amoah A, Owusu S, Acheampong J, Agyenim-Boateng K, Eghan BA, Oli J, Okafor G, Ofoegbu E, Osotimehin B, Abbiyesuku F, Johnson T, Rufus T, Fasanmade O, Kittles R, Daniel H, Chen Y, Dunston G, Collins FS: **A Genome-Wide Search for Type 2 Diabetes Susceptibility Genes in West Africans: The Africa America Diabetes Mellitus (AADM) Study** . *Diabetes* 2004, **53** (3):838–841.

66. Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, Rafnar T, Gudbjartsson D, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Blondal T, Stacey SN, Helgason A, Gunnarsdottir S, Olafsdottir A, Kristinsson KT, Birgisdottir B, Ghosh S, Thorlacius S, Magnusdottir D, Stefansdottir G, Kristjansson K, Bagger Y, Wilensky RL, Reilly MP, Morris AD, Kimber CH, et al.: **Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes**. *Nat Genet* 2007, **39**:977–983.

67. Helgason A, Palsson S, Thorleifsson G, Grant SFA, Emilsson V, Gunnarsdottir S, Adeyemo A, Chen Y, Chen G, Reynisdottir I, Benediktsson R, Hinney A, Hansen T, Andersen G, Borch-Johnsen K, Jorgensen T, Schafer H, Faruque M, Doumatey A, Zhou J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Sigurdsson G, Hebebrand J, Pedersen O, Thorsteinsdottir U, Gulcher JR, et al.: **Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution**. *Nat Genet* 2007, **39**:218–225.

68. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, Baker A, Snorradottir S, Bjarnason H, Ng MCY, Hansen T, Bagger Y, Wilensky RL, Reilly MP, Adeyemo A, Chen Y, Zhou J, Gudnason V, Chen G, Huang H, Lashley K, Doumatey A, So W-Y, Ma RCY, Andersen G, Borch-Johnsen K, et al.: **A variant in CDKAL1 influences insulin response and risk of type 2 diabetes**. *Nat Genet* 2007, **39**:770–775.

69. Cauchi S, Achhab Y, Choquet H, Dina C, Krempler F, Weitgasser R, Nejjari C, Patsch W, Chikri M, Meyre D, Froguel P: **TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis**. *J Mol Med* 2007, **85**:777–782.

70. Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, Lopez-Bigas N, Ouzounis C, Perez-Iratxeta C, Andrade-Navarro MA, Adeyemo A, Patti ME, Semple CAM, Hide W: **Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes**. *Nucleic Acids Res* 2006, **34** (10):3067–3081.

71. Audouze K, Brunak S, Grandjean P: **A computational approach to chemical etiologies of diabetes**. *Sci Rep* 2013, **3**.

72. Botstein D, Risch N: **Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease**. *Nat Genet* .

73. Lunetta KL: **Genetic Association Studies**. *Circ* 2008, **118** (1):96–101.

74. Hattersley AT, McCarthy MI: **What makes a good genetic association study?** *Lancet* 2005, **366**:1315–23.

75. Hanis CL, Boerwinkle E, Chakraborty R, Ellsworth DL, Concannon P, Stirling B, Morrison VA, Wapelhorst B, Spielman RS, Gogolin-Ewens KJ, Shephard JM, Williams SR, Risch N, Hinds D, Iwasaki N, Ogata M, Omori Y, Petzold C, Rietzsch H, Schroder H-E, Schulze J, Cox NJ, Menzel S, Boriraj V V, Chen X, Lim LR, Lindner T, Mereu LE, Wang Y-Q, Xiang K, et al.: **A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2**. *Nat Genet* 1996, **13**:161–166.

76. Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PEH, del Bosque-Plata L, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI: **Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus**. *Nat Genet* 2000, **26**:163–175.

77. Gloyn AL, Weedon MN, Owen KR, Turner MJ, Knight BA, Hitman G, Walker M, Levy JC, Sampson M, Halford S, McCarthy MI, Hattersley AT, Frayling TM: **Large-Scale Association Studies of Variants in Genes Encoding the Pancreatic β-Cell KATP Channel Subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) Confirm That the KCNJ11 E23K Variant Is Associated With Type 2 Diabetes** . *Diabetes* 2003, **52** (2):568–572.

78. Leeper NJ, Kullo IJ, Cooke JP: **Genetics of peripheral artery disease.** *Circulation* 2012, **125**:3220–8.

79. Stankiewicz P, Lupski JR: **Structural Variation in the Human Genome and its Role in Disease**. *Annu Rev Med* 2010, **61**:437–455.

80. Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A: **Segmental copy number variation shapes tissue transcriptomes**. *Nat Genet* 2009, **41**:424–429.

81. Bae JS, Cheong HS, Kim J-H, Park BL, Kim J-H, Park TJ, Kim JY, Pasaje CFA, Lee JS, Park Y-J, Park M, Park C, Koh I, Chung Y-J, Lee J-Y, Shin HD: **The Genetic Effect of Copy Number Variations on the Risk of Type 2 Diabetes in a Korean Population**. *PLoS One* 2011, **6**:e19091.

82. Jeon J-P, Shim S-M, Nam H-Y, Ryu G-M, Hong E-J, Kim H-L, Han B-G: **Copy number variation at leptin receptor gene locus associated with metabolic traits and the risk of type 2 diabetes mellitus.** *BMC Genomics* 2010, **11**:426.

83. Rakyan VK, Beck S: **Epigenetic variation and inheritance in mammals.** *Curr Opin Genet Dev* 2006, **16**:573–7.

84. Lund AH, van Lohuizen M: **Epigenetics and cancer**. *Genes Dev* 2004, **18** (19):2315–2335.

85. Gräff Dohoon Dobbin, Matthew M. Tsai, Li-Huei JK: **Epigenetic Regulation of Gene Expression in Physiological and Pathological Brain Processes**. *Physiol Rev* 2011, **91**:603–649.

86. Bell CG, Finer S, Lindgren CM, Wilson GA, Rakyan VK, Teschendorff AE, Akan P, Stupka E, Down TA, Prokopenko I, Morison IM, Mill J, Pidsley R, Deloukas P, Frayling TM, Hattersley AT, McCarthy MI, Beck S, Hitman GA, Consortium IT 2 D 1q: **Integrated Genetic and Epigenetic Analysis Identifies Haplotype-Specific Methylation in the <italic>FTO</italic> Type 2 Diabetes and Obesity Susceptibility Locus**. *PLoS One* 2010, **5**:e14040.

87. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE, Lumey LH: **Persistent epigenetic differences associated with prenatal exposure to famine in humans**. *Proc Natl Acad Sci* 2008, **105** (44):17046–17049.

88. Rusk N, Kiermer V: **Primer: Sequencing[mdash]the next generation**. *Nat Meth* 2008, **5**:15.

89. Metzker ML: **Sequencing technologies [mdash] the next generation**. *Nat Rev Genet* 2010, **11**:31–46.

90. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ: **Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing**. *Sci* 2010, **328** (5978):636–639.

91. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, Ha C, Johnson S, Kennemer MI, Mohan S, Nazarenko I, Watanabe C, Sparks AB, Shames DS, Gentleman R, de Sauvage FJ, Stern H, Pandita A, Ballinger DG, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z: **The mutation spectrum revealed by paired genome sequences from a lung cancer patient**. *Nature* 2010, **465**:473–477.

92. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin M-L, Ordonez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, et al.: **A comprehensive catalogue of somatic mutations from a human cancer genome**. *Nature* 2010, **463**:191–196.

93. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, Arai Y, Takahashi H, Shirakihara T, Nagasaki M, Shibuya T, Nakano K, Watanabe-Makino K, Tanaka H, Nakamura H, Kusuda J, Ojima H, Shimada K, Okusaka T, Ueno M, Shigekawa Y, Kawakami Y, Arihiro K, Ohdan H, Gotoh K, Ishikawa O, et al.: **Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators**. *Nat Genet* 2012, **44**:760–764.

94. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, de Bakker PIW, Purcell SM, Sunyaev SR: **Exome sequencing and the genetic basis of complex traits**. *Nat Genet* 2012, **44**:623–630.

95. Tanaka D, Nagashima K, Sasaki M, Funakoshi S, Kondo Y, Yasuda K, Koizumi A, Inagaki N: **Exome sequencing identifies a new candidate mutation for susceptibility to diabetes in a family with highly aggregated type 2 diabetes.** *Mol Genet Metab* 2013, **109**:112–7.

96. Hale PJ, López-Yunez AM, Chen JY: **Genome-wide meta-analysis of genetic susceptible genes for Type 2 Diabetes.** *BMC Syst Biol* 2012, **6 Suppl 3**(Suppl 3):S16.

97. Pers TH, Hansen NT, Lage K, Koefoed P, Dworzynski P, Miller ML, Flint TJ, Mellerup E, Dam H, Andreassen OA, Djurovic S, Melle I, Børglum AD, Werge T, Purcell S, Ferreira MA, Kouskoumvekaki I, Workman CT, Hansen T, Mors O, Brunak S: **Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes**. *Genet Epidemiol* 2011, **35**:318–332.

98. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DPK, Holmkvist J, Borch-Johnsen K, Jorgensen T, Sandbaek A, Lauritzen T, Hansen T, Nurbaya S, Tsunoda T, Kubo M, Babazono T, Hirose H, Hayashi M, Iwamoto Y, Kashiwagi A, Kaku K, Kawamori R, Tai ES, Pedersen O, Kamatani N, Kadowaki T, Kikkawa R, Nakamura Y, Maeda S: **SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations**. *Nat Genet* 2008, **40**:1098–1102.

99. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y, Yamagata K, Hinokio Y, Wang H-Y, Tanahashi T, Nakamura N, Oka Y, Iwasaki N, Iwamoto Y, Yamada Y, Seino Y, Maegawa H, Kashiwagi A, Takeda J, Maeda E, Shin HD, Cho YM, Park KS, Lee HK, Ng MCY, Ma RCW, et al.: **Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus**. *Nat Genet* 2008, **40**:1092–1097.

100. Silander K, Mohlke KL, Scott LJ, Peck EC, Hollstein P, Skol AD, Jackson AU, Deloukas P, Hunt S, Stavrides G, Chines PS, Erdos MR, Narisu N, Conneely KN, Li C, Fingerlin TE, Dhanjal SK, Valle TT, Bergman RN, Tuomilehto J, Watanabe RM, Boehnke M, Collins FS: **Genetic Variation Near the Hepatocyte Nuclear Factor-4α Gene Predicts Susceptibility to Type 2 Diabetes**. *Diabetes* 2004, **53** (4):1141–1149.

101. Love-Gregory LD, Wasson J, Ma J, Jin CH, Glaser B, Suarez BK, Permutt MA: **A Common Polymorphism in the Upstream Promoter Region of the Hepatocyte Nuclear Factor-4α Gene on Chromosome 20q Is Associated With Type 2 Diabetes and Appears to Contribute to the Evidence for Linkage in an Ashkenazi Jewish Population**. *Diabetes* 2004, **53** (4):1134–1140.

102. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: **A navigator for human genome epidemiology**. *Nat Genet* 2008, **40**:124–125.

103. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.** *Nucleic Acids Res* 2014, **42**(Database issue):D1001–6.

104. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *Proc Natl Acad Sci* 2009, **106** (23):9362–9367.

105. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D: **The Human Genome Browser at UCSC**. *Genome Res* 2002, **12** (6):996–1006.

106. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30** (1):207–210.

107. R Core Team: **R: A Language and Environment for Statistical Computing**. 2013.

108. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**:R80.

109. Davis S, Meltzer PS: **GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor**. *Bioinforma* 2007, **23** (14):1846–1847.

110. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy—analysis of Affymetrix GeneChip data at the probe level**. *Bioinforma* 2004, **20** (3):307–315.

111. Smyth GK, Michaud J, Scott HS: **Use of within-array replicate spots for assessing differential expression in microarray experiments**. *Bioinforma* 2005, **21** (9):2067–2075.

112. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostat* 2003, **4** (2):249–264.

113. Bolstad BM, Irizarry RA, Åstrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinforma* 2003, **19** (2):185–193.

114. Murie C, Woody O, Lee A, Nadon R: **Comparison of small n statistical tests of differential expression applied to microarrays**. *BMC Bioinformatics* 2009, **10**:1–18.

115. Lönnstedt I, Speed T: **Replicated Microarray Data**. *Stat Sin* 2002, **12**:31–4.

116. SAGE Publications I: **T Test. The SAGE Glossary of the Social and Behavioral Sciences. SAGE Publications, Inc.** 2009.

117. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *J R Stat Soc Ser B* 1995, **57**:289–300.

118. Gentleman R: **annotate: Annotation for microarrays**. .

119. The 1000 Genomes Project Consortium: **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**:56–65.

120. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses**. *Am J Hum Genet* 2007, **81**:559–575.

121. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2008, **4**:44–57.

122. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 2009, **37**:1–13.

123. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Res* 2005, **33** (suppl 1):D514–D517.

124. Becker KG, Barnes KC, Bright TJ, Wang SA: **The Genetic Association Database**. *Nat Genet* 2004, **36**:431–432.

125. Jantzen SG, Sutherland BJ, Minkley DR, Koop BF: **GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets.** *BMC Res Notes* 2011, **4**:267.

126. Becker KG, White SL, Muller J, Engel J: **BBID: the biological biochemical image database**. *Bioinforma* 2000, **16** (8):745–746.

127. Nishimura D: **BioCarta**. *Biotech Softw Internet Rep* 2001, **2**:117–120.

128. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 2000, **28** (1):27–30.

129. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD: **The PANTHER database of protein families, subfamilies, functions and pathways**. *Nucleic Acids Res* 2005, **33** (suppl 1):D284–D288.

130. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: A Library of Protein Families and Subfamilies Indexed by Function**. *Genome Res* 2003, **13** (9):2129–2141.

131. Bland JM, Altman DG: **Multiple significance tests: the Bonferroni method**. *BMJ* 1995, **310**.

132. Choudhury A, Hazelhurst S, Meintjes A, Achinike-Oduaran O, Aron S, Gamieldien J, Jalali Sefid Dashti M, Mulder N, Tiffin N, Ramsay M: **Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance**. *BMC Genomics* 2014, **15**:437.

133. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P: **Ensembl BioMarts: a hub for data retrieval across taxonomic space**. *Database* 2011, **2011** .

134. Wright S: **The genetical structure of populations**. *Ann Hum Genet* 1949, **15**:323–354.

135. Holsinger KE, Weir BS: **Genetics in geographically structured populations: defining, estimating and interpreting FST**. *Nat Rev Genet* 2009, **10**:639–650.

136. Clayton D: **snpStats: SnpMatrix and XSnpMatrix classes and methods**. 2013.

137. Wright S: *Evolution and the Genetics of Populations*. Chicago: University of Chicago Press; 1969.

138. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A Map of Recent Positive Selection in the Human Genome**. *PLoS Biol* 2006, **4**:e72.

139. Chen H, Patterson N, Reich D: **Population differentiation as a test for selective sweeps**. *Genome Res* 2010, **20**:393–402.

140. **WHAMM** [http://coruscant.itmat.upenn.edu/whamm/index.html]

141. Choudhury A, Hazelhurst S, Meintjes A, Achinike-Oduaran O, Shaun A, Gamieldien J, Jalali M, Mulder N, Tiffin N, Ramsay M: **Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance**. *Submitt (In Rev* 2014.

142. The International HapMap Consortium: **The International HapMap Project**. *Nature* 2003, **426**:789–796.

143. The International HapMap Consortium: **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2007, **449**:851–861.

144. The International HapMap 3 Consortium: **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, **467**:52–58.

145. McVean G, Auton A: *LDhat 2.2: A Package for the Population Genetic Analysis of Recombination*. 2011.

146. Matise TC, Chen F, Chen W, De La Vega FM, Hansen M, He C, Hyland FCL, Kennedy GC, Kong X, Murray SS, Ziegle JS, Stewart WCL, Buyske S: **A second-generation combined linkage–physical map of the human genome**. *Genome Res* 2007, **17** (12):1783–1786.

147. Wolfe D, Dudek S, Ritchie M, Pendergrass S: **Visualizing genomic information across chromosomes with PhenoGram**. *BioData Min* 2013, **6**:1–12.

148. Ma RCW, Hu C, Tam CH, Zhang R, Kwan P, Leung TF, Thomas GN, Go MJ, Hara K, Sim X, Ho JSK, Wang C, Li H, Lu L, Wang Y, Li JW, Lam VKL, Wang J, Yu W, Kim YJ, Ng DP, Fujita H, Panoutsopoulou K, Day-Williams AG, Lee HM, Ng ACW, Fang Y-J, Kong APS, Jiang F, Ma X, et al.: **Genome-wide association study in a Chinese population identifies a susceptibility locus for type 2 diabetes at 7q32 near PAX4**. *Diabetologia* 2013, **56**:1291–1305.

149. Saxena R, Saleheen D, Been LF, Garavito ML, Braun T, Bjonnes A, Young R, Ho WK, Rasheed A, Frossard P, Sim X, Hassanali N, Radha V, Chidambaram M, Liju S, Rees SD, Ng DP-K, Wong T-Y, Yamauchi T, Hara K, Tanaka Y, Hirose H, McCarthy MI, Morris AP, DIAGRAM, MuTHER, AGEN, Basit A, Barnett AH, Katulanda P, et al.: **Genome-Wide Association Study Identifies a Novel Locus Contributing to Type 2 Diabetes Susceptibility in Sikhs of Punjabi Origin From India**. *Diabetes* 2013, **62** (5):1746–1755.

150. Imamura M, Maeda S, Yamauchi T, Hara K, Yasuda K, Morizono T, Takahashi A, Horikoshi M, Nakamura M, Fujita H, Tsunoda T, Kubo M, Watada H, Maegawa H, Okada-Iwabu M, Iwabu M, Shojima N, Ohshige T, Omori S, Iwata M, Hirose H, Kaku K, Ito C, Tanaka Y, Tobe K, Kashiwagi A, Kawamori R, Kasuga M, Kamatani N, Consortium DGR and M (DIAGRAM), et al.: **A single-nucleotide polymorphism in ANK1 is associated with susceptibility to type 2 diabetes in Japanese populations**. *Hum Mol Genet* 2012, **21** (13):3042–3049.

151. Huang J, Ellinghaus D, Franke A, Howie B, Li Y: **1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data**. *Eur J Hum Genet* 2012, **20**:801–805.

152. Li H, Gan W, Lu L, Dong X, Han X, Hu C, Yang Z, Sun L, Bao W, Li P, He M, Sun L, Wang Y, Zhu J, Ning Q, Tang Y, Zhang R, Wen J, Wang D, Zhu X, Guo K, Zuo X, Guo X, Yang H, Zhou X, Consortium D, Consortium A-T, Zhang X, Qi L, Loos RJF, et al.: **A Genome-Wide Association Study Identifies GRK5 and RASGRP1 as Type 2 Diabetes Loci in Chinese Hans**. *Diabetes* 2013, **62** (1):291–298.

153. Perry JRB, Voight BF, Yengo L, Amin N, Dupuis J, Ganser M, Grallert H, Navarro P, Li M, Qi L, Steinthorsdottir V, Scott RA, Almgren P, Arking DE, Aulchenko Y, Balkau B, Benediktsson R, Bergman RN, Boerwinkle E, Bonnycastle L, Burtt NP, Campbell H, Charpentier G, Collins FS, Gieger C, Green T, Hadjadj S, Hattersley AT, Herder C, Hofman A, et al.: **Stratifying Type 2 Diabetes Cases by BMI Identifies Genetic Risk Variants in <italic>LAMA1</italic> and Enrichment for Risk Variants in Lean Compared to Obese Cases**. *PLoS Genet* 2012, **8**:e1002741.

154. Rasmussen-Torvik LJ, Guo X, Bowden DW, Bertoni AG, Sale MM, Yao J, Bluemke DA, Goodarzi MO, Chen YI, Vaidya D, Raffel LJ, Papanicolaou GJ, Meigs JB, Pankow JS: **Fasting Glucose GWAS Candidate Region Analysis Across Ethnic Groups in the Multiethnic Study of Atherosclerosis (MESA)**. *Genet Epidemiol* 2012, **36**:384–391.

155. Chen G, Bentley A, Adeyemo A, Shriner D, Zhou J, Doumatey A, Huang H, Ramos E, Erdos M, Gerry N, Herbert A, Christman M, Rotimi C: **Genome-wide association study identifies novel loci association with fasting insulin and insulin resistance in African Americans**. *Hum Mol Genet* 2012, **21** (20):4530–4536.

156. Palmer ND, McDonough CW, Hicks PJ, Roh BH, Wing MR, An SS, Hester JM, Cooke JN, Bostrom MA, Rudock ME, Talbert ME, Lewis JP, Ferrara A, Lu L, Ziegler JT, Sale MM, Divers J, Shriner D, Adeyemo A, Rotimi CN, Ng MCY, Langefeld CD, Freedman BI, Bowden DW, Consortium D, Investigators M: **A Genome-Wide Association Search for Type 2 Diabetes Genes in African Americans**. *PLoS One* 2012, **7**:e29202.

157. Tabassum R, Chauhan G, Dwivedi OP, Mahajan A, Jaiswal A, Kaur I, Bandesh K, Singh T, Mathai BJ, Pandey Y, Chidambaram M, Sharma A, Chavali S, Sengupta S, Ramakrishnan L, Venkatesh P, Aggarwal SK, Ghosh S, Prabhakaran D, Srinath RK, Saxena M, Banerjee M, Mathur S, Bhansali A, Shah VN, Madhu SV, Marwaha RK, Basu A, Scaria V, McCarthy MI, et al.: **Genome-Wide Association Study for Type 2 Diabetes in Indians Identifies a New Susceptibility Locus at 2q21**. *Diabetes* 2013, **62** (3):977–986.

158. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, Rybin D, Petrie JR, Travers ME, Bouatia-Naji N, Dimas AS, Nica A, Wheeler E, Chen H, Voight BF, Taneera J, Kanoni S, Peden JF, Turrini F, Gustafsson S, Zabena C, Almgren P, Barker DJP, Barnes D, Dennison EM, Eriksson JG, Eriksson P, Eury E, Folkersen L, Fox CS, Frayling TM, et al.: **Genome-Wide Association Identifies Nine Common Variants Associated With Fasting Proinsulin Levels and Provides New Insights Into the Pathophysiology of Type 2 Diabetes**. *Diabetes* 2011, **60** (10):2624–2634.

159. Irvin MR, Wineinger NE, Rice TK, Pajewski NM, Kabagambe EK, Gu CC, Pankow J, North KE, Wilk JB, Freedman BI, Franceschini N, Broeckel U, Tiwari HK, Arnett DK: **Genome-Wide Detection of Allele Specific Copy Number Variation Associated with Insulin Resistance in African Americans from the HyperGEN Study**. *PLoS One* 2011, **6**:e24052.

160. Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang W, Frossard P, Been LF, Chia K-S, Dimas AS, Hassanali N, Jafar T, Jowett JBM, Li X, Radha V, Rees SD, Takeuchi F, Young R, Aung T, Basit A, Chidambaram M, Das D, Grundberg E, Hedman AK, Hydrie ZI, Islam M, Khor C-C, Kowlessur S, Kristensen MM, Liju S, Lim W-Y, et al.: **Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci**. *Nat Genet* 2011, **43**:984–989.

161. Sim X, Ong RT-H, Suo C, Tay W-T, Liu J, Ng DP-K, Boehnke M, Chia K-S, Wong T-Y, Seielstad M, Teo Y-Y, Tai E-S: **Transferability of Type 2 Diabetes Implicated Loci in Multi-Ethnic Cohorts from Southeast Asia**. *PLoS Genet* 2011, **7**:e1001363.

162. Parra EJ, Below JE, Krithika S, Valladares A, Barta JL, Cox NJ, Hanis CL, Wacher N, Garcia-Mena J, Hu P, Shriver MD, Kumate J, McKeigue PM, Escobedo J, Cruz M: **Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas**. *Diabetologia* 2011, **54**:2038–2046.

163. Cui B, Zhu X, Xu M, Guo T, Zhu D, Chen G, Li X, Xu L, Bi Y, Chen Y, Xu Y, Li X, Wang W, Wang H, Huang W, Ning G: **A Genome-Wide Association Study Confirms Previously Reported Loci for Type 2 Diabetes in Han Chinese**. *PLoS One* 2011, **6**:e22353.

164. Cho YS, Chen C-H, Hu C, Long J, Hee Ong RT, Sim X, Takeuchi F, Wu Y, Go MJ, Yamauchi T, Chang Y-C, Kwak SH, Ma RCW, Yamamoto K, Adair LS, Aung T, Cai Q, Chang L-C, Chen Y-T, Gao Y, Hu FB, Kim H-L, Kim S, Kim YJ, Lee JJ-M, Lee NR, Li Y, Liu JJ, Lu W, Nakamura J, et al.: **Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians**. *Nat Genet* 2012, **44**:67–72.

165. Below JE, Gamazon ER, Morrison J V, Konkashbaev A, Pluzhnikov A, McKeigue PM, Parra EJ, Elbein SC, Hallman DM, Nicolae DL, Bell GI, Cruz M, Cox NJ, Hanis CL: **Genome-wide association and meta-analysis in populations from Starr County, Texas, and Mexico City identify type 2 diabetes susceptibility loci and enrichment for expression quantitative trait loci in top signals**. *Diabetologia* 2011, **54**:2047–2055.

166. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, Denny JC, Peissig PL, Miller AW, Wei W-Q, Bielinski SJ, Chute CG, Leibson CL, Jarvik GP, Crosslin DR, Carlson CS, Newton KM, Wolf WA, Chisholm RL, Lowe WL: **Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study**. *J Am Med Informatics Assoc* 2012, **19** (2):212–218.

167. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segre A V, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, et al.: **Twelve type 2 diabetes susceptibility loci**

identified through large-scale association analysis. *Nat Genet* 2010, **42**:579–589.

168. Yamauchi T, Hara K, Maeda S, Yasuda K, Takahashi A, Horikoshi M, Nakamura M, Fujita H, Grarup N, Cauchi S, Ng DPK, Ma RCW, Tsunoda T, Kubo M, Watada H, Maegawa H, Okada-Iwabu M, Iwabu M, Shojima N, Shin HD, Andersen G, Witte DR, Jorgensen T, Lauritzen T, Sandbaek A, Hansen T, Ohshige T, Omori S, Saito I, Kaku K, et al.: **A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B**. *Nat Genet* 2010, **42**:864–868.

169. Shu XO, Long J, Cai Q, Qi L, Xiang Y-B, Cho YS, Tai ES, Li X, Lin X, Chow W-H, Go MJ, Seielstad M, Bao W, Li H, Cornelis MC, Yu K, Wen W, Shi J, Han B-G, Sim XL, Liu L, Qi Q, Kim H-L, Ng DPK, Lee J-Y, Kim YJ, Li C, Gao Y-T, Zheng W, Hu FB: **Identification of New Genetic Risk Variants for Type 2 Diabetes**. *PLoS Genet* 2010, **6**:e1001127.

170. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, Lindgren CM, Magi R, Morris AP, Randall J, Johnson T, Elliott P, Rybin D, Thorleifsson G, Steinthorsdottir V, Henneman P, Grallert H, Dehghan A, Hottenga JJ, Franklin CS, Navarro P, Song K, Goel A, Perry JRB, Egan JM, Lajunen T, et al.: **New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk**. *Nat Genet* 2010, **42**:105–116.

171. Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, Pankow JS, Dupuis J, Florez JC, Fox CS, Paré G, Sun Q, Girman CJ, Laurie CC, Mirel DB, Manolio TA, Chasman DI, Boerwinkle E, Ridker PM, Hunter DJ, Meigs JB, Lee C-H, (MAGIC) M-A of G and I traits C, Consortium DGR and M (DIAGRAM), van Dam RM, Hu FB: **Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes**. *Hum Mol Genet* 2010, **19** (13):2706–2715.

172. Tsai F-J, Yang C-F, Chen C-C, Chuang L-M, Lu C-H, Chang C-T, Wang T-Y, Chen R-H, Shiu C-F, Liu Y-M, Chang C-C, Chen P, Chen C-H, Fann CSJ, Chen Y-T, Wu J-Y: **A Genome-Wide Association Study Identifies Susceptibility Variants for Type 2 Diabetes in Han Chinese**. *PLoS Genet* 2010, **6**:e1000847.

173. Takeuchi F, Serizawa M, Yamamoto K, Fujisawa T, Nakashima E, Ohnaka K, Ikegami H, Sugiyama T, Katsuya T, Miyagishi M, Nakashima N, Nawata H, Nakamura J, Kono S, Takayanagi R, Kato N: **Confirmation of Multiple Risk Loci and Genetic Impacts by a Genome-Wide Association Study of Type 2 Diabetes in the Japanese Population**. *Diabetes* 2009, **58** (7):1690–1699.

174. Rung J, Cauchi S, Albrechtsen A, Shen L, Rocheleau G, Cavalcanti-Proenca C, Bacot F, Balkau B, Belisle A, Borch-Johnsen K, Charpentier G, Dina C, Durand E, Elliott P, Hadjadj S, Jarvelin M-R, Laitinen J, Lauritzen T, Marre M, Mazur A, Meyre D, Montpetit A, Pisinger C, Posner B, Poulsen P, Pouta A, Prentki M, Ribel-Madsen R, Ruokonen A, Sandbaek A, et al.: **Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia**. *Nat Genet* 2009, **41**:1110–1115.

175. Chambers JC, Zhang W, Zabaneh D, Sehmi J, Jain P, McCarthy MI, Froguel P, Ruokonen A, Balding D, Jarvelin M-R, Scott J, Elliott P, Kooner JS: **Common Genetic Variation Near Melatonin Receptor MTNR1B Contributes to Raised Plasma Glucose and Increased Risk of Type 2 Diabetes Among Indian Asians and European Caucasians**. *Diabetes* 2009, **58** (11):2703–2708.

176. Rich SS, Goodarzi MO, Palmer ND, Langefeld CD, Ziegler J, Haffner SM, Bryer-Ash M, Norris JM, Taylor KD, Haritunians T, Rotter JI, Chen Y-DI, Wagenknecht LE, Bowden DW, Bergman RN: **A genome-wide association scan for acute insulin response to glucose in Hispanic-Americans: the Insulin Resistance Atherosclerosis Family Study (IRAS FS)**. *Diabetologia* 2009, **52**:1326–1333.

177. Palmer ND, Langefeld CD, Ziegler JT, Hsu F, Haffner SM, Fingerlin T, Norris JM, Chen YI, Rich SS, Haritunians T, Taylor KD, Bergman RN, Rotter JI, Bowden DW: **Candidate loci for insulin sensitivity and disposition index from a genome-wide association analysis of Hispanic participants in the Insulin Resistance Atherosclerosis (IRAS) Family Study**. *Diabetologia* 2010, **53**:281–289.

178. Timpson NJ, Lindgren CM, Weedon MN, Randall J, Ouwehand WH, Strachan DP, Rayner NW, Walker M, Hitman GA, Doney ASF, Palmer CNA, Morris AD, Hattersley AT, Zeggini E, Frayling TM, McCarthy MI: **Adiposity-Related Heterogeneity in Patterns of Type 2 Diabetes Susceptibility Observed in Genome-Wide Association Data**. *Diabetes* 2009, **58** (2):505–510.

179. Bouatia-Naji N, Rocheleau G, Van Lommel L, Lemaire K, Schuit F, Cavalcanti-Proença C, Marchand M, Hartikainen A-L, Sovio U, De Graeve F, Rung J, Vaxillaire M, Tichet J, Marre M, Balkau B, Weill J, Elliott P, Jarvelin M-R, Meyre D, Polychronakos C, Dina C, Sladek R, Froguel P: **A Polymorphism Within the G6PC2 Gene Is Associated with Fasting Plasma Glucose Levels**. *Sci* 2008, **320** (5879):1085–1088.

180. Chen W-M, Erdos MR, Jackson AU, Saxena R, Sanna S, Silver KD, Timpson NJ, Hansen T, Orrù M, Grazia Piras M, Bonnycastle LL, Willer CJ, Lyssenko V, Shen H, Kuusisto J, Ebrahim S, Sestu N, Duren WL, Spada MC, Stringham HM, Scott LJ, Olla N, Swift AJ, Najjar S, Mitchell BD, Lawlor DA, Smith GD, Ben-Shlomo Y, Andersen G, Borch-Johnsen K, et al.: **Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels** . *J Clin Invest* 2008, **118**:2620–2628.

181. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PIW, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding C-J, Doney ASF, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, et al.: **Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes**. *Nat Genet* 2008, **40**:638–645.

182. Prokopenko I, Langenberg C, Florez JC, Saxena R, Soranzo N, Thorleifsson G, Loos RJF, Manning AK, Jackson AU, Aulchenko Y, Potter SC, Erdos MR, Sanna S, Hottenga J-J, Wheeler E, Kaakinen M, Lyssenko V, Chen W-M, Ahmadi K, Beckmann JS, Bergman RN, Bochud M, Bonnycastle LL, Buchanan TA, Cao A, Cervino A, Coin L, Collins FS, Crisponi L, de Geus EJC, et al.: **Variants in MTNR1B influence fasting glucose levels**. *Nat Genet* 2009, **41**:77–81.

183. Bouatia-Naji N, Bonnefond A, Cavalcanti-Proenca C, Sparso T, Holmkvist J, Marchand M, Delplanque J, Lobbens S, Rocheleau G, Durand E, De Graeve F, Chevre J-C, Borch-Johnsen K, Hartikainen A-L, Ruokonen A, Tichet J, Marre M, Weill J, Heude B, Tauber M, Lemaire K, Schuit F, Elliott P, Jorgensen T, Charpentier G, Hadjadj S, Cauchi S, Vaxillaire M, Sladek R, Visvikis-Siest S, et al.: **A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk**. *Nat Genet* 2009, **41**:89–94.

184. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT and Novartis Institutes of BioMedical Research LU, Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PIW, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Boström K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Råstam L, Speliotes EK, Taskinen M-R, et al.: **Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels**. *Sci* 2007, **316** (5829):1331–1336.

185. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**:661–678.

186. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky A V, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P: **A genome-wide association study identifies novel risk loci for type 2 diabetes**. *Nature* 2007, **445**:881–885.

187. Meigs JB, Manning AK, Fox CS, Florez JC, Liu C, Cupples LA, Dupuis J: **Genome-wide association with diabetes-related traits in the Framingham Heart Study.** *BMC Med Genet* 2007, **8 Suppl 1**(Suppl 1):S16.

188. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JRB, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney ASF, (WTCCC) TWTCCC, McCarthy MI, Hattersley AT: **Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes**. *Sci* 2007, **316** (5829):1336–1341.

189. Hayes MG, Pluzhnikov A, Miyake K, Sun Y, Ng MCY, Roe CA, Below JE, Nicolae RI, Konkashbaev A, Bell GI, Cox NJ, Hanis CL: **Identification of Type 2 Diabetes Genes in Mexican Americans Through Genome-Wide Association Studies**. *Diabetes* 2007, **56** (12):3033–3044.

190. Hanson RL, Bogardus C, Duggan D, Kobes S, Knowlton M, Infante AM, Marovich L, Benitez D, Baier LJ, Knowler WC: **A Search for Variants Associated With Young-Onset Type 2 Diabetes in American Indians in a 100K Genotyping Array**. *Diabetes* 2007, **56** (12):3045–3052.

191. Rampersaud E, Damcott CM, Fu M, Shen H, McArdle P, Shi X, Shelton J, Yin J, Chang Y-PC, Ott SH, Zhang L, Zhao Y, Mitchell BD, O'Connell J, Shuldiner AR: **Identification of Novel Candidate Genes for Type 2 Diabetes From a Genome-Wide Association Scan in the Old Order Amish: Evidence for Replication From Diabetes-Related Quantitative Traits and From Independent Populations** . *Diabetes* 2007, **56** (12):3053–3062.

192. Florez JC, Manning AK, Dupuis J, McAteer J, Irenze K, Gianniny L, Mirel DB, Fox CS, Cupples LA, Meigs JB: **A 100K Genome-Wide Association Scan for Diabetes and Related Traits in the Framingham Heart Study: Replication and Integration With Other Genome-Wide Datasets** . *Diabetes* 2007, **56** (12):3063–3074.

193. Salonen JT, Uimari P, Aalto J-M, Pirskanen M, Kaikkonen J, Todorova B, Hyppönen J, Korhonen V-P, Asikainen J, Devine C, Tuomainen T-P, Luedemann J, Nauck M, Kerner W, Stephens RH, New JP, Ollier WE, Gibson JM, Payton A, Horan MA, Pendleton N, Mahoney W, Meyre D, Delplanque J, Froguel P, Luzzatto O, Yakir B, Darvasi A: **Type 2 diabetes whole-genome association study in four populations: the DiaGen consortium.** *Am J Hum Genet* 2007, **81**:338–45.

194. Patti ME, Butte AJ, Crunkhorn S, Cusi K, Berria R, Kashyap S, Miyazaki Y, Kohane I, Costello M, Saccone R, Landaker EJ, Goldfine AB, Mun E, DeFronzo R, Finlayson J, Kahn CR, Mandarino LJ: **Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1**. *Proc Natl Acad Sci* 2003, **100** (14):8466–8471.

195. Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC: **Gene Expression in Peripheral Blood Mononuclear Cells from Children with Diabetes**. *J Clin Endocrinol Metab* 2007, **92**:3705–3711.

196. Wu X, Wang J, Cui X, Maianu L, Rhees B, Rosinski J, So WV, Willi S, Osier M, Hill H, Page G, Allison D, Martin M, Garvey WT: **The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle**. *Endocrine* 2007, **31**:5–17.

197. Frederiksen CM, Højlund K, Hansen L, Oakeley EJ, Hemmings B, Abdallah BM, Brusgaard K, Beck-Nielsen H, Gaster M: **Transcriptional profiling of myotubes from patients with type 2 diabetes: no evidence for a primary defect in oxidative phosphorylation genes**. *Diabetologia* 2008, **51**:2068–2077.

198. Pihlajamäki J, Boes T, Kim E-Y, Dearie F, Kim BW, Schroeder J, Mun E, Nasser I, Park PJ, Bianco AC, Goldfine AB, Patti ME: **Thyroid Hormone-Related Regulation of Gene Expression in Human Fatty Liver**. *J Clin Endocrinol Metab* 2009, **94**:3521–3529.

199. Marselli L, Thorne J, Dahiya S, Sgroi DC, Sharma A, Bonner-Weir S, Marchetti P, Weir GC: **Gene Expression Profiles of Beta-Cell Enriched Tissue Obtained by Laser Capture Microdissection from Subjects with Type 2 Diabetes**. *PLoS One* 2010, **5**:e11499.

200. Gallagher I, Scheele C, Keller P, Nielsen A, Remenyi J, Fischer C, Roder K, Babraj J, Wahlestedt C, Hutvagner G, Pedersen B, Timmons J: **Integration of microRNA changes in vivo identifies novel molecular features of muscle insulin resistance in type 2 diabetes**. *Genome Med* 2010, **2**:1–18.

201. Misu H, Takamura T, Takayama H, Hayashi H, Matsuzawa-Nagata N, Kurita S, Ishikura K, Ando H, Takeshita Y, Ota T, Sakurai M, Yamashita T, Mizukoshi E, Yamashita T, Honda M, Miyamoto K, Kubota T, Kubota N, Kadowaki T, Kim H-J, Lee I, Minokoshi Y, Saito Y, Takahashi K, Yamada Y, Takakura N, Kaneko S: **A Liver-Derived Secretory Protein, Selenoprotein P, Causes Insulin Resistance**. *Cell Metab* 2014, **12**:483–495.

202. Dominguez V, Raimondi C, Somanath S, Bugliani M, Loder MK, Edling CE, Divecha N, da Silva-Xavier G, Marselli L, Persaud SJ, Turner MD, Rutter GA, Marchetti P, Falasca M, Maffucci T: **Class II Phosphoinositide 3-Kinase Regulates Exocytosis of Insulin Granules in Pancreatic β Cells**. *J Biol Chem* 2011, **286** (6):4216–4225.

203. Jin W, Goldfine AB, Boes T, Henry RR, Ciaraldi TP, Kim E-Y, Emecan M, Fitzpatrick C, Sen A, Shah A, Mun E, Vokes M, Schroeder J, Tatro E, Jimenez-Chillaron J, Patti M-E: **Increased SRF transcriptional activity in human and mouse skeletal muscle is a signature of insulin resistance**. *J Clin Invest* 2011, **121**:918–929.

204. Storey JD: **The positive false discovery rate: a Bayesian interpretation and the q-value**. 2003:2013–2035.

205. Weir GC, Marselli L, Marchetti P, Katsuta H, Jung MH, Bonner-Weir S: **Towards better understanding of the contributions of overwork and glucotoxicity to the β-cell inadequacy of type 2 diabetes**. *Diabetes, Obes Metab* 2009, **11**:82–90.

206. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci* 2001, **98** (9):5116–5121.

207. Kuwahara K, Barrientos T, Pipes GCT, Li S, Olson EN: **Muscle-Specific Signaling Mechanism That Links Actin Dynamics to Serum Response Factor**. *Mol Cell Biol* 2005, **25** (8):3173–3181.

208. Chen R, Corona E, Sikora M, Dudley JT, Morgan AA, Moreno-Estrada A, Nilsen GB, Ruau D, Lincoln SE, Bustamante CD, Butte AJ: **Type 2 Diabetes Risk Alleles Demonstrate Extreme Directional Differentiation among Human Populations, Compared to Other Diseases**. *PLoS Genet* 2012, **8**:e1002621.

209. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis**. *Bioinforma* 2005, **21** (16):3439–3440.

210. Gibbons A: **Diabetes Genes Decline Out of Africa**. *Sci* 2011, **334** (6056):583.

211. Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A: **Adaptations to Climate-Mediated Selective Pressures in Humans**. *PLoS Genet* 2010, **7**:e1001375.

212. Balaresque PL, Ballereau SJ, Jobling MA: **Challenges in human genetic diversity: demographic history and adaptation**. *Hum Mol Genet* 2007, **16**:R134–R139.

213. Neel J: **Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress."***Am J Hum Genet* 1962, **14**:353–362.

214. Chakravarthy M V, Booth FW: **Eating, exercise, and "thrifty" genotypes: connecting the dots toward an evolutionary understanding of modern chronic diseases**. *J Appl Physiol* 2004, **96** (1):3–10.

215. Ling C, Groop L: **Epigenetics: A Molecular Link Between Environmental Factors and Type 2 Diabetes**. *Diabetes* 2009, **58** (12):2718–2725.

216. Gorlov IP, Gallick GE, Gorlova OY, Amos C, Logothetis CJ: **GWAS Meets Microarray: Are the Results of Genome-Wide Association Studies and Gene-Expression Profiling Consistent? Prostate Cancer as an Example**. *PLoS One* 2009, **4**:e6511.

217. Fedor MJ: **Alternative Splicing Minireview Series: Combinatorial Control Facilitates Splicing Regulation of Gene Expression and Enhances Genome Diversity**. *J Biol Chem* 2008, **283** (3):1209–1210.

218. Perseghin G, Caumo A, Arcelloni C, Benedini S, Lanzi R, Pagliato E, Sereni LP, Testolin G, Battezzati A, Comi G, Comola M, Luzi L: **Contribution of Abnormal Insulin Secretion and Insulin Resistance to the Pathogenesis of Type 2 Diabetes in Myotonic Dystrophy**. *Diabetes Care* 2003, **26** (7):2112–2118.

219. The International HapMap Consortium: **A haplotype map of the human genome**. *Nature* 2005, **437**:1299–1320.

220. Hofer T, Foll M, Excoffier L: **Evolutionary forces shaping genomic islands of population differentiation in humans**. *BMC Genomics* 2012, **13**:107.

221. Willing E-M, Dreyer C, van Oosterhout C: **Estimates of Genetic Differentiation Measured by F_{ST} Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers**. *PLoS One* 2012, **7**:e42649.

222. Wu JK, Kitajewski JK: **A Potential Role for Notch Signaling in the Pathogenesis and Regulation of Hemangiomas**. *J Craniofac Surg* 2009, **20**.

223. Apelqvist A, Li H, Sommer L, Beatus P, Anderson DJ, Honjo T, de Angelis MH, Lendahl U, Edlund H: **Notch signalling controls pancreatic cell differentiation**. *Nature* 1999, **400**:877–881.

224. Shih HP, Kopp JL, Sandhu M, Dubois CL, Seymour PA, Grapin-Botton A, Sander M: **A Notch-dependent molecular circuitry initiates pancreatic endocrine and ductal cell differentiation**. *Dev* 2012.

225. Wittmann T, Hyman A, Desai A: **The spindle: a dynamic assembly of microtubules and motors**. *Nat Cell Biol* 2001, **3**:E28–E34.

226. Furukawa Y, Shimada T, Furuta H, Matsuno S, Kusuyama A, Doi A, Nishi M, Sasaki H, Sanke T, Nanjo K: **Polymorphisms in the IDE-KIF11-HHEX Gene Locus Are Reproducibly Associated with Type 2 Diabetes in a Japanese Population**. *J Clin Endocrinol Metab* 2008, **93**:310–314.

227. Hu C, Zhang R, Wang C, Wang J, Ma X, Lu J, Qin W, Hou X, Wang C, Bao Y, Xiang K, Jia W: **PPARG, KCNJ11, CDKAL1, CDKN2A-CDKN2B, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 Are Associated with Type 2 Diabetes in a Chinese Population**. *PLoS One* 2009, **4**:e7643.

228. Winkler C, Bonifacio E, Grallert H, Henneberger L, Illig T, Ziegler AG: **BMI at age 8 years is influenced by the type 2 diabetes susceptibility genes HHEX-IDE and CDKAL1.** *Diabetes* 2010, **59**:2063–2067.

229. Biankin A V, Waddell N, Kassahn KS, Gingras M-C, Muthuswamy LB, Johns AL, Miller DK, Wilson PJ, Patch A-M, Wu J, Chang DK, Cowley MJ, Gardiner BB, Song S, Harliwong I, Idrisoglu S, Nourse C, Nourbakhsh E, Manning S, Wani S, Gongora M, Pajic M, Scarlett CJ, Gill AJ, Pinho A V, Rooman I, Anderson M, Holmes O, Leonard C, Taylor D, et al.: **Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes**. *Nature* 2012, **491**:399–405.

230. Zhao J, Deliard S, Aziz A, Grant S: **Expression analyses of the genes harbored by the type 2 diabetes and pediatric BMI associated locus on 10q23**. *BMC Med Genet* 2012, **13**:1–5.

231. Rebhan M: **GeneCards: integrating information about genes, proteins and diseases**. *Trends Genet* 1997, **13**:163.

232. Boutant M, Ramos OHP, Tourrel-Cuzin C, Movassat J, Ilias A, Vallois D, Planchais J, Pégorier J-P, Schuit F, Petit PX, Bossard P, Maedler K, Grapin-Botton A, Vasseur-Cognet M: **COUP-TFII Controls Mouse Pancreatic β-Cell Mass through GLP-1-β-Catenin Signaling Pathways**. *PLoS One* 2012, **7**:e30847.

233. Okamura M, Kudo H, Wakabayashi K, Tanaka T, Nonaka A, Uchida A, Tsutsumi S, Sakakibara I, Naito M, Osborne TF, Hamakubo T, Ito S, Aburatani H, Yanagisawa M, Kodama T, Sakai J: **COUP-TFII Acts Downstream of Wnt/β-**

Catenin Signal to Silence PPARγ Gene Expression and Repress Adipogenesis. *Proc Natl Acad Sci U S A* 2009, **106**:5819–5824.

234. Qin L, Chen Y, Niu Y, Chen W, Wang Q, Xiao S, Li A, Xie Y, Li J, Zhao X, He Z, Mo D: **A deep investigation into the adipogenesis mechanism: profile of microRNAs regulating adipogenesis by modulating the canonical Wnt/beta-catenin signaling pathway.** *BMC Genomics* 2010, **11**:320.

235. Lowe CE, O'Rahilly S, Rochford JJ: **Adipogenesis at a glance**. *J Cell Sci* 2011, **124** (16):2681–2686.

236. Xu Z, Yu S, Hsu C-H, Eguchi J, Rosen ED: **The orphan nuclear receptor chicken ovalbumin upstream promoter-transcription factor II is a critical regulator of adipogenesis**. *Proc Natl Acad Sci* 2008, **105** (7):2421–2426.

237. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P: **The Reactome pathway knowledgebase**. *Nucleic Acids Res* 2013.

238. Nakae J, Biggs WH, Kitamura T, Cavenee WK, Wright CVE, Arden KC, Accili D: **Regulation of insulin action and pancreatic [beta]-cell function by mutated alleles of the gene encoding forkhead transcription factor Foxo1**. *Nat Genet* 2002, **32**:245–253.

239. Desbois-Mouthon C: **Insulin and IGF-1 stimulate the bold beta-catenin pathway through two signalling cascades involving GSK-3bold beta inhibition and Ras activation**. *Nat Onc* 2001, **20**:252–259.

240. Feldstein O, Ben-Hamo R, Bashari D, Efroni S, Ginsberg D: **RBM38 Is a Direct Transcriptional Target of E2F1 that Limits E2F1-Induced Proliferation**. *Mol Cancer Res* 2012, **10** (9):1169–1177.

241. Dayeh T, Volkov P, Salö S, Hall E, Nilsson E, Olsson AH, Kirkpatrick CL, Wollheim CB, Eliasson L, Rönn T, Bacos K, Ling C: **Genome-Wide DNA Methylation Analysis of Human Pancreatic Islets from Type 2 Diabetic and Non-Diabetic Donors Identifies Candidate Genes That Influence Insulin Secretion**. *PLoS Genet* 2014, **10**:e1004160.

242. Qian Y, Lu F, Dong M, Lin Y, Li H, Chen J, Shen C, Jin G, Hu Z, Shen H: **Genetic Variants of <italic>IDE-KIF11-HHEX</italic> at 10q23.33 Associated with Type 2 Diabetes Risk: A Fine-Mapping Study in Chinese Population**. *PLoS One* 2012, **7**:e35060.

243. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM: **The Genetic Structure and History of Africans and African Americans**. *Science (80-)* 2009, **324**:1035–1044.

244. Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Guldemann T, Kure B, Mpoloka SW, Nakagawa H, Naumann C, Lipson M, Loh P-R, Lachance J, Mountain J, Bustamante CD, Berger B, Tishkoff SA, Henn BM, Stoneking M, Reich D, Pakendorf B: **The genetic prehistory of southern Africa**. *Nat Commun* 2012, **3**:1143.

245. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation.** *Genome Biol* 2006, **7 Suppl 1**:S12.1–14.

246. Li J, Hawkins IC, Harvey CD, Jennings JL, Link AJ, Patton JG: **Regulation of Alternative Splicing by SRrp86 and Its Interacting Proteins**. *Mol Cell Biol* 2003, **23** (21):7437–7447.

247. Henn BM, Cavalli-Sforza LL, Feldman MW: **The great human expansion**. *Proc Natl Acad Sci* 2012, **109**:17758–17764.

248. Khrameeva EE, Bozek K, He L, Yan Z, Jiang X, Wei Y, Tang K, Gelfand MS, Prufer K, Kelso J, Paabo S, Giavalisco P, Lachmann M, Khaitovich P: **Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans**. *Nat Commun* 2014, **5**.

249. Chen M, Manley JL: **Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches**. *Nat Rev Mol Cell Biol* 2009, **10**:741–754.

250. Shalev A, Blair PJ, Hoffmann SC, Hirshberg B, Peculis BA, Harlan DM: **A Proinsulin Gene Splice Variant with Increased Translation Efficiency Is Expressed in Human Pancreatic Islets**. *Endocrinology* 2002, **143**:2541–2547.

251. Kaminska D, Pihlajamäki J: **Regulation of alternative splicing in obesity and weight loss**. *Adipocyte* 2013, **2**:143–147.

252. Kulseth MA, Berge KE, Bogsrud MP, Leren TP: **Analysis of LDLR mRNA in patients with familial hypercholesterolemia revealed a novel mutation in intron 14, which activates a cryptic splice site**. *J Hum Genet* 2010, **55**:676–680.

253. Péterfy M, Phan J, Reue K: **Alternatively Spliced Lipin Isoforms Exhibit Distinct Expression Pattern, Subcellular Localization, and Role in Adipogenesis**. *J Biol Chem* 2005, **280** (38):32883–32889.

254. Belfiore A, Frasca F, Pandini G, Sciacca L, Vigneri R: **Insulin Receptor Isoforms and Insulin Receptor/Insulin-Like Growth Factor Receptor Hybrids in Physiology and Disease**. *Endocr Rev* 2009, **30**:586–623.

255. Pihlajamäki J, Lerin C, Itkonen P, Boes T, Floss T, Schroeder J, Dearie F, Crunkhorn S, Burak F, Jimenez-Chillaron JC, Kuulasmaa T, Miettinen P, Park PJ, Nasser I, Zhao Z, Zhang Z, Xu Y, Wurst W, Ren H, Morris AJ, Stamm S, Goldfine AB, Laakso M, Patti ME: **Expression of the Splicing Factor Gene SFRS10 Is Reduced in Human Obesity and Contributes to Enhanced Lipogenesis**. *Cell Metab* 2011, **14**:208–218.

256. Hansson O, Zhou Y, Renström E, Osmark P: **Molecular Function of TCF7L2: Consequences of TCF7L2 Splicing for Molecular Function and Risk for Type 2 Diabetes**. *Curr Diab Rep* 2010, **10**:444–451.

257. Le Bacquer O, Shu L, Marchand M, Neve B, Paroni F, Kerr Conte J, Pattou F, Froguel P, Maedler K: **TCF7L2 splice variants have distinct effects on β-cell turnover and function**. *Hum Mol Genet* 2011, **20** (10):1906–1915.

258. Kaminska D, Kuulasmaa T, Venesmaa S, Käkelä P, Vaittinen M, Pulkkinen L, Pääkkönen M, Gylling H, Laakso M, Pihlajamäki J: **Adipose Tissue TCF7L2 Splicing Is Regulated by Weight Loss and Associates With Glucose and Fatty Acid Metabolism**. *Diabetes* 2012, **61** (11):2807–2813.

259. Morgan MJ: **A Brief (If Insular) History of the Human Genome Project**. *PLoS Biol* 2011, **9**:e1000601.

260. Hu FB: **Globalization of Diabetes: The role of diet, lifestyle, and genes** . *Diabetes Care* 2011, **34** (6):1249–1257.

APPENDICES

The electronic appendices can be found at http://www.bioinf.wits.ac.za/data/oduaran/

APPENDIX I – GWAS Genes and SNPs Lists

GWAS Genes List

MARCH1	CRY2	HUNK	PCBP3	SLC30A8
ABCB11	CSMD1	IDE	PCK1	SLC44A3
ACHE	CTCFL	IG2BP2	PCNXL2	SND1
ADAM30	DCD	IGF1	PCSK1	SNX7
ADAMTS9	DGKB	IGF2BP2	PDGFC	SPRY2
ADCY5	DPYSL5	IRS1	PDX1	SRR
ADRA2A	DUSP9	ITGB6	PEPD	ST6GAL1
AKAP2	F3	JAZF1	PEX5L	SULF1
AKAP6	FADS1	KCNJ11	PLS1	SYK
ANK1	FAM58A	KCNK16	PPARG	SYN2
AP3B1	FBXL10	KCNQ1	PPP1R3B	TCERG1L
AP3S2	FITM2	KCNU1	PPP2R2C	TCERGIL
ARAP1	FLJ16165	KIF11	PRC1	TCF2
ARF5	FOXA2	KLF14	PRDM5	TCF7L2
ATP10A	FOXN3	LAMA1	PROX1	TGFBR3
BARX2	FTO	LARP6	PSMD6	THADA
BCL11A	G6PC2	LGR5	PTCHD3	TMEM163
C14orf70	GALNTL4	LMO1	PTPRD	TMEM195
C2CD4A	GAS1	LOC64673	PTTG1	TMEM45B
C2CD4B	GCC1	LOC72901	R3HDML	TP53INP1
C6orf57	GCK	LOC729013	RASGRP1	TRIAP1
CACNA1D	GCKR	LPIN2	RBM38	TSPAN8
CAMK1D	GLIS3	LYPLAL1	RBM43	TUBA3C
CDC123	GRB10	MADD	RBMS1	UHRF1BP1
CDKAL	GRB14	MAEA	RHOU	VEGFA
CDKAL1	GRHL3	MAP3K1	RND3	VPS13C
CDKN2A	GRK5	MGC21675	RREB1	VPS26A
CDKN2B	HHEX	MRPL33	SACS	WFS1
CENTD2	HLA-DQA2	MTNR1B	SC4MOL	WISP1
CETN3	HMG1L1	NOTCH2	SEZ6L	WWOX
CHCHD9	HMG20A	NXN	SFMBT2	ZBED3
CHL1	HMGA2	OR4S1	SGCG	ZFAND3
CMIP	HNF1A	PALM2	SGSM2	ZFAND6
COBLL1	HNF1B	PAX4	SLC10A6	ZMAT4
CR2	HNF4A	PCBP3	SLC2A2	ZPLD1

GWAS SNPs List

rs10050311	rs2383208	rs7903146	rs5219	rs163182	rs10741243	rs4689388
rs10229583	rs2501677	rs7944584	rs1111875	rs11642841	rs472265	rs2943641
rs10248619	rs2714337	rs7981942	rs13266634	rs849134	rs3773506	rs6235
rs10461617	rs2785980	rs8004664	rs9300039	rs231362	rs11677370	rs4502156
rs10811661	rs2943634	rs8050136	rs7756992	rs243021	rs7630877	rs4790333
rs10814916	rs308971	rs8090011	rs12304921	rs13292136	rs17045328	rs9727115
rs10829848	rs328506	rs8182584	rs7659604	rs7578326	rs1048886	rs10838687
rs10830963	rs340874	rs9552416	rs358806	rs4760790	rs7593730	rs10501320
rs10849893	rs35747	rs9552911	rs9465871	rs11634397	rs4712523	rs1549318
rs10885122	rs3736594	rs9841287	rs1495377	rs1552224	rs515071	
rs10886471	rs3802177	rs9939609	rs4607103	rs896854	rs1327796	
rs11041816	rs3916765	rs4506565	rs9472138	rs972283	rs10993738	
rs11165354	rs4402960	rs35767	rs7020996	rs10965250	rs7656416	
rs11257655	rs4430796	rs11071657	rs5215	rs10440833	rs4712524	
rs11558471	rs4527850	rs563694	rs6931514	rs13081389	rs6769511	
rs11603334	rs4607517	rs180730	rs17036101	rs1531343	rs2237897	
rs11605924	rs4646949	rs10510634	rs1153188	rs7957197	rs17584499	
rs11708067	rs4691380	rs7731657	rs10923931	rs1801214	rs391300	
rs11920090	rs4841132	rs2722425	rs7578597	rs4457053	rs2237895	
rs12010175	rs5015480	rs2166706	rs7961581	rs8042680	rs7018475	
rs13179048	rs560887	rs1387153	rs12779790	rs649891	rs2722769	
rs13273088	rs5945326	rs7043482	rs864745	rs1333051	rs7107217	
rs1334893	rs6048205	rs17589516	rs7901695	rs7305618	rs7560163	
rs1371614	rs623323	rs9792548	rs564398	rs730570	rs7542900	
rs1470579	rs6426514	rs1401492	rs10946398	rs1374910	rs6467136	
rs1483121	rs6670533	rs16962638	rs1436955	rs7119	rs6815464	
rs17046216	rs6723108	rs6576507	rs1359790	rs2833610	rs1535500	
rs17053082	rs7034200	rs2407314	rs10906115	rs2063640	rs17797882	

rs174550	rs7173964	rs12655917	rs3923113	rs6583826	rs9470794
rs1801282	rs7178572	rs4819143	rs16861329	rs9295474	rs3786897
rs1895320	rs7403531	rs17431357	rs4812829	rs12027542	rs831571
rs2191349	rs7607980	rs591044	rs2028299	rs10460009	rs6017317
rs2237892	rs7754840	rs2407103	rs1802295	rs3792615	rs7041847
rs2293941	rs7766070	rs6712932	rs7172432	rs642858	rs16955379
rs2300835	rs780094	rs5219	rs1436953	rs7636	rs12518099

APPENDIX II – Expression Genes List

SEPT6	ARPC2	CASP1	CLTC	EMR1	GATAD1	HOXD1
SEPT9	ASH2L	CASP10	CMAHP	ENG	GATAD2A	HSP90B1
ABCC10	ATP5B	CASQ1	CNN2	ENSA	GCFC1	HSPC072
ABCF2	ATP5E	CAT	CNOT8	ENTPD3	GCNT1	HUWE1
ACOT1	ATP5G2	CBX4	COG2	EPOR	GFER	ID2
ACSL1	ATP6AP2	CC2D1A	COX5A	ERAP1	GIF	IDE
ACSL6	ATP6V1E1	CCDC15	CRIP1	ERCC4	GIT2	IDH3G
ADAM28	ATP8B1	CCDC85B	CRYM	ERGIC2	GK	IDS
ADIPOR2	ATXN2L	CCPG1	CSNK2A1	ERH	GLS	IDUA
ADNP2	AURKAIP1	CCR1	CSTF1	ERLIN1	GMFB	IGHA1
ADRB3	B4GALT5	CCR2	CTDSP2	EXOC7	GNAQ	IKZF1
AHCY	BBIP1	CCR4	CUL4A	EYA2	GNAS	IQCB1
AHSP	BLCAP	CD200	CYC1	F5	GOLGA1	IRF4
AKR7A2	BLM	CD7	CYP2B6	FABP5	GPHN	ITFG1
AKT2	BOP1	CDC123	DARS	FABP7	GPS2	ITPR1
AKT3	BRD2	CDC16	DARS2	FAM115A	GRB10	ITPR2
ALPPL2	BTF3	CDC40	DBT	FAM47E	GRSF1	ITSN2
ANAPC13	BTN1A1	CDC5L	DCAF15	FAR2	GSN	IVD
ANKFY1	BTN2A1	CDK14	DHX15	FBXO11	GTF3C2	JTB
ANKHD1	C11orf30	CDS2	DICER1	FBXO28	GZMM	JUN
ANP32A	C12orf47	CEACAM1	DLG1	FBXO9	HCFC1R1	KCNAB1
AP3B1	C14orf45	CEP350	DLX4	FCF1	HEATR1	KCNJ1
APC	C15orf44	CFDP1	EDEM3	FCGR2B	HEXA	KDM4A
APOE	C16orf42	CHN2	EDF1	FCGR2C	HIRA	KDM4C
APOL2	C19orf60	CHPT1	EEF1D	FHL1	HLA-DPA1	KHSRP
ARF5	C1orf9	CHTOP	EEF1E1	FKBP1A	HMOX2	KIAA0125
ARFIP1	C1QBP	CIRBP	EGR1	FLJ11292	HNRNPA0	KIDINS220
ARHGAP15	C8orf33	CIZ1	EIF1	FOXC1	HNRNPA3	KLC1
ARHGEF1	CACYBP	CLCN4	EIF1AX	FOXN3	HNRNPD	KLF9
ARHGEF12	CALM1	CLK1	EIF4A3	FRY	HNRNPK	KRT33A
ARHGEF7	CAPZB	CLTB	EIF4H	GADD45B	HNRNPU	LAMA4

LAMP1	MFN1	NF1	PEMT	PTPRC	RPL12	SNRPD1
LAT	MFNG	NFIB	PEX5	PTPRCAP	RPL14	SNRPE
LEPROT	MINK1	NOMO1	PFKM	PVRL3	RPL35A	SNX1
LIMK1	MIOS	NONO	PGK1	QKI	RPL7	SNX13
LMBRD1	MLX	NOP10	PGRMC2	RAB11FIP2	RTF1	SOCS5
LMF1	MMP14	NOS1	PHKG1	RAB14	RUFY3	SON
LPCAT4	MMP24	NPRL3	PHLPP2	RAB3GAP2	RWDD1	SORL1
LSM12	MON2	NR2F2	PHTF1	RAB6A	SAP130	SOX2
LSM14A	MPHOSPH6	NR3C1	PICALM	RABGGTA	SART1	SPO11
LSM2	MPPE1	NR4A1	PIK3R4	RAD17	SCAMP1	SPOP
LTBP4	MR1	NRG1	PML	RALGDS	SEC23B	SRCAP
LYST	MRP63	NSF	PMVK	RANBP2	SECISBP2L	SRR
LZTFL1	MRPL28	NTRK3	PNISR	RASGRP2	SEMA3F	SRRM1
MAFB	MRPL9	NUCB2	POLDIP3	RASSF7	SESN1	SRRM2
MAN1A2	MRPS10	NUP98	POLG	RBCK1	SETX	SRSF1
MAP4K1	MTDH	NUPL1	POLR2B	RBM3	SF1	SRSF9
MAPRE2	MTF1	OSBP	POM121L1P	RBM38	SF3A2	SSBP3
MAST2	MTOR	PAFAH1B1	PPIA	RBMY1A1	SFI1	ST6GALNAC4
MAU2	MTRR	PBX2	PPIB	RC3H2	SFPQ	STAU1
MAZ	MYL4	PCBP1	PPP1R2	RCC1	SFXN3	STK24
MBD3	MYO7A	PCBP2	PPPDE1	REST	SLC25A17	STRAP
MCFD2	MZT2B	PCDH9	PREPL	RGS14	SLC2A4RG	STX16
MCL1	NAA60	PCM1	PRKACA	RGS4	SLC35A2	SULT1B1
MDH2	NAMPT	PCNX	PRKAG1	RIN3	SLC39A4	SUMO1
MEAF6	NBPF10	PDCD4	PRKCSH	RNF126	SLC39A8	SUMO2
MED16	NCL	PDE10A	PRPF31	RNF34	SLC7A4	SUMO3
MED21	NCOA1	PDE4DIP	PSMA2	RNGTT	SLIRP	SUPT4H1
MED27	NCOR1	PDE8A	PSMA3	RNPS1	SMAD3	TAF10
MEF2A	NDRG2	PDPK1	PSMD1	ROCK2	SMARCA4	TAF1C
MEF2C	NDUFA5	PEBP1	PTBP1	RPH3A	SNF8	TANK
MFAP3	NDUFB8	PEF1	PTPN12	RPL11	SNRNP70	TAOK2

TAP2	TROVE2
TBC1D5	TTC12
TBK1	U2AF2
TCEB1	UBE2D2
TCERG1	UBR2
TCL6	UBR5
TERF2IP	UFM1
TF	USP34
TFRC	USP9X
TGFB1	UTP14C
TKT	VAMP1
TLR7	WDFY3
TMBIM6	WIPI2
TMEM30B	WNK1
TMF1	WSB1
TMOD1	XAB2
TNFRSF14	XRCC6
TOB1	YBX1
TOM1L1	YIPF6
TOMM20	YTHDC2
TOR1B	YWHAE
TOX3	YWHAZ
TOX4	YY1
TP63	ZC3HAV1
TPI1	ZCCHC6
TPM2	ZFP36L2
TPR	ZIC1
TRAPPC2	ZNF10
TRAPPC9	ZNF277
TRIM2	ZNF287
TRIM38	ZNF473
TRIM44	

APPENDIX III – Core “Genes List” (used for study analyses)

ADAM30	GCK	NOTCH2	SF1
ADAMTS9	GCKR	NR2F2	SF3A2
ANKHD1	GPS2	NRG1	SFPQ
ATP8B1	GRSF1	PAX4	SLC2A2
BLM	HEATR1	PCBP1	SLC30A8
BOP1	HHEX	PCBP2	SMAD3
CACNA1D	HNF1A	PCK1	SMARCA4
CAMK1D	HNF1B	PCSK1	SNRNP70
CBX4	HNF4A	PDCD4	SNRPD1
CCDC85B	HNRNPA0	PDX1	SNRPE
CDC123	HNRNPA3	PEBP1	SOX2
CDC16	HNRNPD	PML	SPOP
CDC40	HNRNPK	POLR2B	SRRM1
CDC5L	HNRNPU	PPARG	SRRM2
CDKAL1	ID2	PROX1	STRAP
CDKN2A	IDE	PRPF31	SUMO1
CDKN2B	IGF1	PSMA2	SUPT4H1
CRYM	IGF2BP2	PSMA3	SYN2
CSTF1	IKZF1	PSMD1	TCERG1
DCD	IRS1	PTBP1	TCF7L2
DHX15	JAZF1	PTPRC	TERF2IP
DICER1	JUN	QKI	TGFB1
EGR1	KCNJ11	RAD17	THADA
EIF4A3	KHSRP	RBM3	TP63
ENG	KIF11	RBM38	TRIAP1
ERCC4	LGR5	RBMY1A1	TSPAN8
F3	LSM2	REST	U2AF2
FADS1	MAEA	RNGTT	UTP14C
FCF1	MBD3	RNPS1	VEGFA
FKBP1A	MEF2C	RPL11	WFS1
FOXA2	MLX	RPL14	XAB2
FOXN3	MTDH	RPL35A	YBX1
FTO	NCOR1	RPL7	YWHAE
GAS1	NONO	SART1	ZCCHC6
GATAD2A	NOP10	SETX	ZFP36L2

APPENDIX IV – List of Genes in the Intercontinental Comparisons with $F_{ST} \geq 0.25$ (63 genes) and the final list of F_{ST} genes (7 genes) and SNPs (228 SNPs)

AFR_nonAFR – Genes ($F_{ST} \geq 0.25$)

Gene	SNP Count		Gene	SNP Count
ADAM30	1		JAZF1	1
ADAMTS9	18		KIF11	20
ANKHD1	3		LGR5	38
ATP8B1	12		MLX	1
BLM	6		NCOR1	1
BOP1	5		NOTCH2	25
CACNA1D	26		NR2F2	4
CAMK1D	49		NRG1	45
CDC16	4		PPARG	35
CDC40	3		PTBP1	2
CDKAL1	22		QKI	24
CDKN2A	1		RBM38	20
CDKN2B	1		RNGTT	146
CRYM	4		RNPS1	3
CSTF1	1		RPL35A	4
ENG	8		SART1	5
ERCC4	6		SF1	3
FCF1	2		SFPQ	1
FOXA2	1		SLC2A2	1
FOXN3	25		SLC30A8	23
FTO	37		SNRPD1	9
GATAD2A	1		STRAP	4
GRSF1	7		SYN2	26
HEATR1	15		TCF7L2	10
HNF1B	8		TERF2IP	9
HNRNPA3	1		THADA	22
HNRNPD	2		TP63	14
ID2	2		TSPAN8	11
IDE	1		U2AF2	1
IGF2BP2	27		WFS1	3
IKZF1	1		YWHAE	12
			ZFP36L2	1

AFR_nonAFR – SNPs and Genes (F_{ST} ≥ 0.25, p ≤ 0.05, overrepresentation score ≥ 2)

SNP	Gene		SNP	Gene
rs11907421	RBM38		rs201153006	RNGTT
rs2426714	RBM38		rs201553456	RNGTT
rs3764719	RBM38		rs2026019	RNGTT
rs3764722	RBM38		rs2064632	RNGTT
rs6014986	RBM38		rs2149459	RNGTT
rs6014987	RBM38		rs2149460	RNGTT
rs6014988	RBM38		rs2181024	RNGTT
rs6025526	RBM38		rs2610705	RNGTT
rs6025527	RBM38		rs2610706	RNGTT
rs6025528	RBM38		rs2610707	RNGTT
rs6025529	RBM38		rs2610709	RNGTT
rs6025530	RBM38		rs2610710	RNGTT
rs6025531	RBM38		rs2610712	RNGTT
rs6025532	RBM38		rs2610714	RNGTT
rs6025533	RBM38		rs2610716	RNGTT
rs6025537	RBM38		rs2610717	RNGTT
rs6025541	RBM38		rs2610718	RNGTT
rs6099605	RBM38		rs2610719	RNGTT
rs6128020	RBM38		rs2610720	RNGTT
rs7264925	RBM38		rs2610721	RNGTT
rs111904268	RPL35A		rs2610723	RNGTT
rs61215870	RPL35A		rs2610724	RNGTT
rs7631002	RPL35A		rs2610726	RNGTT
rs9818493	RPL35A		rs2610728	RNGTT
rs1011591	RNGTT		rs2610729	RNGTT
rs1321085	RNGTT		rs2610730	RNGTT
rs143448457	RNGTT		rs2610732	RNGTT
rs148728841	RNGTT		rs2610733	RNGTT
rs149726691	RNGTT		rs2610736	RNGTT
rs1590253	RNGTT		rs2610744	RNGTT
rs1928064	RNGTT		rs2610746	RNGTT
rs1928065	RNGTT		rs2610747	RNGTT
rs199533287	RNGTT		rs2610749	RNGTT
rs200127482	RNGTT		rs2610751	RNGTT
rs200217831	RNGTT		rs2610752	RNGTT
rs200515942	RNGTT		rs2610753	RNGTT

SNP	Gene		SNP	Gene
rs2610754	RNGTT		rs2756398	RNGTT
rs2610755	RNGTT		rs2756399	RNGTT
rs2610756	RNGTT		rs2756400	RNGTT
rs2610758	RNGTT		rs2756401	RNGTT
rs2610759	RNGTT		rs2756403	RNGTT
rs2610760	RNGTT		rs2756404	RNGTT
rs2610761	RNGTT		rs2756405	RNGTT
rs2610767	RNGTT		rs2756407	RNGTT
rs2756348	RNGTT		rs2756408	RNGTT
rs2756349	RNGTT		rs2756409	RNGTT
rs2756350	RNGTT		rs2756410	RNGTT
rs2756351	RNGTT		rs2756411	RNGTT
rs2756353	RNGTT		rs2756412	RNGTT
rs2756355	RNGTT		rs2756413	RNGTT
rs2756357	RNGTT		rs2756414	RNGTT
rs2756360	RNGTT		rs34693837	RNGTT
rs2756361	RNGTT		rs35095368	RNGTT
rs2756362	RNGTT		rs35792075	RNGTT
rs2756363	RNGTT		rs36063561	RNGTT
rs2756364	RNGTT		rs36112710	RNGTT
rs2756369	RNGTT		rs3839421	RNGTT
rs2756377	RNGTT		rs56277084	RNGTT
rs2756379	RNGTT		rs57988674	RNGTT
rs2756381	RNGTT		rs58588716	RNGTT
rs2756383	RNGTT		rs60138176	RNGTT
rs2756384	RNGTT		rs66869054	RNGTT
rs2756385	RNGTT		rs67781583	RNGTT
rs2756386	RNGTT		rs6902454	RNGTT
rs2756388	RNGTT		rs6915642	RNGTT
rs2756389	RNGTT		rs6919539	RNGTT
rs2756390	RNGTT		rs6919725	RNGTT
rs2756391	RNGTT		rs6920037	RNGTT
rs2756392	RNGTT		rs71554791	RNGTT
rs2756393	RNGTT		rs71661540	RNGTT
rs2756394	RNGTT		rs71681491	RNGTT
rs2756395	RNGTT		rs76042624	RNGTT
rs7739245	RNGTT		rs2493410	NOTCH2
rs7747388	RNGTT		rs2493416	NOTCH2
rs7750038	RNGTT		rs2493419	NOTCH2
rs7759650	RNGTT		rs2641316	NOTCH2
rs7760061	RNGTT		rs2793830	NOTCH2
rs7760588	RNGTT		rs327197	NOTCH2
rs7760910	RNGTT		rs4659250	NOTCH2

rs79967308	RNGTT		rs5025718	NOTCH2
rs911596	RNGTT		rs56216048	NOTCH2
rs9344843	RNGTT		rs6688004	NOTCH2
rs9344844	RNGTT		rs67062239	NOTCH2
rs9351171	RNGTT		rs699780	NOTCH2
rs9353583	RNGTT		rs7414396	NOTCH2
rs9353592	RNGTT		rs7530844	NOTCH2
rs9353633	RNGTT		rs835574	NOTCH2
rs9362538	RNGTT		rs11187085	KIF11
rs9362539	RNGTT		rs112853133	KIF11
rs9444639	RNGTT		rs11815573	KIF11
rs9444650	RNGTT		rs11817621	KIF11
rs9451046	RNGTT		rs12259474	KIF11
rs9451053	RNGTT		rs12261518	KIF11
rs9451062	RNGTT		rs12264712	KIF11
rs9451065	RNGTT		rs144328956	KIF11
rs9451067	RNGTT		rs56111269	KIF11
rs9451108	RNGTT		rs58045955	KIF11
rs975973	RNGTT		rs59877288	KIF11
rs10494235	NOTCH2		rs61011943	KIF11
rs10923926	NOTCH2		rs6583828	KIF11
rs10923929	NOTCH2		rs6583831	KIF11
rs1493694	NOTCH2		rs7069680	KIF11
rs1493695	NOTCH2		rs7070990	KIF11
rs2364166	NOTCH2		rs7079583	KIF11
rs2453042	NOTCH2		rs7079602	KIF11
rs2453044	NOTCH2		rs7089765	KIF11
rs2453055	NOTCH2		rs7914248	KIF11
rs2453056	NOTCH2		rs2847117	SNRPD1
rs2847139	SNRPD1			
rs2850556	SNRPD1			
rs2850558	SNRPD1			
rs2850568	SNRPD1			
rs2959525	SNRPD1			
rs2959527	SNRPD1			
rs3017641	SNRPD1			
rs34202260	SNRPD1			
rs142499350	NR2F2			
rs2398260	NR2F2			
rs73471288	NR2F2			
rs73471290	NR2F2			