Integration of Transcriptomics and Genome Wide Association Studies Data to Study the Biological Processes Involved in Type 2 Diabetes Mellitus

Junior practical training

Manuel Alejandro González Hernández

Id number: 6083141

Period January 26th to June 12th 2015

Department of Bioinformatics-BiGCaT
NUTRIM School of Nutrition and Translational Research in Metabolism
Maastricht University

Supervisors:
Martina Summer-Kutmon, PhD
Lars Eijssen, PhD
Elisa Cirillo, MSc
Table of contents

Abstract .................................................................................................................. 1
Introduction ........................................................................................................... 2
Materials and Methods ......................................................................................... 5
Results ................................................................................................................... 9
Discussion ............................................................................................................ 14
Conclusion and Limitations .................................................................................. 20
Future perspectives ............................................................................................... 21
Appendix ............................................................................................................... 22
References ............................................................................................................ 23
Abstract

Type 2 Diabetes Mellitus (T2DM) is a chronic disease characterized by hyperglycemia, caused by defects in beta cell function or insulin resistance, or both. T2DM is the most prevalent form of diabetes and according to the World Health Organization (WHO), it affects more than 550 million people worldwide. It is a multifactorial disease that involves several metabolic pathways and the exact pathophysiology still needs to be elucidated.

Aim: In the current study the objective was to re-analyze a transcriptomics dataset from hepatic biopsies of T2DM patients and integrate the results with Genome Wide Association Studies (GWAS) data. Hypothesis: We hypothesized that this integrative approach can strengthen understanding and provide insight in the mechanisms of T2DM.

Results: In the transcriptomics data, both pathway and Gene Ontology (GO) analyses reported pathways and biological processes associated to lipid and cholesterol metabolism. Moreover, five transcription factors (USF2, SP2, SREBF1, TRIM28 and KAT2A) of the significantly altered genes reported involvement of pathways related to T2DM. By integrating the GWAS data with the significantly altered pathways from the transcriptomics data, the statin pathway shows several genes with known T2DM related variations, like CETP, APOC1, APOB, LIPC, LDLR and LPL. Then, using GWAS3D to obtain the significant variants (P-value 10E-5), PPARG, KCNJ11 and TCF7L2 were reported among the loci with leading variants. Lastly, a pathway-based analysis with the Curated Human Collection from WikiPathways and the prioritized variants occurred in relevant pathways, such as statin pathway, Wnt signaling, T2DM pathway, SREBP signaling, Il-1 signaling, apoptosis and Toll-like receptor signaling pathway.

Conclusion: Analysis of existing transcriptomics and GWAS data aided to pinpoint relevant loci associated to T2DM. We can expect that further investigations with large-scale experiments add genetic risk factors involved. However, genetic testing should not distract from modifiable risk factors, such as diet, lifestyle and medication in the therapy of T2DM.
1. Introduction

Type 2 Diabetes Mellitus (T2DM) is a chronic disease, characterized by hyperglycemia, caused by defects in beta cell function or insulin resistance, or both (1). It is considered a multifactorial disease due to the interplay of environmental factors, lifestyle risk factors and genetic predisposition in its onset. T2DM is intricate and involves several metabolic pathways and the exact pathophysiology still needs to be elucidated (2, 3).

T2DM usually goes undiagnosed for years at early stages, due to the gradual progress and moderate insulin deficiency that delay the onset of the apparent symptoms. Diagnosis can be conducted with the following criteria: (i) HbA1C, a clinical marker of protein glycation reflecting glucose levels for a 2-3 month period, ≥6.5% certified with the NGSP (National Glycohemoglobin Standardization Program) in the USA (ii) Fasting Plasma Glucose (FPG) ≥ 7 mmol (126 mg/dl), (iii) Oral Glucose Tolerance Test (OGTT) with a 2 hour postprandial plasma glucose ≥ 11 mmol (200 mg/dl) or (iv) a random plasma glucose ≥ 11 mmol (200 mg/dl) in a patient with classic symptoms of hyperglycemic crisis (1, 4). Classical symptoms of T2DM consist of polyuria, polydipsia, weight loss and polyphagia. Long-term effects can lead to microvascular complications, such as retinopathy, nephropathy and neuropathy. Moreover, macrovascular complications may arise, such as coronary artery disease (CAD), renal failure (RF), myocardial infarction (MI), cerebrovascular disease and peripheral artery disease (PAD) possibly leading to amputation (5). T2DM is associated with different risk factors, which can be modifiable or non-modifiable. Modifiable risk factors include hypertension, abdominal obesity, lack of physical activity and dyslipidemia. Non-modifiable risk factors include age, sex, ethnicity, genetic predisposition and family history (6) (Figure 1).

T2DM is the most prevalent form of diabetes; it comprises 90 to 95 % of the people with this condition. In 2014, according to the World Health Organization, (WHO) the global prevalence was 9% in adults 18 years or older, affecting more than 550 million people worldwide (7, 8). Consequently, mortality related to T2DM has increased, mainly in low and middle-income nations, which have reported more than 80% of diabetes related deaths (9). Based on the state of affairs, WHO projects that diabetes will be the 7th leading cause of death in 2030 (10).
With respect to therapy and prevention of T2DM, current guidelines provide recommendations for diet and physical activity. Systematic reviews and meta-analyses have reported that aerobic exercise reduces HbA1c (glycated hemoglobin), and this has been further associated with a reduction in microvascular complications and cardiovascular (CDV) events. Moreover, resistance training ameliorates plasma glucose, lipids, blood pressure and cardiovascular risk. Based on that, both the American Heart Association and the European Society of Cardiology recommend 150 minutes of moderate intensity exercise distributed over at least 5 sessions a week (11, 12). Similarly, the Canadian Diabetes Association suggests that people with diabetes should accumulate a minimum of 150 minutes of moderate-vigorous intensity aerobic exercise each week, spread over at least 3 days of the week, with no more than 2 consecutive days without exercise (13).

In regard to Medical Nutrition Therapy the American Diabetes Association recommends an individualized diet and active engagement of the patient with the health care provider, claiming that a cooperative effort should facilitate the prevention of complications (14). According to the American Dietetic Association, a Registered Dietitian (RD) should be involved in the treatment from the first and follow-up encounters of T2DM patients. Ideally an RD should conduct a nutritional assessment, diagnosis, devise interventions, as well as being in charge of monitoring and evaluation to encourage lifestyle modification (15). Concerning drug therapy, metformin and/or insulin therapy are considered as standard treatments in patients with persistent fasting blood glucose >6 mmol/L, because hyperglycaemia reduction lowers HbA1c levels and microvascular complications (16). Other drugs such as the α-Glucosidase inhibitors, incretin agonists and Dipeptidyl peptidase-4 (DPP4) inhibitors have been employed in the treatment in T2DM. In conclusion Glucose-dependent insulinotropic peptide (GIP) and glucagon-like peptide-1 (GLP-1) are gut hormones that regulate the amount of insulin exocytosis (17).

Despite all the therapy available, it is of utmost importance to fathom the underlying mechanisms to develop a better treatment. Thus, the understanding of the pathophysiology in different tissues, such as liver and skeletal muscle is still relevant. For this matter, high throughput technology, such as cytogenomic array, has aided to unravel such mechanisms of the genetic and molecular basis of T2DM (18).
An enormous wealth of data has been obtained, not only with transcriptomics data analysis, but also with Genome Wide Association (GWA) studies. Genetic variation includes SNPs (Single Nucleotide Polymorphisms), CNVs (Copy Number Variants), aneuploidies among others (19). GWA studies aim to identify Trait Associated SNPs (TAS) or haplotypes (clusters of SNPs) associated with various traits. A SNP (pronounced “snip”) is a variation at a single position in the genome which can be in strong correlation known as Linkage Disequilibrium (LD) with other SNPs. SNPs can lead to synonymous mutations, when a modification in a codon codes for the same aminoacid, non-synonymous mutations lead to an amino acid modification and might result in truncated proteins or frameshift mutations, which can alter both the structure and the function of the protein (20).

A large cohort study, The Wellcome Trust Case Control Consortium (WTCCC) explored T1DM and T2DM and 5 other complex human diseases of major public health relevance. It performed a GWAS that examined 2,000 individuals for each of the diseases and a set of 3,000 shared controls. In regard to T2DM, it reported robust associations of the following SNPs: rs1801282 in the PPARG gene, rs5219 in the KCNJ11 gene and rs7903146 in the TCF7L2 gene (21). Moreover, several variants have been associated to diabetic kidney disease, retinopathy, nephropathy and CDV disease (22).

**Research question:** In the current study, the aim was to re-analyze a transcriptomics dataset from hepatic biopsies in T2DM patients to seek for relevant mechanisms of the disease. We focused on the liver due to its important role in insulin resistance and glucose homeostasis in general (23). Additionally, a GWAS dataset from a meta-analysis in T2DM was merged with the transcriptomics dataset in order to look for associations between both analysis. We hypothesized that this integrative approach can strengthen the comprehension and provide insight of mechanisms in T2DM.
2 Materials and Methods

2.1 Transcriptomics dataset

A dataset from a study published by Misu et al (2010) was selected from the Gene Expression Omnibus (GEO) repository (accessed using the number GSE23343) (4). Their study aimed to identify hepatic secretory proteins correlated with insulin resistance. Biopsies were obtained with ultrasonography-guided percutaneous needle in 10 diabetic subjects and 7 healthy controls. Gene expression was measured with Affymetrix Human Genome U133 microarrays.

2.1.2 Pre-processing and statistical analysis

In the present report, the raw data was reexamined using ArrayAnalysis.org (24), an online data analysis workflow that allows pre-processing, quality control and statistical analysis of gene expression data. The data was normalized with the GC-RMA algorithm, and its quality evaluated using the report generated by the website (Appendix A1 and A2). Based on results of quality control procedures, 2 samples from the healthy controls were removed from the analysis. As a result 10 diabetic subjects and 5 healthy controls were used for the statistical analysis. It compared the T2DM group versus the control group. Genes were considered significantly altered with a cutoff on the 2logged Fold Change logFC > 0.58 or <-0.58 (corresponding to a 50% change in expression on original scale) and a significant p-value < 0.05.

2.1.3 Pathway analysis

For the pathway analysis, PathVisio version 3.1.3 was utilized to conduct the pathway statistics, analysis and visualization (PathVisio is available at http://www.pathvisio.org) (25). The analysis was performed using the statistics results with the Curated Human Collection of 277 pathways, (downloaded from WikiPathways at http://wikipathways.org on February 9th 2015) to identify the pathways with an overrepresentation of significantly altered genes, as defined above. Pathway statistics permits to identify the significantly altered pathways, following three criteria; P-value <0.05, a Z score >1.96 and a minimum of three changed genes.

In Table 1 pathways are ordered by the Z-score, where a positive Z-score indicates a pathway with more significantly altered genes than expected by chance. A negative Z-score indicates less significantly altered genes in the pathway in question.
2.1.4 Gene Ontology analysis

In order to fathom and unify the characteristics and functions of the genes, a gene ontology (GO) analysis was used. It describes how genes encode biological functions at the molecular, cellular or biological process levels, by reporting GO terms. Based on the same criteria as in PathVisio, the list of differentially expressed genes was used to identify the GO terms of biological processes affected in T2DM. This approach, not only helps to attain a wider perspective of the pathological processes, but also to acquire a biological interpretation. For this matter, three gene ontology tools were utilized and the results were compared between each other.

1.- GO-Elite (www.genmapp.org/go_elite), requires a target gene list, which consists of the significantly altered genes. Moreover, a background list with all the measured genes is needed. The pruned summary report contains the main statistics from the Overrepresentation analysis (ORA) GO-Elite does not show any graphical display but data can be visualized as a network in Cytoscape (www.cytoscape.org) using generated output files (26, 27).

2.- Gorilla, Gene Ontology Enrichment Analysis and visualization tool (available from www.cbl-gorilla.cs.technion.ac.il) displays the results color-coded on the GO tree, with the significantly changed elements highlighted. Similar to GO-Elite, it requires a background and target files (28).

3.- The App ClueGO (apps.cytoscape.org/apps/cluego) not only uses the visualization framework of Cytoscape, but also creates a functionally organized GO/pathway term network that can be used for further analysis. It does not require a denominator or background file to be uploaded, only a simple text format file with the target genes (29). Finally, in order to compare the results visually, diagrams were generated in Venny 2.0 (www.bioinfogp.cnb.csic.es/tools/venny/) and Bio Venn (http://www.cmbH.nl/cdd/biovenn/) to analyze the differences and similarities between the tools described above (Appendix B) (30, 31).

2.2 GWAS dataset

In order to combine the gene expression data with genetic variation data we utilized a GWAS dataset from Johnson A et al (2009). They collected articles through PubMed searches (GWAS, GWA, WGAS,WGA, genome-wide, genome wide, whole genome, all terms plus association or plus scan), finally they built a gene-annotated database with 56,411 significant SNPs from 118 GAWS selected articles.
The SNPs listed in the dataset are associated with different phenotypes/traits including T2DM. Each article was selected based on the following criteria: 1) the SNP has an identifiable ID or verifiable genomic position, 2) the SNP reported statistical P-value 3) P-value is less or equal to 0.001 and 4) the P-value is less than or equal to 0.05 if the association is derived from replication, fine mapping or re-sequencing efforts or if it is identified as belonging to a locus or region previously identified by the authors. In cases of large amounts of data, they performed the processing of the associations by using a customized Perl program to facilitate the process.

2.2.1 Analysis of GWAS dataset

In the present study we extracted from the dataset 2,022 SNPs linked to T2DM that constituted the unprocessed GWAS dataset (32). Moreover, the GWAS dataset was processed for SNPs prioritization and visualization (Figure 2) using the online tool GWAS3D (33). The tool required specific inputs to run the analysis such as the list of SNP id (rsnumber) associated with P-value from the GWAS dataset, a P-value cutoff (10E-5), a SNP data set to fetch LD information of leading variants (HapMap I+II+III), and the population of interest (CEU: European American (CEPH)) (33).

Figure 2 Visualization after prioritization using GWAS 3D. Circos-Style plotting) Top variants with highest regulatory signals and distal interaction regions are displayed in the outer circle. The genes or genomic locations connected to respective SNVs are showed in the inner circle. Red lines connect variants with interactions. Wideness of the lines represents the intensity of interaction.
2.3 Network and Integrated analysis

First of all, the significantly altered pathways from the transcriptomics dataset were merged using Cytoscape 3.2.0. (Figure 5). The resulting network was extended with both transcription factors (Appendix C) and approved drugs using Cytargetlinker App (Appendix D) (34). Moreover, the first network was merged with the 2,022 SNPs linked to T2DM, as shown in Figure 6. Finally a pathway-based analysis using the prioritized variants and the curated Wikipathway human collection to identify relevant pathways was performed. The SNPs were associated to their corresponding locus/gene by using biomart ensembl (http://www.ensembl.org/biomart).

![Diagram](image.png)

**Figure 3** Transcriptomics and GWAS network biology workflow. The workflow includes seven different tools that can connect the output by importing it to the next tool.
3 Results

3.1.1 Transcriptomics dataset

In the transcriptomics dataset 20,009 genes were measured. The statistical analysis results showed a total of 4,166 genes significantly altered (P-value <0.05) in T2DM patients versus controls, of which 866 genes were at least 50% modified. From this, 311 genes were up-regulated and 555 genes were down-regulated in diabetic hepatic cells.

3.1.2 Gene ontology

GO analysis was performed with various tools to seek for biological processes affected by the altered genes. First, an analysis with both up- and down-regulated genes identified GO terms associated metabolic processes, such as lipid and cholesterol metabolism. Only the GO term of apoptotic cell clearance (GO:0043277) was identified in all the tools. Figure 4 shows a Bio Venn diagram with the overlap of GO terms reported by ClueGO, GO Elite and Gorilla. Additionally, an enrichment analysis with the up-regulated genes, with all tools identified cholesterol biosynthetic process (GO:0006695). Moreover, the analysis with all the tools of the down-regulated genes reported organic acid metabolic process (GO:0006082) and base-excision repair, AP site formation (GO:00062285).

Figure 4 Diagram obtained from Bio Venn. Image displays the overlap of GO analysis of different tools.
3.1.3 Pathway analysis

Although, the study of single genes can shed light in the comprehension of relevant mechanisms, pathway analysis can lead to a more holistic and valuable understanding of the biological processes. Pathway analysis, was performed in PathVisio 3 using the Curated Human Collection from WikiPathways. The statistical criteria previously described aided to pinpoint eight different pathways. In Table 1, the pathways are ranked according to the highest Z-score and are related to Cholesterol and Lipid metabolism. Table 1 also reports genes inside the pathways that fulfill the criteria. Appendix E shows all significantly altered pathways.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Z-score</th>
<th>P-value</th>
<th># Genes</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol Biosynthesis</td>
<td>6.36</td>
<td>0.001</td>
<td>6/31</td>
<td>↑ HMGCR, HMGCS1, SC5DL, IDI1, SQLE, FDFT1</td>
</tr>
<tr>
<td>SREBF and miR33 in Cholesterol and Lipid homeostasis</td>
<td>4.7</td>
<td>0.001</td>
<td>5/19</td>
<td>↑ HMGCS1, HMGCR ↓ NR1H3, SREBF1, MED15</td>
</tr>
<tr>
<td>NAD Biosynthesis</td>
<td>4.12</td>
<td>0.001</td>
<td>3/34</td>
<td>↓ KMO, HAAO, NADSYN1</td>
</tr>
<tr>
<td>Statin pathway</td>
<td>3.88</td>
<td>0.001</td>
<td>6/46</td>
<td>↑ HMGCR, SQLE, FDFT1 ↓ DGAT1, LRP1, ABCG8</td>
</tr>
<tr>
<td>Alanine and aspartate metabolism</td>
<td>3.3</td>
<td>0.005</td>
<td>3/62</td>
<td>↓ GPT, PC, ASPA</td>
</tr>
<tr>
<td>Osteopontin Signaling</td>
<td>3.3</td>
<td>0.012</td>
<td>3/14</td>
<td>↑ SPP1, IKK-alpha, ITGAV</td>
</tr>
<tr>
<td>Osteoblast signaling</td>
<td>3.1</td>
<td>0.005</td>
<td>3/19</td>
<td>↑ PDGFRA, ITGAV ↓ PTH receptor</td>
</tr>
<tr>
<td>SREBP signaling</td>
<td>2.86</td>
<td>0.01</td>
<td>8/71</td>
<td>↑ HMGCR, HMGCS, FDTF, SQLE ↓ SCAP, SREBP1a-c, nSREB, ARC105</td>
</tr>
</tbody>
</table>

Table 1 List of eight significantly altered pathways from the statistical analysis in PathVisio. All altered pathways meet the following criteria: Z-score > 1.96, P-value < 0.05, and minimum of 3 changed genes. The (# Genes) column represents the number of differentially expressed genes in the pathway compared to the measured number of genes in each pathway. The arrows indicate up ↑ and down-regulation ↓
3.2.3 Network Building

Networks or graphs, not only provide a broader perspective of the biological processes through its topology, but they can also aid to unravel crucial interactions between pathways (35). Figure 5 shows the significant pathways merged together. Some of the highly connected nodes are HMGCR, FDFT1, SQLE, HMGCS1 and Acetyl-CoA which are present in at least three different pathways. Moreover, other components found in at least two pathways, are MTOR, LDLR, LPL, IDI1 and L-Glutamine.

Figure 5  Network of relevant processes of T2DM. Significant pathways from pathway analysis merged together. Hubs are shown in red.

3.3.1 Integrated network of altered pathways with T2DM variants

The network in Figure 6 shows the integration of the significantly altered pathways (orange nodes), their participating genes (light blue) merged with the SNPs from the GWAS dataset. The analysis depicts 15 different SNPs linked to genes in 4 of the pathways altered in T2DM. The statin pathway shows most of the associated variations in genes, such as CETP, APOC1, APOB, LIPC, LDLR and LPL. The SREBP signaling pathway includes associated variations in PPARG, CREB, SECE31B and LPL. Moreover, the Osteopontin signaling pathway shows associated variations in NFKB1 and IKK-alpha (CHUK).
Finally in the SREBF and miR33 in Cholesterol and Lipid homeostasis pathway variations linked to LDLR gene are found. Additionally, LDLR and LPL associated to variants rs6413504 and rs328 respectively, are highly connected to statin signaling, SREBP signaling and SREBF and miR33 in Cholesterol and Lipid homeostasis.

Ensembl biomart ([http://www.ensembl.org/biomart](http://www.ensembl.org/biomart)) was used to obtain scores and predictions if the found variants might affect protein function. Only rs1801282 in PPARG reported a 0.05 and 0.0 for SIFT and Polyphen scores respectively, however results are contradictory. A SIFT score from 0.00-0.05 predicts a deleterious effect, while a Polyphen score of 0.00-0.99 are classified as benign. Moreover, s328 reported a stop gained as a functional consequence (Appendix F).
3.2.1 GWAS dataset analysis using GWAS3D

GWA studies have generated large amounts of data, therefore, it is necessary to discriminate the relevant TAS that can aid in screening and diagnosis of complex diseases (36). For this purpose, GWAS3D was utilized to process the GWAS dataset for prioritization and visualization of the variants associated to T2DM. The analysis was performed with a P-value cut-off of 10E-5, and reported 864 SNPs (Appendix G). A cluster with rs7903146 in the transcription factor 7-like 2 gene (TCF7L2) as the leading variant showed the most significant association. It included rs4506565 (p-value of 1.19E-47), rs7074440 (1.94E-47), rs7901695 (p-value of 2.67E-47) and rs4132670 (p-value of 1.04E-46) in the same gene.

Three highly associated genes were the insulin-like growth factor 2 mRNA-binding protein (IGF2BP2), the FTO gene and the Cyclin-dependent kinase 5 regulatory subunit-associated protein 1-like (CDKAL1) gene. GWAS3D reported 32 SNPs for IGF2BP2 with rs4402960 as the leading variant, 38 SNPs for FTO with rs8050136 as the leading variant, and 18 SNPs for CDKAL1 with rs7754840 as the leading variant. Finally, rs1800775 (CETP) and rs4420638 (APOC1) leading variants were associated with T2DM before and after prioritization.

3.3.2 Pathway collection and prioritized variants

In order to identify relevant pathways linked to the genes of the GWAS dataset, the WikiPathways curated human pathway collection and the prioritized variants (P-value of 10E-5) were integrated into one network using Cytoscape. Both, FTO and IGF2BP2 are not present in any of the pathways in the collection. Appendix H shows the genes with their T2DM associated SNPs and links them to known biological pathways in which they are present.

3.3 Investigation of drug targets in altered pathways

In the network of altered pathways, we identified several genes with variants associated to T2DM. CyTargetLinker was used to extend the network with known drug-target interactions from DrugBank. This showed that three of the genes with variants, LPL, PPARG and CHUK, are known to be targeted by different drugs. PPARG is targeted by hypolipidemic, hypoglycemic and anti-inflammatory agents. LPL is targeted by adsorbent and non-ionic detergent agents. Finally, CHUK is targeted by anti-inflammatory and expectorator antiviral agents.
4 Discussion

4.1 Transcriptomics data

The development of T2DM is related to obesity, and excess in adiposity is known to lead to augmented free fatty acids (FFA) and insulin resistance. Together, these factors lead to an increased risk of CDV disease which plays an important role in the pathogenesis and complications of T2DM (37). Accordingly, pathway and GO analysis of the transcriptomics dataset showed altered pathways related to T2DM, including cholesterol biosynthesis as the most significant (Table 1). In relation, studies in animals have shown different results, for instance a study in mice conducted by Kim et al (2004) shows down-regulation of the cholesterol biosynthesis in the liver after a high fat diet (HFD) (38). another study in rats conducted by Sun et al (2014) shows up-regulation of cholesterol biosynthesis in pancreatic islets in response to glucose (39). Consistent with the nutrition affecting hepatic gene expression, Sanderson et al (2008) identified a regulatory role of PPARα in mice in response to a higher degree of unsaturation and increased fatty chain length in the diet (40). Although diet seems to play a role in different tissues, an in vitro study, conducted by Xinrui et al (2012), described up-regulation of the cholesterol biosynthesis regardless of the dietary treatment (41). Therefore, it is of interest to understand to what extend diet might regulate gene expression in vivo.

Linked to nutrients playing a role in gene expression, beta-cells constantly monitor nutrient availability and can react in response to the signals (42). In the present study, the alanine and aspartate metabolism pathway shows down-regulation of Glutamic-Pyruvate Transaminase (GPT), Aspartoacylase (ASPA) and Pyruvate Carboxylase (PC). PC deficiency in particular, has shown a pivotal role in diabetes through enhanced gluconeogenesis in mice and rat studies (43).

The identification of relevant transcription factors highlighted Sterol regulatory element-binding protein (SREBP), This gene has three isoforms due to specific promoters and alternative splicing, and is known to regulate more than 30 genes required for cholesterol and fatty acid synthesis. SREBP-1a has an important role in fatty acid and cholesterol synthesis, SREBP-1c and SREBP-2 are the predominant forms in the liver. They mediate transcription of genes encoding insulin and enzymes involved in glycolysis, lipogenesis and gluconeogenesis (44). Interestingly, an increase of cholesterol biosynthesis and uptake, regulated by SREBP feeds back and inhibits its activity.
Although, this might explain the decrease of expression of SREBF1 observed in this study, other co-activators might play a role in the transcriptional regulation. Additionally, glycogen synthase kinase 3b (GSK3b) has been reported to induce SREBP phosphorylation that promotes ubiquitylation and its degradation (45). However, GSK3b did not show modification in expression in the present study.

*Upstream Transcription Factor 2 (USF2)*, which belongs to the Myc family of transcription factors, is ubiquitously expressed in mammals. It has recently been associated to cancerogenesis, due to inducing cell migration after being phosphorylated by GSK3b which is in the SREBP signaling pathway (46). USF2 is furthermore linked to diabetic complications, for instance, Liu *et al* (2007) in a mouse study demonstrated that overexpression of USF2 in the kidney increased Thrombospondin 1 (TSP1) gene expression and subsequently Transforming Growth Factor beta (TGFβ). TGFβ has a direct involvement in diabetic nephropathy through fibrogenic components including fibronectin, collagen, Osteopontin among others (47). Although, USF2 is down-regulated in the hepatic tissue in the present study, pathway analysis revealed that Osteopontin signaling pathway is significantly altered (Table 1). Interestingly, Bertola, A *et al* (2011) in a human study found that increased expressions of Osteopontin and its receptor (CD44) were strongly correlated with both insulin resistance and hepatic steatosis in obese patients and mice, possibly due to the accumulation of triglycerides (48).

SP2 another transcription factor has also been associated to tumors in a study in rats (49). SP2 transcription factor was found to be up-regulated in the present study. It encodes a protein with several transactivation domains. Not only it interacts with E2F1 but also mediates cell proliferation and p53-dependent/independent apoptosis which have been linked to both T1DM and T2DM (50).

*TRIM28*, forms a chromatin remodeling scaffold protein that plays a pivotal role in immunity, in particular for Tregs differentiation and function (51). In relation, an imbalance in pro-inflammatory and anti-inflammatory T cells including Tregs, has been reported in T2DM. Moreover, the study conducted by Zheng *et al* (2012) reported that T cell polarization can be modulated by HDL levels (52). Although the levels of HDL of the patients in the study are not available, the most significantly altered pathway in the transcriptomics dataset is cholesterol biosynthesis.
In relation to epigenetic mechanisms, K (lysine) acetyltransferase 2A (KAT2A) was down-regulated. It is a histone acetyltransferase (also known as GCN5) that plays a role in transcriptional modulation, nucleosome assembly, cell cycle regulation, DNA repair, telomere maintenance and genome integrity maintenance. KAT2A together with SIRT1 have been discovered to be regulators of gluconeogenic gene expression. A study in db/db mice treated with metformin revealed that KAT2A protein and its mRNA were increased and a reduction in glucose and insulin were observed after the treatment. Thus, the treatment with metformin mediates gluconeogenesis inhibition, and as a result reduces endogenic glucose production typically observed in T2DM (53).

4.2 Network of the altered pathways and GWAS dataset

WTCCC (21) reported relevant genes associated with T2DM, such as PPARG, KCNJ11 and TCF7L2. In the network generated from the transcriptomics data and the 2,022 SNPs (Figure 6) only PPARG was identified, showing two associated variants, rs10510418 (P-value 0.048) and rs1801282 also called P12A (1.7E-6). Based on Ensembl Biomart analysis, the P12A variant had a SIFT score of 0.05. SIFT predicts substitutions with values less than 0.05 as deleterious, however, some authors claim that a score of 0.1 provides better sensitivity (20). Moreover, OMIM (Online Mendelian Inheritance in Man ,http://omim.org), confirms that P12A polymorphism (rs1801282) is associated with diabetes susceptibility as it is reported by a variety of international consortia; such as: WTCCC itself, Diabetes Genetics Initiative (DGI) and Finland-United States Investigation of NIDDM Genetics (FUSION) (54).

Another relevant gene associated with four variants in the transcriptomics network is Cholesteryl ester transfer protein (CETP). Its function is to mediate the exchange of lipids between lipoproteins and subsequent uptake of cholesterol by the hepatocytes in a process known as reverse cholesterol transport (55). In this study variant rs1800775 was present for CETP. It is located in the promoter of the gene and might modify the gene function possibly by changing the binding site and as a result repress the expression (56). Moreover, rs1800775 has been significantly associated with higher plasma HDL levels in Chinese men (57). Additionally, another CETP variant, rs5882, is linked to high levels of HDL and a reduction in risk of dementia including a lower incidence of Alzheimer disease in elderly (58).
In relation to lipid metabolism, various Apolipoprotein genes showed associations with genetic variation. rs4420638 in Apolipoprotein C1 (ApoC1) has a strong LD with Apolipoprotein E (ApoE) and has been strongly associated with Alzheimer disease (59). ApoE (synthetized in the liver and brain) removes excess of cholesterol and triglycerides, thus involved in mediating clearing of lipids (60). Interestingly for the lipid metabolism, a cohort study obtained a genotype score to assess cardiovascular risk based on the measurement of 9 SNPs that affected either HDL or LDL levels. The score included rs693 in ApoB (P-value 2E-11), rs328 in LPL (P-value 3E-12) rs1800775 in CETP (P-value 2E-29) and rs4420638 in APOE cluster (P-value 3E-21). Increasing genotype scores were linked to increases in LDL and decreased HDL levels, thus predisposing of cardiovascular risk (61).

Another relevant locus, CHUK is a ubiquitously expressed kinase that modulates the NFKB transcription factor dependent activation of several genes involved in insulin sensitivity (62). Interestingly, a metanalysis identified rs11597086 in CHUK and 8 other variants in ERLIN1 and CWF19L1 to be associated with high susceptibility to both simple steatosis and hepatic steatosis with inflammation (63).

In relation to NFKB1 a study in Japanese individuals identified that a haplotype block which contains rs230539, appeared to be related to gastric cancer. Lu R et al (2012) reported that people who carry the homozygous genotype GG in the rs4648068 in NFKB1, located in the haplotype block, show increased risk for gastric cancer (64). In association to HDL levels, the rs261332 variant in LIPC gene was identified in a meta-analysis to be associated, as indicated by a P-value of 5.02E-7 with a 95% confidence interval of 1.03-1.06 (65).

### 4.3 Variant prioritization

The GWAS analysis in Johnson et al dataset (2010) identified highly associated regions of the genome, replicated loci and novel loci suggestive of a variety of traits. With respect to T2DM the analysis conducted by Johnson et al reported rs4740283 as the most statistically significant associated SNP. However, rs4740283 was not identified in the present study before and after prioritization. Appendix G shows both the leading and associated variants reported after prioritization. Both FTO and IGF2BP2 were associated to several variants and rs7903146 in TCF7L2 showed the most significant P-value (1.00E-48).
4.4 Pathway collection and prioritized variants

Although, FTO and IGF2BP2 were not connected to any pathway in the network, they have reported association with T2DM which implies the necessity to update the pathways constantly. First, FTO gene (rs9939609) has been correlated with higher BMI, systolic Blood Pressure (BP), fasting insulin and glucose, HbA1C, Triglycerides (TG), C-reactive Protein and lower HDL-C (66). According to HapMap, rs9939609 is correlated with 45 additional SNPs, including rs7202116 and rs17817449 with strong LD (67). Another relevant variant with robust associations to T2DM is rs4402960 in the second intron of insulin-like growth factor 2 mRNA-binding protein (IGF2BP2) (54).

Among the variant-containing genes associated to the pathway collection, the most relevant are TCF7L2, GCKR, CDKAL1, KCNJ11, ABCC8, and the already discussed CEPT and APOC1. TCF7L2 cluster has previously been associated to T2DM in diverse ethnic groups, in fact some studies have reported p-values of 10E-80 and 10E-140 (68). TCF7L2 is a pleiotropic gene that plays a pivotal role in glucose homeostasis. Its mechanism in T2DM is thought to be related to beta-cell dysfunction within the pancreas, through proglucagon repression via Wnt signaling pathway which might also affect insulin resistance (69). Interestingly, GCKR belongs to the Wnt signaling pathway. GCKR overexpression has been linked to increased triglycerides levels. Additionally, rs780094 in GCKR has been associated to lower glucose (P < 0.10, P < 0.02 respectively), less insulin resistance (HOMA-IR P < 0.05 and P < 0.01), and lower risk of T2D (P < 0.20, P < 0.03)(54).

The Cyclin-dependent kinase 5 regulatory subunit-associated protein 1-like (CDKAL1)gene is highly expressed in the human pancreas, skeletal muscle, and brain tissue. rs7756992 A/G polymorphism has been associated with T2DM. CDKAL1 controls first-phase insulin exocytosis in β cells by facilitating ATP generation, K (ATP) channel responsiveness and the activity of Ca (2+) channels. Although a mutation in CDKAL1 might alter insulin secretion, the exact mechanism is still unclear. In a metanalysis, rs7756992 A/G polymorphism was associated with increased risk of T2DM, particularly in Caucasian and Asian populations but not in African descent population (60).
Other relevant genes in T2DM are *KCNJ11* and its adjacent gene, *ABCC8*. *KCNJ11* gene encodes an ATP-sensitive potassium channel, whereas *ABCC8* encodes the functionally related sulfonylurea receptor (*SUR1*). rs5219 in *KCNJ11* and rs757110 in *ABCC8* are in strong LD and they have been found to play a role in the pathogenesis of T2DM, in Caucasian population in particular. Moreover, rs5219 seemed to predispose women to develop Gestational Diabetes Mellitus in Scandinavian women (70, 71).

As previously explained, *PPARG, KCNJ11 and TCF7L2* have reported the most robust associations. Interestingly, in this study both rs5219 and rs5215 in *KCNJ11* were found with a final p-value of 1.19E-12 and 1.32E-12, respectively. Moreover, not only *KCNJ11* but also *TCF7L2* was significantly associated, in the processed GWAS dataset.
5. Conclusion

In this study we aimed to combine a transcriptomics dataset with a GWAS dataset in an attempt to seek for associations between genetic variations and patterns of gene expression in T2DM. In the transcriptomics dataset, both pathway analysis and gene ontology reported enriched pathways and biological processes associated with T2DM. Similarly, a pathway-based analysis using the curated Wikipathway Human collection with the prioritized variants revealed pathways related to T2DM (i.e. statin pathway, Wnt signaling, T2DM pathway, SREBP signaling among others). In addition Il-1 signaling, apoptosis and Toll-like receptor signaling pathway were identified, which have been associated with the immune system activation in T2DM. Moreover, network building aided in the identification of hubs and relevant transcription factors, which reported associations to fatty acid synthesis, cholesterol biosynthesis, Osteopontin signaling, cell proliferation/apoptosis, as well as immune interactions and epigenetic mechanisms.

Among the variants identified, they report associations with metabolic disturbances related to the pathological mechanisms in T2DM, PPARG, KCNJ11 and TCF7L2 in particular show significant associations with T2DM as previously reported by International consortia. In conclusion both analyses of transcriptomics and GWAS data aided to pinpoint relevant loci associated to mechanisms in T2DM.

Limitations

A reductionist mindset in biology veered towards a systems biology approach as a result of progress in molecular biology with high throughput technology. However, this approach has advantages and disadvantages. For instance high throughput technology requires standardization to reduce possible batch variations, pre-processing to evaluate the quality of the raw data and adequate statistical analysis. Moreover, in regard to GWA studies is of utmost importance to specify that the associations originate from large cohort studies at the population level and an extrapolation to the individual level should always be done with caution, due to the complexity of T2DM and its multifactorial nature.
**Future perspectives**

Undoubtedly, the measurement of gene expression and genetic variation with high throughput technology has evolved to clarify the molecular basis of T2DM and other diseases. Although other technologies have gained territory, such as next generation sequencing, microarray technology is well-established. Not only because specialized arrays have become more accurate and reliable but also are expected to provide valuable insight in biomarker detection, screening, prevention and diagnostics which contribute to the improvement of personalized medicine.

In regard to T2DM, although numerous variants in several loci have been identified, they still poorly explain the mechanisms. In clinical settings, genetic testing is linked to little predictive power, thus it should not distract from modifiable risk factors, such as diet, lifestyle and medication in the therapy of T2DM. As a result it is of utmost importance to interpret the effect of the variants, which not only can provide more insight of the mechanisms but also to classify subtypes to T2DM. Finally, larger studies in non-European populations should be conducted to obtain insight of genetics and the differences among populations.
Appendix A-H

http://projects.bigcat.unimaas.nl/data/students/ManuelGonzalez/Appendix.pdf
References

12. Rydén L, Grant PJ, Anker SD, Berne C, Cosentino F, Danchin N, et al. ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD2013 2013-08-31 00:00:00.


