DOCTOR OF PHILOSOPHY

Multi-Omics to Study Causes and Consequences of Type 2 Diabetes

Davtian, David

Award date:
2024

University of Dundee

Multi-Omics to Study Causes and Consequences of Type 2 Diabetes

David Davtian

Supervised by Dr. Andrew Brown and Prof. Ewan Pearson

20th of March 2024

Table of Contents

Acknowledgments

I would like to thank first my supervisors, Dr. Andrew Brown and Prof. Ewan Pearson for their advices, continuous support and patience during my PhD. They encouraged me throughout these 4 years and shared with me their knowledge and experience needed for this project. During this period, even through the Covid-19 crisis, they never stopped helping me whenever I needed support, either professionally or personnally, until the end. I want to thank also Dr. Ana Viñuela who tremendously supported me for my research whether it was for scientific questions, corrections or personal advices. Thank you also Theo, the last member of our group and my fellow frenchman for his help and the great moments during these years. My gratitude extends to the University of Dundee and the Faculty of Medicine and members of the Division of Population health and Genomics, Prof. Colin Palmer, Dr. Suzanne Grant, Jennifer Watson, who also supported me and without whom I wouldn’t be able to submit this work. Many thanks also to Tenovus Scotland for the funding opportunity to underake this study. I would also like to thank my friends and colleagues, Moneeza, Ally, Dafni, Aline, Ambra, Margherita, Dina, for a cherished time spent together working in person or remotely, and in social settings.

Of course I would not have managed to do this without the continuous support from my parents, thank you mom and dad, for everything. As well as my friends from France, merci a la banana, Daz, Jujine, Mamine, Nanas, Thomas, Rahim and Soub, le meilleur groupe d’amis qu’on peut souhaiter avoir ; but also Ilo, Jess, Doli, Robert, Ines, Jade, Mel, I am very grateful to have all of you in my life.

I, David Davtian, declare that I am the author of this thesis. All references cited were consulted by me, the work of this thesis was done by me and has not been accepted for any higher degree.

Abstract

This thesis discusses the benefits and limitations of using molecular studies to understand Type 2 diabetes (T2D). As a complex disease, T2D is caused by a combination of multiple genetic and environmental factors. Its consequences on the glucose homeostasis often trigger a variety of co-morbidities, including cardiovascular diseases, stroke, and neuropathies, generating a major health burden. In the first chapter, we looked for genes which were differentially expressed with T2D across 49 tissues using data created by the GTEx consortium. In twelve of those tissues, we found between 1 and 5174 genes to be significantly differentially expressed (FDR=0.05). The relevance of the tissue was not a major factor in the number of genes discovered. In fact, in some cases the tissue suggested that the differential expression was due to comorbidities with T2D rather than the disease itself. For example, the most differentially expressed genes were found in the tibial nerve, likely due to comorbidities with neuropathy. We also found a larger number of differentially expressed genes with biological factors such as sex or age (a maximum of 14319 and 12897 respectively, FDR=0.05), though this could be due to the limited number of individuals with T2D in the dataset. In chapter two we looked to discover genes responsible for T2D development using a class of methods called transcriptome wide association studies (TWAS), which combine information from genetic studies into disease with information from reference molecular datasets. One question we wanted to answer was about the choice of the molecular reference: was it better to collect data from a tissue that was directly relevant for the tissue, or could a well powered study in a non-relevant accessible tissue also be informative? For this we used expression data from 49 tissues previously mentioned, a mix of relevant and non relevant tissues, expression data from the directly relevant pancreatic islets produced by the Inspire consortium, and data from whole blood produced by the DIRECT consortium, non relevant but with a sample size more than three times larger than the others. We found that the relevance of the underlying molecular data had no effect on the number of T2D causal genes we discovered, instead observing a strong correlation with sample size. More genes found using DIRECT data were confirmed using multiple instrument Mendelian randomisation than with any other molecular dataset. (31 genes confirmed through MR compared to 1 to 9 for the other molecular datasets). However, important GWAS loci were missed when using non relevant tissues; for example, the well known *TCF7L2* locus was only found using pancreatic islets. Finally, our last chapter focused on the observable effects of metformin, a commonly used medication for T2D. We took three approaches to discover genes which interacted with the drug. Firstly, using GWAS summary statistics from a study into metformin efficacy, we applied two methods to find genes associated with drug efficacy: MAGMA, which is based on purely genetic information, and the TWAS methods from the previous chapter. Using MAGMA we discovered 880 associated genes (FDR=0.05) compared to only 38 using TWAS (though TWAS associations contained important information on tissue of action and direction of effect). Finally, we used longitudinal data to look for genes whose expression changed after taking metformin and discovered 348 genes at a less stringent threshold of FDR=0.2. Comparing our differentially expressed genes to TWAS associated genes for efficacy and for causing T2D, we found that metformin tended to produce changes in expression which would promote drug efficacy, a treatment virtuous circle, and counteracted expression patterns predicted to cause the disease, explaining the treatment benefits. However, neither of these properties was statistically significant, due to the small numbers of genes involved. In summary, this thesis has explored the molecular causes and consequences of T2D, as well as the genes and pathways involved in treating it.

Introduction

A brief history of genetics

The origin of modern genetic studies can be traced back to the first works of Gregor Mendel on plant hybridization published in his memoir (Mendel, n.d.). Mendel showed by mating together different varieties of peas that the resulting offspring consistently displayed one of the two traits. He observed that the proportions of these traits in the offspring were approximately 3:1 and classed the more common as a dominant trait and the other as recessive. The term ‘genetics’ was first used by Wiliam Bateson (1906) who connected Mendel’s work with heredity and later expanded the concept not only to plants but also animals. From this time, several important terms were introduced to the scientific vocabulary, which are still used today but in many cases with an altered meaning. In 1909, the word ‘gene’ was suggested by Wilhelm Johannsen, who used it to describe Mendel’s discrete units of inheritance inherited from one generation to the next. He also introduced the term ‘phenotype’ to describe the observable characteristics of an individual, ‘dominant’ and ‘recessive’ gene , respectively a gene where one or two alleles are required to observe the phenotype. Even without understanding how information is passed from parent to child, much of the work of these early geneticists is still relevant today (Gayon 2016) (Figure 1).

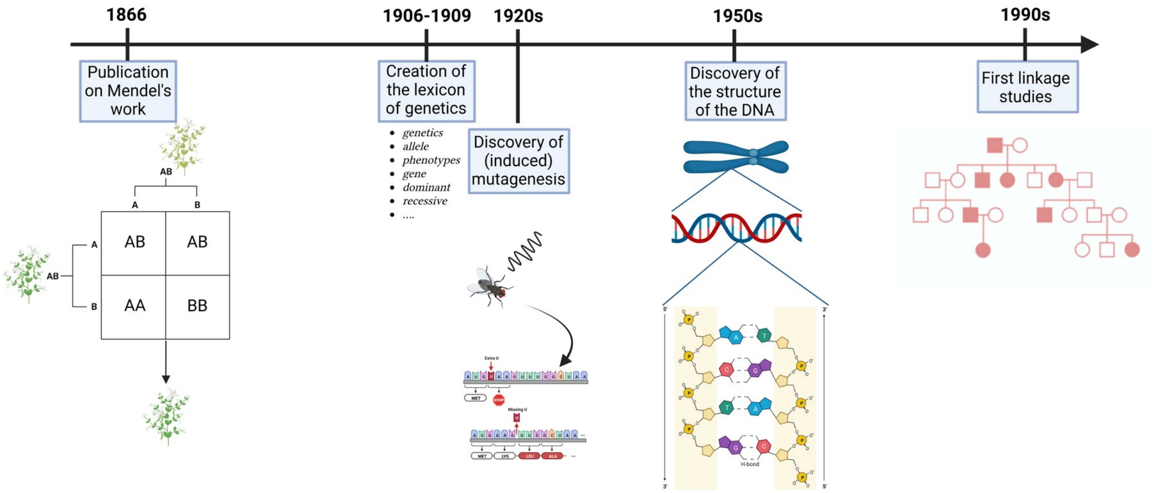

1866
1906-1909
1920s
1950s
1990s
Publication on Mendel's work
AB
A
B
A
B
AB
AB
AB
AA
BB
Creation of the lexicon of genetics
genetics
allele
phenotypes
gene
dominant
recessive
....
Discovery of (induced) mutagenesis
Discovery of the structure of the DNA
First linkage studies

Figure 1: ***Early history of genetics.*** *Figure made in Biorender.*

Understanding how information was passed from parent to child came more slowly. Herman Muller working on drosophila in the 1920s identified alterations to the chromosomes of the flies, but did not know exactly where these alterations were happening. The big breakthrough came in the 1950s, when work by Rosalind Franklin, Francis Crick and James Watson revealed the structure of the molecule by which information is passed down through generations, the double helix of DNA (Watson and Crick 1953). DNA was seen to be made up to four molecules, Adenine, Cytosine, Guanine and Thymine, and the order of these molecules defines the genetic code. Each human inherits 23 chromosomes of DNA from their parents, and some mitocrondial DNA from their mother, and this is known as the human genome. This helped to understand some concept exposed earlier such as ‘alleles’ which we define as the specific amino acid base of the DNA sequence at a given location, and ‘heterozygote’ and ‘homozygote’, having either two different or two identical copies of a specific allele.

First attempts for DNA sequencing

Since then, scientists have been actively working to read this genetic code, first for increasingly large regions of the human genome, more recently for larger numbers of individuals. Starting from the 1970s, methods based on the chemical cleavage of the DNA strands such as the Maxam-Gilbert method emerged as a technique to sequence the DNA. At the same time, techniques such as Southern blot (suggested by Edwin Southern (Southern 1975)) allowed researchers to target specific DNA sequences using complementary probes. Of course, all of these technologies were expensive, labour-intensive and limited in the ability to sequence long sections of the DNA. It is only in the late 1970s that these technologies were developed enough to push towards the expansion of the sequencing era, with Frederick Sanger proposing the first DNA sequencing technique, known as the Sanger sequencing method. This revolutionary method allowed scientists to read the sequence of nucleotide bases in DNA and became the gold standard for sequencing for several decades. Then in the 1990s, The Human Genome Project started globally with the goal to fully sequence the human genome over the next decade (Collins and Fink 1995). Completed in 2003, they provided the first draft of the human genome, which was then used as a basis for many studies on genetics and disease. During that time, sequencing methods shifted from the initial Sanger method to what is now called Next-Generation Sequencing, which significantly increased the amounts of DNA that could be read and reduced costs (Schuster 2008). With this technology, it became feasible to read an individual’s entire DNA for the cost of around $400, compared to the $3 billion cost of the human genome project. The technology to read the code allowed researchers to investigate the function of the human genome, one of the first information we learned was that DNA has coding sequences and non-coding sequences. The first contain instructions on how to produce mRNA molecules later to be converted into proteins and the second is responsible for tasks such as regulation of the gene expression or chromosomal strucure and represents the majority of the genome.

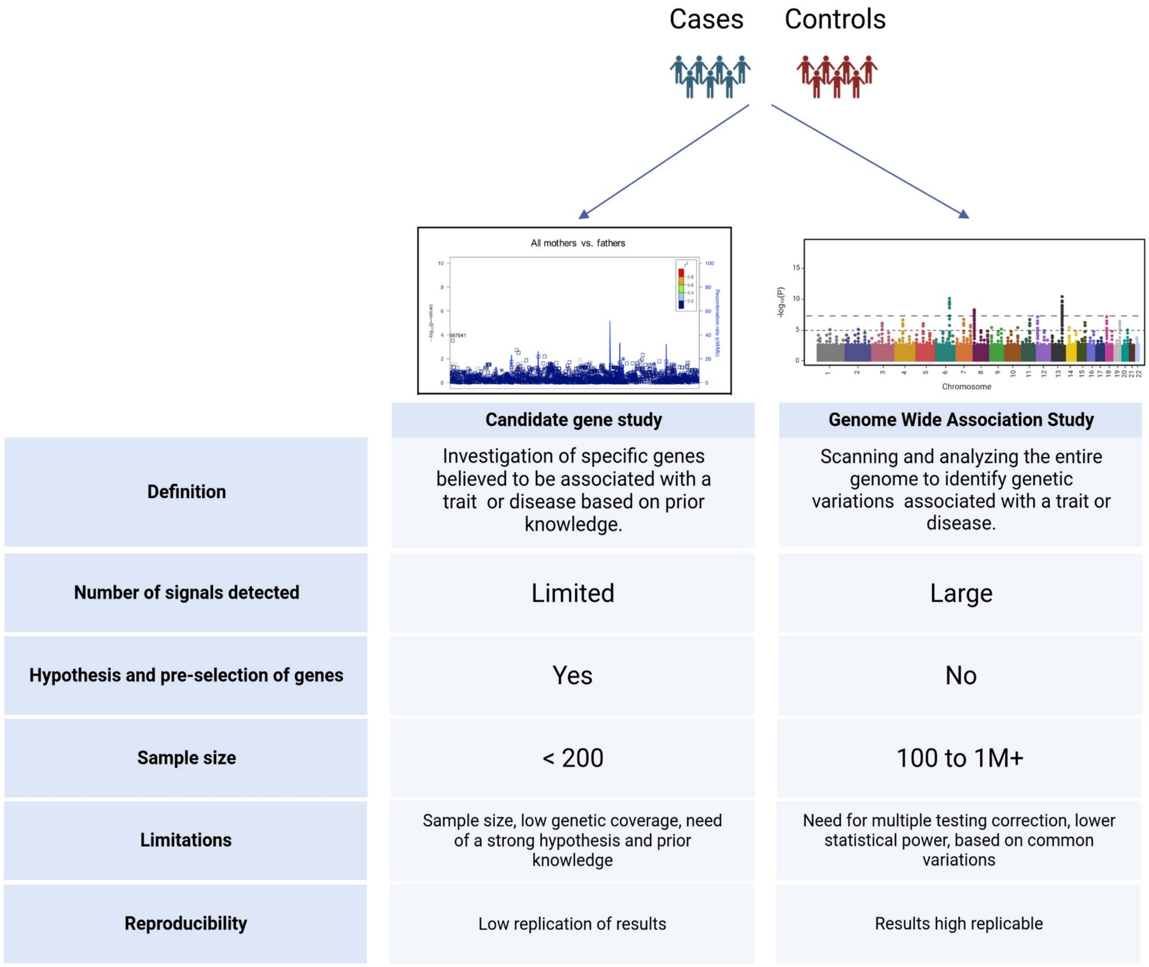
Linkage and candidate gene studies to investigate human traits

Mendel and the early geneticists had shown that offspring inherit traits from their parents, one important trait is risk of disease. Now, with knowledge of the genetic code, researchers moved on to investigate whether these inherited risks could be linked to particular parts of the human genome. Linkages studies were one of the first attempts to do this. These looked at families and were based on the fact that within such families some individuals would get a particular disease while others would not. Researchers were interested in finding out whether all the patients within a family would inherit the same sequence of DNA from their parents, while all the controls would inherit a different sequence. One of the most commonly used genetic marker for these studies were microsatellites repeats, short repeated DNA sequences displaying a high level of polymorphism, therefore easier to characterize and replicate. An early success of linkage studies was the discovery of a strong association between risk of T2D and a region on the chromosome 10q, which turned out later to be the *TCF7L2* risk locus, since validated several times by multiple different approaches (Duggirala et al. 1999; Grant et al. 2006). Doing that, researchers started to step into the complexity of diseases such as T2D which are complex, polygenic traits by opposition to Mendelian traits controlled by a single gene. Therefore understanding each of these individual effects and their inheritance is crucial to rightfully characterize patients with the disease.

Candidate gene studies were another way researchers tried to tie regions of DNA with disease. Often, researchers would have defined hypotheses of which genes and regions of DNA they thought were responsible for disease. By collecting a cohort of healthy individuals and individuals with the disease, they could compare the DNA around and in this gene for all these individuals. Then, at each position they had read, they could ask the question: is one allele more common in cases than controls? The advantage of such studies was that the sample size required was quite low, since they were concentrating on a limited part of the human genome and already had an idea of the mechanisms underlying this particular trait (Kwon 2000). For example, a candidate gene study for Type 2 Diabetes (T2D) could focus on the region around the gene encoding for the insulin receptor or genes involved in insulin secretion and sensitivity. Several candidate gene studies in T2D identified genetic variants associated with the disease, including variants in the *PPARG* and *KCNJ11* genes. A study by (Barroso et al. 2003) investigated the association between 35 genes known to have an influence on insulin action and showed significant association with diabetes for many of them, including *KCNJ11, HNF4A* and *INSR*. Flaws of these candidate genes studies later became apparent. Results from these studies were based on small sample size studies, and frequently failed to replicate in other studies. It became apparent that frequently the original hypothesis of which gene was involved was not as strong as the researcher had thought.

The GWAS era

To address this, research became focused on Genome Wide Association Studies (GWAS). Rather than examining a handful of variants in one region of the genome, in a GWAS, researchers looked at millions of genetic variants throughout the entire human genome. They compared the frequencies of these variants between individuals who had the disease and those without (Bush and Moore 2012). Because they were testing millions of statistical hypotheses, it was not sufficient to use standard P value criteria for rejecting null hypotheses, accepting the standard P < 0.05 would mean concluding one in 20 of the millions of variants were related to the disease, even if none truly were. To address this they came up with the concept of genome-wide significance : to conclude a variant was associated with disease required very strong statistical evidence, typically a P value < 5e-8. This was estimated by correcting the traditional 0.05 pvalue by the estimated number of independent common SNPs (Single - Nucleotid Polymorphisms) across the human genome (Dudbridge and Gusnanto 2008; Jannot, Ehret, and Perneger 2015) (Figure 2).



	Candidate gene study	Genome Wide Association Study
Definition	Investigation of specific genes believed to be associated with a trait or disease based on prior knowledge.	Scanning and analyzing the entire genome to identify genetic variations associated with a trait or disease.
Number of signals detected	Limited	Large
Hypothesis and pre-selection of genes	Yes	No
Sample size	< 200	100 to 1M+
Limitations	Sample size, low genetic coverage, need of a strong hypothesis and prior knowledge	Need for multiple testing correction, lower statistical power, based on common variations
Reproducibility	Low replication of results	Results high replicable

Figure 2: ***Differences between GWAS and candidate gene studies.*** *Although both based on the detection of changes at a global omic scale within a population of individuals, they present differences in term of target, power, hypothesis and requirements. Figure made in Biorender.*

Several large-scale GWAS have been conducted, involving hundreds of thousands of individuals with T2D and control subjects in order to identify genetic variants associated with the disease (Visscher et al. 2012). These studies have identified multiple genetic loci associated with T2D, including the confirmation of the importance of *TCF7L2* discovered previously in genome-wide linkage study (ref of the grant 2006 paper), but also new genes such as *CDKAL1* or *KCNQ1*, among many others (Mahajan et al. 2018; Spracklen et al. 2020). The latest study identified 318 novel regions of the genome associated with the disease and these associations have provided insight into the biological mechanisms underlying type 2 diabetes, such as beta-cell dysfunction and insulin resistance (Vujkovic et al.

2019). GWAS have greatly expanded our understanding of the genetic basis of type 2 diabetes and have opened up new avenues for personalised treatment and prevention strategies.

Because most of the genetic effects were small, the ability of researchers to capture significant associations was limited and often required very large sample sizes, logistically and financially demanding. To address this, many groups worked together in large international consortia, and more recently national biobanks have emerged, allowing many diseases and traits to be investigated in the same study. But even with these massive sample sizes, GWAS studies have mostly only been able to capture common variants with low effect sizes and account only for a fraction of the observed heritability (Manolio et al. 2009). Since GWAS only identifies regions of the genome associated with a trait, it can be difficult to know how the DNA change is related to disease. Almost 90% of GWAS results fall in non-coding regions of the genomes, regions of the DNA which are not transcribed and then translated into proteins but rather control gene expression via regulatory elements such as silencers or enhancers. For example, enhancers, in response to binding with transcription factors, can interact with promoters and initiate the transcription mechanisms. Given that these GWAS hits lie outside the coding region of proteins and do not affect their sequence, it can be difficult to know which genes they affect, and how they affect them.

Technologies evolved to solve the missing link issue

To better understand how GWAS hits affect risk of disease, we can go back to the Central Dogma of Molecular Biology. This states that information is transferred in the cell from DNA to RNA to proteins and then used or converted in other macromolecules. In the case of a GWAS variant, this means that the first consequence of a genetic variant should be on the process of transcription, from DNA to RNA. Therefore, the assumption must be that the initial consequence of non coding GWAS associations should be found by observing gene expression changes. Undestanding these GWAS variants and how they affect risk of disease can start by investigating their impact on gene expression. But to do this, it was necessary to produce a quantification for gene expression, a way of measuring how much DNA was being converted to mRNA.

Many technologies were developed for this purpose and are still nowadays used to generate large scale data. For expression, microarrays and RNA sequencing are two methods which vary in term of scale, speed and cost. Although both are used to provide gene expression quantification, they are based on different techniques. Microarrays use hybridization between probes and labeled RNA as well as fluorescence to derive expression levels (Baldi and Hatfield 2002). RNA-Seq uses fragmented RNA molecules collected from tissues and converted into cDNA, which are then sequenced and aligned to a reference genome to quantify the expression (Wang, Gerstein, and Snyder 2009). Their use depends on the scope of the study, the objectives and also the cost available. Small scale studies with a very well defined hypothesis may prefer microarray for their efficiency at a lower cost, while larger studies with large cohorts could make a use of the bulk power of RNA-Seq to allow the discovery of novel signals at a transcriptome wide level. In general, RNA-Seq is preferred as it avoids technical issues in microarray studies related to probe performance, such as cross-hybridization, limited detection range of individual probes, as well as non-specific hybridization. Nonetheless, RNA-Seq can still present flaws such as cross-hybridization, especially in the case of short-read sequencing which is the best cost-efficiency technology. Mismapping of the reads on the wrong gene when sequences are highly similar and reading length short can happen. In these cases, usually reads mapping to multiple location are filtered out and only uniquely mapped reads are kept for downstream analyses. Moving to long read technologies also allow to solve these issues by reducing the ambiguities of mapping. All these technologies are constantly evolving, and new methods are being developed to improve the sensitivity, accuracy, and speed of molecular phenotyping.

Instead of looking at gene expression changes, another approach is to try to identify proteins mediating GWAS disease risk by quantifying protein levels. The most common method to measure protein levels is mass spectrometry. Proteins are first digested using enzymes (usually trypsin) and isolated using a mass analyser separating them based on their weights. Then, after ionization, the mass-to-charge ratio for the resulting molecules is quantified and compared to known profiles to estimate protein levels. Another method is to use affinity-based methods and immunoassays based on antibodies targeting specific proteins. One example is the Olink method, which uses a pair of antibodies to tag a target protein which, after hybridization, can be extended into quantifiable DNA polymerase. Unlike expression quantification, these are not proteome wide methods but rather targeted for specific sets of proteins or panels (Suhre, McCarthy, and Schwenk 2021).

And finally, another molecular phenotype often measured is metabolites levels. These small molecules are closer to certain disease traits than RNA or proteins and give insights into the different biochemical pathways or therapeutic responses. To measure metabolites, the main method is a combination of chromatography to separate the macromolecules and mass spectrometry to quantify these molecules based on their concentrations. These can be done either in a targeted context, focused on pre-defined specific set of metabolites, or untargeted for a more comprehensive, but also less accurate, overview of the whole metabolome (Clish 2015).

Molecular studies expanding GWAS results

Once we have our quantifications we can test genetic variants for association, where individuals with one allele produce more mRNA, proteins or metabolites than someone with a different one. These are called Quantitative Trait Locus or QTLs. For example, we call an eQTL (expression QTL) a region of the genome where an individual’s expression of the gene depends on their genotype at a particular locus; the eQTL is referred to as a cis eQTL if the locus is within 1MB of the gene, a trans eQTL if it is more distant (Figure 3).

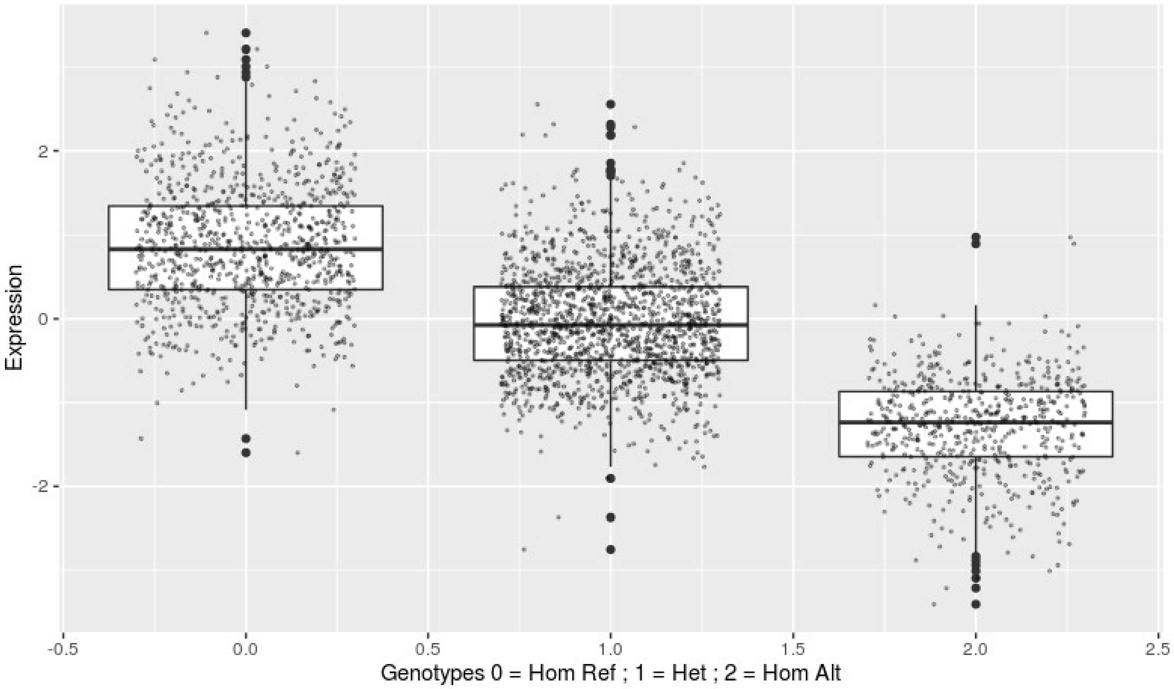


Figure 3: ***Representation of an eQTL.*** *Distribution of individuals' expression for a specific gene separated by genotypes. Either homozygotes for the reference allele (0), heterozygotes (1) or homozygotes for the alternative allele (2).*

The identification of QTLs is referred to as QTL mapping. In cis, this is done looking at each genetic variant in within 1MB of a gene and testing to see if expression is associated with genotype using a linear model. Then, after correction of p-values for multiple testing with methods such as FDR, betas representing the effect of each variant on the phenotype are extracted. In addition, permutation steps can be included in the analysis by randomly shuffling the phenotype data a defined number of times in order to estimate p values adjusted for multiple testing. The advantage here is the ability to identify causal variants for a specific trait in addition to regulatory variants such as enhancers or silencers affecting gene expression (respectively, sequences increasing or decreasing the expression of a gene) (Khetan et al. 2018).

Similarily to GWAS, QTL studies have their own limitations. First, the sample size requirements are relatively high in order to reach enough statistical power.

Without reasonable sample size, signals with low effect sizes can be missed. It is also highly dependent on the population composition and the interpretation for one study might not apply and be replicable in another study, if the other study is based on a population with different characteristics or uses a different tissue. In addition, when studying complex traits, environment usually plays a major role on the phenotype and QTL studies rarely take in consideration the gene-environment interactions. Efforts are being made to solve this by studying interactions QTLs and context-specific QTLs with traits such as age, cell type or other phenotypes (Kasela et al. 2023).

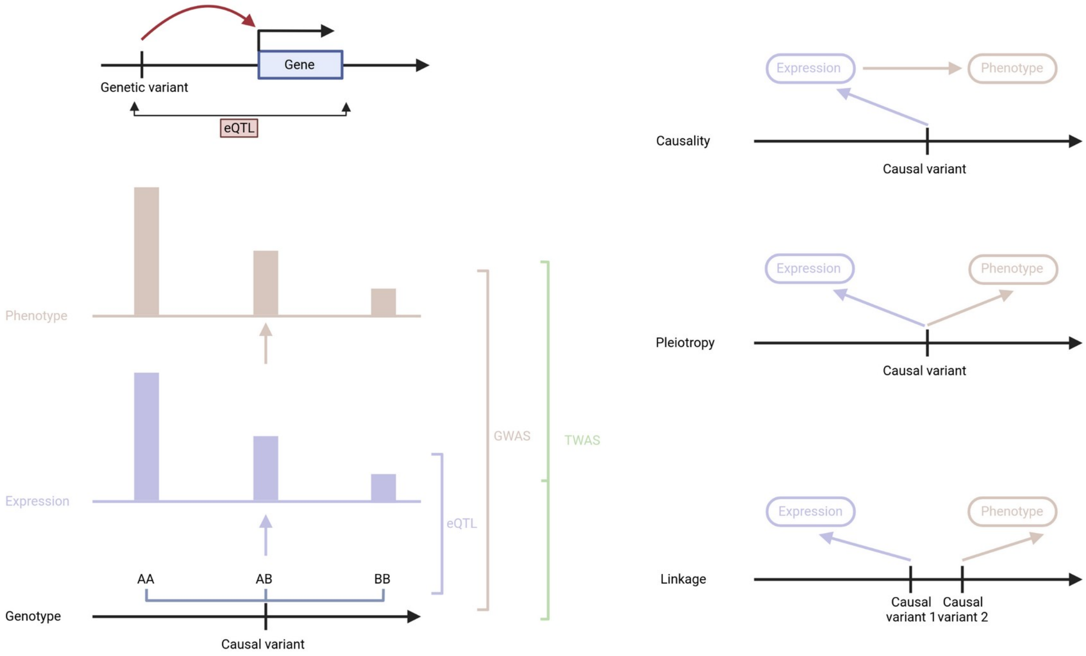
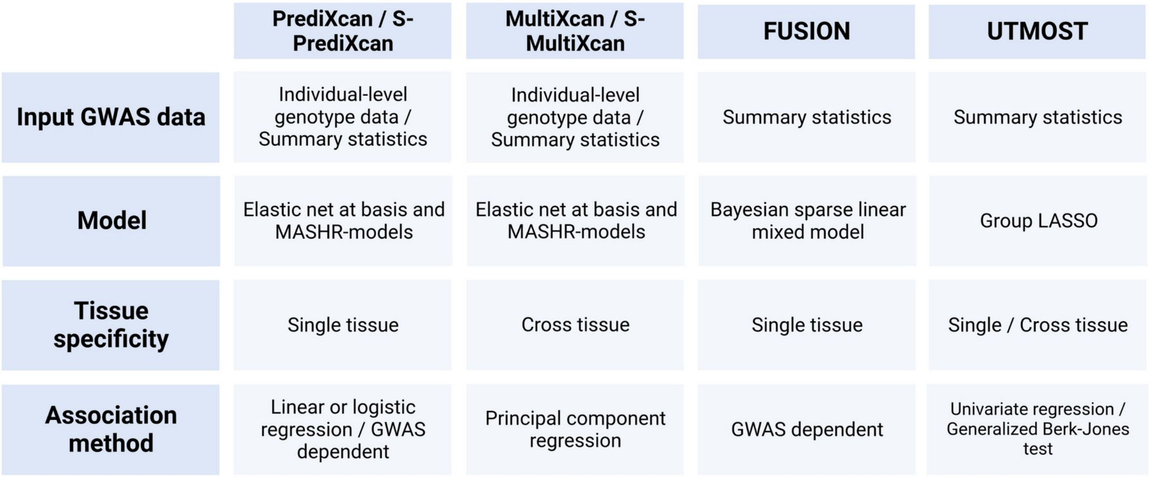


Figure 4: ***Molecular studies.*** *Top left is a schematic representation of an eQTL. Below, comparison of GWAS, eQTL and TWAS in terms of phenotypes. On the right, effects observed in molecular studies. Figure made in Biorender.*

To combine GWAS results and gene expression information to identify genes causal at a GWAS loci for a disease or a trait, methods such as Transcriptome-Wide Association Studies (TWAS) were developed. TWAS are a class of methods that leverage reference expression datasets to construct predictive models of gene expression using genetic information (Gamazon et al. 2015). These models are subsequently applied to data from GWAS studies to generate cis-predicted expression phenotypes for thousands of genes and all the individuals in the study, which include typically hundreds of thousands participants. Using this individual level predicted expression phenotypes, researchers can identify differentially expressed genes between cases and controls. In situations where individual-level genotype data is not available, approximations can be made using summary statistics from the disease-genetics associations. This approach shifts the variant-level analysis of a GWAS to a gene-level analysis. By doing so, it reduces the multiple testing burden from millions of tests (SNPs) to just a few thousands (genes). Because the expression phenotypes are derived from genotypes, a reverse causality where the disease effects the expression of the gene can be excluded, and identified genes are more likely to be on the causal pathway to disease, which is important when developing new therapies (Wainberg et al. 2019).



	PrediXcan / S-PrediXcan	MultiXcan / S-MultiXcan	FUSION	UTMOST
Input GWAS data	Individual-level genotype data / Summary statistics	Individual-level genotype data / Summary statistics	Summary statistics	Summary statistics
Model	Elastic net at basis and MASHR-models	Elastic net at basis and MASHR-models	Bayesian sparse linear mixed model	Group LASSO
Tissue specificity	Single tissue	Cross tissue	Single tissue	Single / Cross tissue
Association method	Linear or logistic regression / GWAS dependent	Principal component regression	GWAS dependent	Univariate regression / Generalized Berk-Jones test

Figure 5: ***Summary table of the different existing methods for TWAS.***

Since the emergence of TWAS, additional methods have been developed to conduct this type of study. Initially, PrediXcan was widely adopted to perform TWAS, followed by FUSION. More recently UTMOST has been released as an alternative method, in particular for cross-tissue analyses (Figure 5). The main differences between these methods lies in the input data, the method used to build predictive models and the statistical methods to calculate gene-trait scores (Gamazon et al. 2015; Mancuso et al. 2018; Rodriguez-Fontenla and Carracedo

2021). Depending on the data available, either individual-level data or summary statistics from GWAS can be used, and analyses conducted in a single tissue or a cross-tissue context. Using summary statistics is usually computational less heavy but can introduce noise because of the need to calculate LD matrices separately to account for the LD structure, which can be different from the actual structure of the cohort. LD, or linkage disequilibrium, is used when alleles at different positions on a chromosome are observed together more often than by random chance representing physical proximity.These regions of the genome where genetic markers are in high LD with each other are called haplotype blocks and need to be taken in consideration when studying any genetic associations with a trait. On the other hand, individual-level data is harder to access and takes more time but can provide more accurate estimates. Cross-tissue methods (MultiXcan, UTMOST) were developed to compensate for some of the single tissue caveats, such as shared eQTL effects across tissues and sample size limitations. However, with cross tissue studies, we lose the ability to identify the tissue-specific effects and increase the computational burden.

Although TWAS implicated genes are likely to be on the causal path to disease, some researchers argue certain assumptions of the methodology preclude TWAS from identifying true causal genes. In addition, even though TWAS methods use covariance tables to correct for the LD structure, results provided by TWAS are still subject to issues that affect the assumption of causality, such as colocalization and pleiotropy (Ndungu et al. 2020). Colocalization addresses that locally variants in the genome are often highly correlated: a GWAS variant and an eQTL in close proximity could be observed as a single effect, when actually they relate to distinct processes. Pleiotropy, in particular horizontal pleiotropy, relates to when genetic variants have multiple consequences and in this case the GWAS variant and the eQTL could lie on distinct causal paths, in particular when the eQTL affects multiple genes. The ability of TWAS methods to build predictive models of expression is also limited, in particular by the heritability of the molecular trait. Finally, predictive models are often based on *cis*-expression which only represent a fraction of the total expression, ignoring rare *cis*-eQTLs and *trans*-eQTLs and therefore miss an opportunity to increase the overall power to discover signals.

Mechanisms and causes of T2D

Type 2 diabetes (T2D), also known as non-insulin dependent diabetes mellitus, is a chronic metabolic disorder which affects populations worldwide. It is the most common form of diabetes, with almost 462 million individuals affected by the disease in 2017 (Khan et al. 2020), and this number increases every year. Due to the multiple complications that present with T2D, the disease carries a large burden, both in terms of death (1.37M in 2017) as well as an economic cost (1.32 trillion dollars in 2015 worldwide) (Bommer et al. 2018). The disease is characterized by elevated levels of glucose in blood, and is responsible for comorbidities such as hypertension, retinopathy, neuropathy, coronary heart disease, and kidney disease. T2D tends to develop later in life, and is strongly associated to genetic, lifestyle and environmental factors such as body fat. The combination of these factors may cause both impaired insulin secretion from the pancreas and failure of insulin to properly stimulate glucose intake in other tissues, known as insulin resistance (Cerf 2013).

Insulin is a peptide hormone secreted exclusively in the beta cells of the pancreatic islets in the pancreas. Pancreatic islets are around 2% of the raw tissue volume of the pancreas, and is composed by hormone producing cells such as beta cells, gamma cells, and others (Da Silva Xavier 2018). Translated from the *INS* gene, a preproinsulin protein is synthesized first and then modified into a proinsulin protein, before being packaged into vesicle for the last post translational modifications which result in a mature insulin protein. Insulin is then released in the blood, either oscillating over a long period of time, or released faster following an increase of blood glucose levels. Circulating insulin can have multiple effects on the overall metabolism, but its main role is to increase the glucose intake in organs such as adipose tissue and muscle (Rahman et al. 2021). In order to do this, insulin proteins bind to specific receptors on the cell membrane (insulin receptors), promoting cellular processes that transform glucose such as glycogenesis (the process converting glucose into glycogen), inhibition of gluconeogenesis and glycogenolysis (production of glucose in the liver) or stimulation of esterification and fat synthesis (conversion of glucose into triglycerides for nutrition). Because insulin plays such a major role in the glucose homeostasis, any disturbance in the synthesis or release process could lead to failures in reducing glucose blood levels and have consequences for the metabolism of the whole body.

In addition to impaired secretion of insulin, insulin resistance also contributes to the disruption of the glucose homeostasis in tissues targeted by insulin (skeletal muscles, adipose tissue and liver) and responsible for the management of blood glucose. The glucose transporter gene (*SLC2A4)* is responsible for producing the intracellular glucose transporter (GLUT4) that captures glucose from intercellular spaces in adipose and skeletal muscle tissues. Other transporters such as GLUT1 and 2 have similar functions but in different tissues (liver, pancreas, etc...). Mutations in this gene can affect the glucose transport and lead to consequences on metabolic processes such as glycogen and lipid synthesis, resulting in hyperglycaemia or hyperlipidemia (higher levels of glucose and lipids in the blood) (Czech 2017; Pearson et al. 2016).

Although traditionally T2D is thought to be the consequence of poor glucose homeostasis, the complex genetic and environmental factors known to define disease risk suggest T2D is a consequence of multiple factors interacting together. For example, T2D can be caused by a combination of high BMI due to a high caloric diet together with a sedentary lifestyle behaviour, which causes insulin resistance. These factors cause the development of inflammation and stress/oxidative stress in response, triggering the slow degradation of pancreatic islets (Galicia-Garcia et al. 2020; Cerf 2013). This is usually a slow process involving many biological pathways in multiple organs and tissues separate from the pancreas. After the development of insulin resistance (IR), beta cells first try to compensate for the reduced glucose intake by increasing insulin production and release in pancreatic islets, inducing cellular stress and triggering inflammation processes. This state, after a long period of time, can lead to the exhaustion of the beta cells, also called glucolipotoxicity, and ultimately cell death (Christensen and Gannon 2019; Halban et al. 2014). Therefore, impaired secretion of insulin can be directly associated with increased insulin resistance. This process, by which the pancreas compensates for the high blood glucose level caused by the disruption of insulin absorption by secreting insulin in excess is called hyperinsulinemia, and it affects insulin sensitivity in other tissues. IR itself is considered highly environmentally dependent as it is strongly correlated with high fat diets and lifestyle. But many genetic factors are also involved in causing IR, including mutations of the *GLUT4* transporter in muscle, and *Foxo1* and Akt protein kinase pathways in liver (Galicia-Garcia et al. 2020; Pearson et al. 2016; Czech 2017).

As a result of the impact of IR , T2D can be described as a multi-tissue disease, with its pathogenesis involving various organs. From a causal point of view, due to its role in insulin secretion, pancreas, and more specifically the pancreatic islets, may be considered the main tissue in the development of T2D. However, insulin resistance, and more generally blood glucose management, can be attributed to a wide range of tissues. For example, adipose tissue, through the release of inflammatory molecules and adipokines, can disrupt insulin sensitivity. In addition to adipose tissue, liver and skeletal muscle are key components of glucose release and consumption; when insulin receptors are inhibited in these tissues the whole body glucose homeostasis is disturbed. This is the reason why these tissues are often studied in T2D studies, to explore the different mechanisms responsible for the disease pathogenesis.

Diagnosis and medication

T2D is often diagnosed by measuring the levels of glucose in fasting individuals’ (usually 6 to 8 hours after the last meal) plasma blood and the levels of glycosylated haemoglobin A1c (HbA1c). Levels of glucose greater than 126 and 200 mg/dL or HbA1c levels greater than 42mmol/mol are indicators of diabetes. Oral Glucose Tolerance Test (OGTT), another test, measures the metabolic response before and after a glucose intake. This test also can identify both impaired glucose tolerance (IGT) and impaired fasting hyperglycaemia (IFG), indicators of a prediabetes status, where blood glucose levels are still lower than what would be observed for a patient with diabetes but potentially in the path of disease. Additionally, other clinical metrics can be used to more precisely define a patient’s condition, in terms of disease severity and subtypes. For example, insulin c-peptide test can be used to measure the ability of the body to produce insulin by quantifying the amount of c-peptide, a cleaved section of the proinsulin, secreted to blood by pancreatic islets (Leighton, Sainsbury, and Jones 2017). Lower levels would suggest treatment supplementing insulin using injections. Glucose clamp methods measure glucose continuously over a period of time in a hyperglycaemic context to establish an individual’s ability to secrete insulin and absorb glucose (DeFronzo, Tobin, and Andres 1979). Similarly, homeostasis model assessment (HOMA) methods can be used to measure insulin resistance (HOMA-IR) or beta cell dysfunction (HOMA-B) using measurements of insulin levels and fasting glucose (Song et al. 2007).

Environmental and genetic factors define the risk of developing T2D. BMI and sex are among the most influential environmental factors on T2D risk. Physiological differences between males and females, such as body fat distribution and sex hormones, have been associated with disease risk. Obesity, defined as a high BMI, is a strong risk factor associated with the development of insulin resistance by increasing the deposition of adipose tissue and visceral fat, promoting inflammation processes and modifying the metabolic micro-environment of cells with consequences for insulin signalling pathways. In addition, high caloric diet, sedentary lifestyle, regular stress and low physical activity are other risk factors, influencing in particular the age of onset of T2D. These factors are confounded with ethnicity and the specific frequency of alleles in a population, which may influence how a particular diet or lifestyle defines the risk of disease. For example, in Western populations the peak incidence of newly diagnosed diabetes is between 55 and 59 years old, but in Indian populations this peak happens around 40 years old, even in those individuals with Western diets and life styles (Lin et al. 2020; Khan et al. 2020). Similarly, although higher BMI is a risk factor for T2D, women show a higher tolerance for BMI than men, with Indian populations also developing T2D with lower BMIs than European populations. The combination of these factors is one reason why T2D is such a heterogeneous and complex disease and also stresses the importance of building a patient profile to have a more adapted prevention system and treatment.

The complexity of disease mechanisms and aetiology make it difficult to consider all patients with T2D equally in terms of diagnosis and treatments. More recent studies have sought to characterize individuals diagnosed with T2D using clinical variables to define new subtypes of T2D. One suggested classification was first introduced in Ahlqvist et al (Ahlqvist et al. 2018; Ahlqvist, Prasad, and Groop 2020; Mansour Aly et al. 2021) and proposed five subgroups of T2D patients using a model based on six clinical parameters (glutamate decarboxylase antibodies, age at diagnosis, BMI, HbA1c, and homoeostatic model assessment 2 estimates of β-cell function and insulin resistance). They defined five distinct subtypes with specific disease progression and risk of complications : 1) Severe autoimmune diabetes (SAID), which includes patients with early-onset diabetes, low BMI, insulin deficiency and presence of GADA (glutamic acid decarboxylase antibody, an antibody acting against the pancreatic cells, this subtype has a large overlap with Type I diabetes) ; 2) Severe insulin-deficient diabetes (SIDD), includes patients similar to the previous group but without GADA antibodies; 3) Severe insulin-resistant diabetes (SIRD) with individuals presenting high BMI and high insulin resistance ; 4) Mild obesity-related diabetes (MOD), patients with obesity but no insulin deficiency ; 5) Mild age-related diabetes (MARD), similar to MOD except a higher overall age distribution.

Treatment of T2D is often based a combination of glucose lowering drugs, behavioural interventions, such as diet management and increased physical activity, and medication for comorbidities such as retinopathy and cardiovascular diseases. Among the glucose lowering medications, the most widely used is metformin (Inzucchi et al. 2012). Its main mechanism of action consists of reducing hepatic glucose production which leads to weight loss without hypoglycaemia. Side effects of the drug including gastrointestinal troubles, and it is contraindicated for patients with CKD or acidosis. Other commonly used drugs include the sulfonylureas, a class of drugs that stimulate insulin secretion by acting on potassium channels in beta cells. A common side effect of these type of drugs is hypoglycaemia and weight gain. Similar to sulfonylureas, drugs based on GLP-1 receptor agonists stimulate the pancreas to increase the insulin secretion and also slow down the absorption of glucose in the intestine. One of the main consequences of this drug is weight loss caused by the decrease of appetite as well as slowing the gastric emptying, and these drugs are also currently being explored as a general weight loss medication (Nauck et al. 2020).

Thesis aims

The aim of this thesis is to use the different computational methods to explore the variety of causes and also the consequences of T2D as a complex disease. In the first chapter we will explore the consequences of T2D on gene expression using a multi-tissue expression dataset. The second chapter will explore TWAS methods as applied to T2D, to propose causal genes for the disease. Finally, we will look in more detail at metformin, using gene expression and genetic studies into the drug to explore the genes which affect its efficacy, and the direct molecular consequences of the medication.

The data

This thesis is a computational project which relies heavily on different datasets to conduct various analyses. We used data from two consortia, the Genotype Tissue Expression (GTEx) consortium and the DIabetes REsearCh on patient straTification (DIRECT) project (Aguet et al. 2017; Koivula et al. 2014). The first is a multi-tissue American cohort, created to investigate the relationship between genetic variation, gene expression, and phenotypic traits. The second is a European consortium focused on T2D and using whole blood samples to identify biomarkers of the disease in order to improve the understanding and treatment of T2D.

GTEx

The GTEx project was first unveiled in 2013 (Lonsdale et al. 2013) as a collaboration between multiple human biology and health centres in the United states (NCI, NCBI, LDACC). The aim was to produce a comprehensive resource to understand how genetic variation influences gene expression across multiple tissues in the human body. For this, they produced a multi-tissue cohort aiming to detect tissue-specific QTLs which could be associated with multiple traits and diseases. Data was released in several stages, starting with a pilot release after ~2.5 years built from 9 tissues and 175 donors. The final release included 49 tissues and more than 800 individuals (Figure 6).

After the selection of the donors based on specific criteria (aged between 21 and 70, with exclusion criteria for HIV, viral hepatitis, metastatic cancer, chemotherapy and radiation therapy for any condition within the past 2 years or a BMI > 35 or < 18.5), organs were sampled and prepared for the next steps, and questionnaires were given to family members to collect phenotype and disease information. Then, after quality control, they extracted DNA and RNA and proceeded to genotyping and producing expression quantifications using Illumina technologies.

Their first major publication in 2015 (GTEx Consortium 2015) provided an overview of the project's goals, methodology, and initial findings. They reported the eQTLs they discovered, describing how these eQTLs were shared across tissues, reported patterns of allele-specific expression as well as relating the eQTLs to functional annotations of the genome produced by ChIP-seq experiments. After this, there were two other main releases of data, accompanied by major publications. One in 2017 was based on the v6p version of the data (Aguet et al. 2017), and described in more detail the sharing of eQTL signals and the functional characterization of these QTLs.

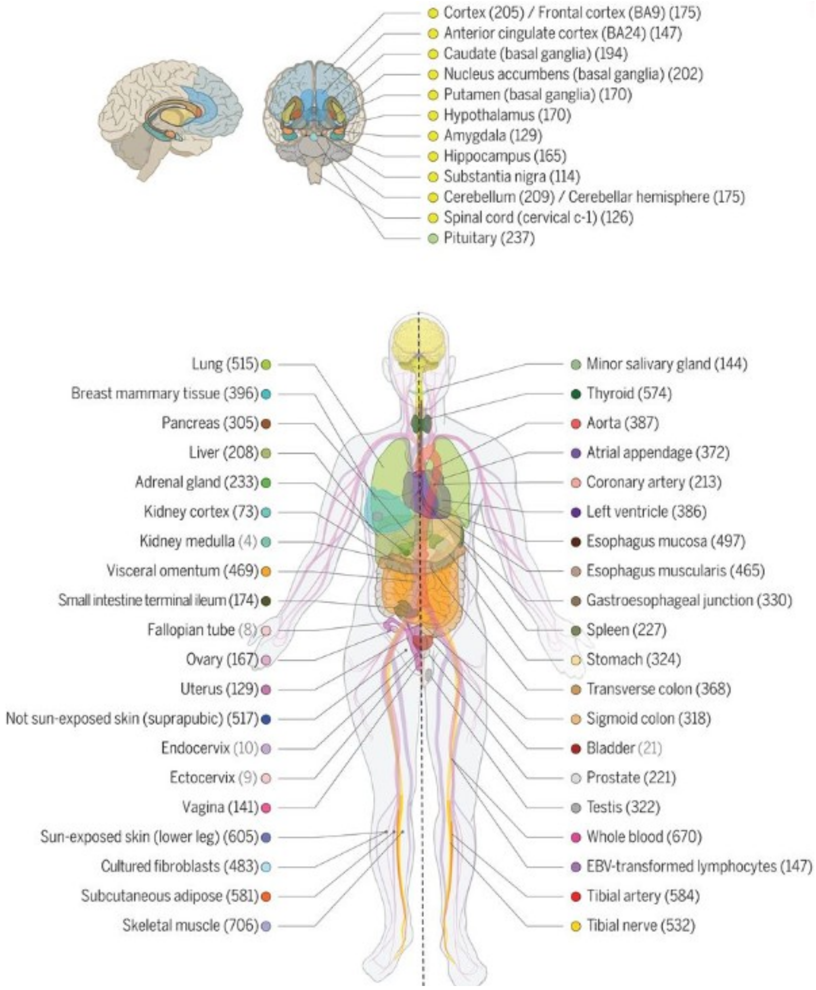


Figure 6: ***Tissues and numbers of samples per tissue included in the GTEx v8 release.*** *Figure from GTEx , 2020*

Finally, in 2020, the full v8 dataset was released, with a full discussion on tissue specificity and sex differences for regulatory effects (The GTEx Consortium 2020). This final release includes expression quantified in 49 tissues from 838 donors, released together with whole genome sequencing of the individuals. In total, mRNA was sequenced from 15,201 tissue samples. The number of samples ranges from 73 for kidney cortex samples up to 706 for skeletal muscle. In total they found 4,278,636 genetic variants were significantly associated with at least one gene in at least one tissue, 23,268 protein coding or lincRNA genes had at least one cis-eQTL and mapped a total of 695,981 independent cis-eQTLs across all genes and tissues (Figure 7).

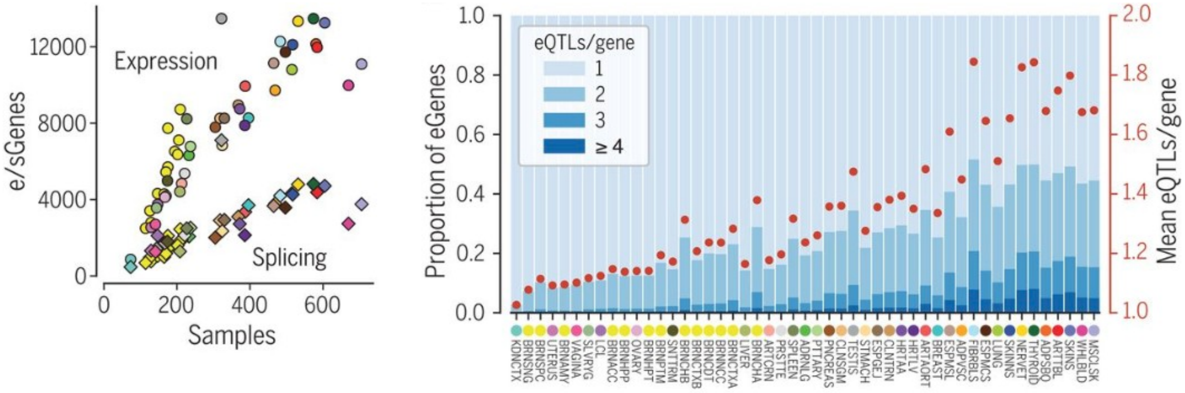


*Figure 7: **Panels from the final release of the GTEx v8 data describing the major findings and reproduced from ref. A Number of genes with a cis eQTL per tissue vs sample size. B Proportion of eGenes with one or more independent cis-eQTLs and mean number of cis-eQTLs per gene.** Tissues ranked by sample size. Figure from GTEx, 2020.*

From GTEx, we used publicly available independent eQTLs to build models that predict expression from genotype data, to associate predicted expression with phenotypes from other cohorts. We also used the gene expression data and the fact that around 160 of the GTEx donors had been diagnosed with T2D to look for genes whose expression was associated with the disease.

DIRECT

The DIRECT consortium was created in 2012 to better understand the development and progression of T2D in patients. In order to shape prevention and treatments for patients with T2D and those at risk of developing the disease, the project had two main goals: 1) observe and understand the variation of glycaemic control in patient with T2D and in participants with a pre-diabetic status over time and 2) quantify the effect of treatment on the glucose homeostasis, enabling a more personalised approach to treatment.

In order to accomplish this, they recruited two cohorts of individuals, one with newly diagnosed diabetes and the other with elevated HbA1c that did not reach the level of a clinical diagnosis of diabetes (referred to as pre-diabetes). These cohorts were recruited across 7 different centres in Europe (Kuopio, Amsterdam, Copenhagen, Lund, Dundee, Exeter and Newcastle). They excluded individuals with prior treatment with insulin or other drugs than metformin, patients with pacemakers or any other significant medical reason for exclusion, patients with T2D, a BMI < 20 or > 50 kg/m2 and HbA1c level higher than 75 mmol/mol. These cohorts, a total of 3029 individuals, were followed up over a period of three to four years, with three clinical visits over this period. Whole blood was taken during these visits, and used to produce genetic, transcriptomic, proteomic and metabolite data (Figure 8). Genotyping was done using the Illumina HumanCore array. For expression, mRNA samples were sequenced using the Illumina HiSeq2000 platform, mapped to the GRCh37 reference genome, and exon and gene quantification were produced using a pipeline from Delaneau et al. 2017. Splicing phenotypes were also generated using LeafCutter. Proteomics data was produced using five different models using the Olink® platforms, namely Cardiometabolic, Cardiovascular II, Cardiovascular III, Development and Metabolism panels.

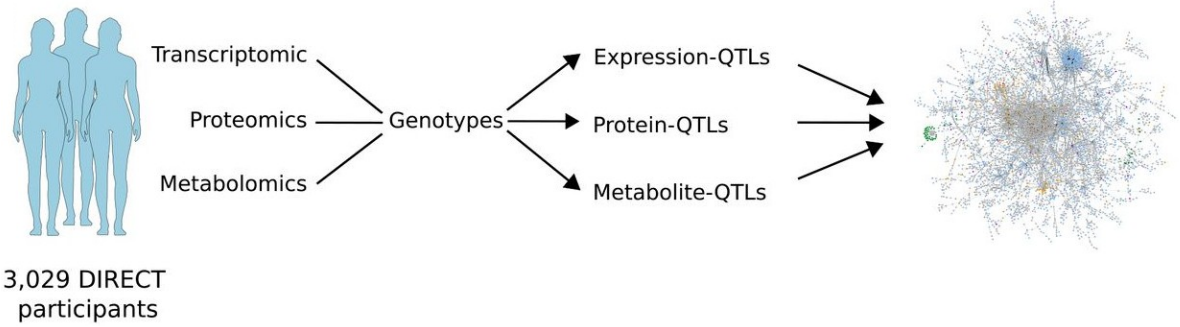


Figure 8: ***Visual representation of the study framework from collecting the data to the analyses of molecular phenotypes.*** *Figure from Viñuela et al 2021*

One recent paper published by this consortium was Brown et al, which mapped QTLs across all omics data types available in DIRECT. They found extensive allelic heterogeneity, where a molecular trait is affected by multiple genetic variants, and also pleiotropy, where a genetic variant affects multiple molecular traits. They also showed that many of the genetic signals across the diverse tissues of GTEx were also active in whole blood (Figure 9); we will use this fact in Chapter 2 to identify T2D genes using a non relevant tissue.

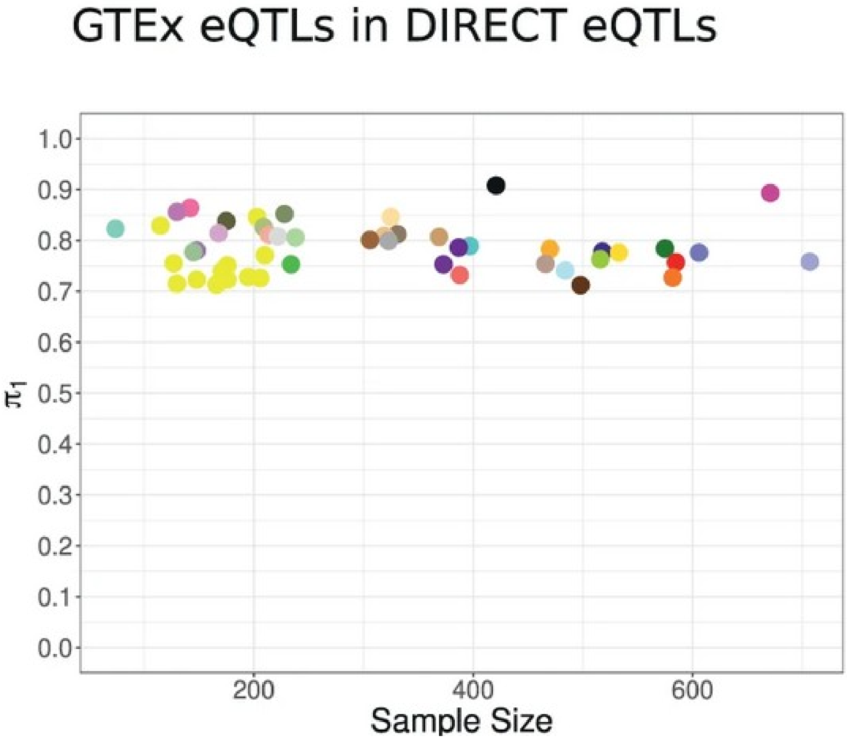


Figure 9: ***π1 values which represent the proportion of genetic effects in each of the GTEx tissues which are also found to be active in DIRECT whole blood.*** *Figure from Brown et al 2023.*

Chapter 1 : Effect of T2D on gene expression in a multi tissue context

Type 2 diabetes as a complex disease

Type 2 diabetes (T2D) is a complex and multifactorial metabolic disorder that affects millions of individuals worldwide. It is characterised by chronic hyperglycaemia resulting from a combination of insulin resistance and impaired insulin secretion (Roden and Shulman 2019). The rising prevalence of T2D presents a significant global health challenge, with its associated complications contributing to increased morbidity and mortality rates. While environmental and lifestyle factors play a crucial role in disease development, it is increasingly evident that genetic factors also contribute significantly to T2D susceptibility and progression (Galicia-Garcia et al. 2020).

Over the years, significant progress has been made in understanding the genetic basis of T2D, with numerous genetic loci and variants associated with disease susceptibility identified through genome-wide association studies (GWAS) (Mahajan et al. 2018; Spracklen et al. 2020; L. Cai et al. 2020). However, the functional consequences of these genetic variants and their impact on gene expression patterns in relevant tissues remain incompletely understood because the majority of genetic associations fall into non-coding regions of the genome. It is therefore difficult to link the disease-associated variant to a particular gene or protein. Whole genome sequencing can help by producing high quality information on all non-coding variants, but fine mapping the exact causal variant and determining its consequences still presents significant challenges (Brown et al. 2017). One approach that has been used is to combine GWAS information with epigenomic information collected from relevant cells or tissues. For example, Weedon et al. used annotations on whole genomes extracted from patients with isolated pancreatic agenesis to discover several mutations affecting the enhancer activity of the *PTF1A* gene, a critical developmental gene and therefore associating these mutations to isolated pancreatic agenesis, a direct cause of neonatal diabetes. However, little is still known about the causes and consequences of the disease.

There are type of diabetes where the pathogenesis scenario follows the example of a monogenic disease, where a single gene in a single tissue is directly causal for the development of the disease. Neonatal diabetes and maturity onset diabetes of the young (MODY) are examples of families of monogenic diabetes, but even there, that specific gene may differ between individuals with the disease (Fernandez-Zapico et al. 2009; Thanabalasingham and Owen 2011). T2D in contrast, is an example of a complex disease. It is the result of many genetic and environmental factors affecting processes such as gene expression, in several tissues such as pancreas, adipose tissue, and skeletal muscle. In individuals with the disease, we could expect to see differences in gene expression in individuals with T2D, in these tissues.

Differential gene expression and T2D

Differential gene expression analysis as been used as a complement to genetic studies to identify genes involved in diseases and its consequences. Discovering genes whose expression is different (up- or downregulated) in individuals with the disease may implicate these genes in the development, progression or consequences of the disease. With the advent of high-throughput technologies to measure thousands of molecular phenotypes, such as RNA-seq, we are now capable to quantify the expression of thousands of genes in cells and tissues (Wang, Gerstein, and Snyder 2009). It is also possible to apply these technologies in a wide range of tissues, including those directly relevant to T2D, and increasingly in single cells as well (allowing us to extract information specific to a cell type). By utilising these large-scale genomic datasets, researchers can identify and quantify changes in gene expression with increased accuracy and sensitivity.

Discovering differentially expressed genes typically involves several key steps. First, biological samples from different conditions or experimental groups are collected and processed to generate sequence data from RNA from expressed genes. The output of the sequencing machine is mapped to the human genome, and the expression of individual genes is quantified by counting reads that map to that particular gene (Birney et al. 2007). Preprocessing steps are applied to filter out low-quality or badly mapped data and to normalise gene expression values across samples to account for technical variations. Once gene quantification data has been obtained, statistical techniques can be applied to analyse gene expression differences between conditions. Given these analyses can look at tens of thousands of genes in a single study, multiple testing corrections must be used to control for false-positive results. Once a set of genes has been discovered to be differentially expressed between conditions, functional enrichment analysis can be performed to gain insights into the biological processes, molecular pathways, or cellular components that are associated with these genes. These analyses can help to understand the underlying biological mechanisms observed by the gene expression changes.

There are a number of methods available to conduct differential gene expression analyses and the choice of the methods is often bound to the type of data used and which questions are asked. For RNA-Seq data, three main methods are widely used in the field (S. Liu et al. 2021) :

- DESeq2: this utilises a negative binomial distribution model to estimate gene expression variation between conditions and perform statistical tests for differential expression (Love, Huber, and Anders 2014).

- EdgeR: Similar to DESeq2, edgeR uses a negative binomial distribution model to estimate gene expression variation. The differences will be in the normalization procedures (scalling for DESeq2 and TMM or quantile normalization for edgeR) and the test to calculate differential expression (GLM for DESeq2 vs exact test for edgeR). Therefore edgeR is usually more fitted to smaller sample sizes (McCarthy, Chen, and Smyth 2012).

- limma-voom. This method combines a voom transformation to normalize the data with a linear modelling approach to analyse RNA-seq data. It provides a framework for handling technical variability using covariates in linear models (Law et al. 2014).

We have chosen to follow the limma-voom framework and work with linear regression models based on quantile normalised gene expression counts. This approach is being widely used in genetic associations analyses such as GWAS and expression quantitative trait loci (eQTL), which can be viewed as a type of differential expression analysis where the condition being research is the particular SNP genotype. The reason not to use the above mentioned packages is that recent work has suggested they produce false positives, especially in datasets as large as the ones we are analysing (Y. Li et al. 2022). Moreover, these methods are often designed for small sample size studies, while population studies, as those used through this thesis, tend to have larger datasets.

In the context of T2D, RNA-Seq and differential gene expression analyses have been used to better understand the pathogenesis of T2D by exploring different mechanisms and potential pathways particularly affected in individuals with T2D. For example, studies have shown reduced expression of genes responsible for fatty acid synthesis and mitochondrial function, along with increased expression of genes associated with innate immunity and transcriptional regulation in the adipose tissue of obese diabetic individuals (Debard et al. 2004; Lackey and Olefsky 2016). In addition to lipid metabolism, immune response and inflammation are important factors altered in individuals with T2D and several studies have identified differentially expressed genes involved in immune cell signalling, inflammatory response, and T-cell activation in individuals with the disease. The NF-kB signalling pathway has also been found to be activated, and the expression of genes associated with immune response and antigen presentation was found to be altered in T2D patients showing another consequence of inflammation and oxidative stress (D. Cai et al. 2005). However, it could be that changes in expression are a result of the disease rather than a cause. Indeed, we are observing correlations between changes in expression and disease status and not necessarily causation (Porcu et al. 2021). In this case, the gene may still be relevant to the symptoms produced by the disease or its progression, and thus still would be interesting drug targets for treatment rather than prevention.

Hypothesis and aims

In this chapter we aim to identify differential gene expression in response to T2D in multiple human tissues to investigate what tissues respond to the development of T2D. The identification of specific tissues which genes respond to the development of T2D would inform about the underlying molecular mechanisms that react to disease such as insulin resistance, in the case of involvement of adipose or skeletal muscle tissue, and beta cell dysfunction in the case of the liver and pancreas involvement. Moreover, by looking in particular tissues we can explore the role of comorbidities associated to the disease. The risk of T2D is known to increase with age, BMI and for smokers. Therefore, the identification of differentially expressed genes must consider whether changes in expression are related to T2D, or other environmental factors such as obesity or older age. In addition, integrating data across multiple tissues can enable the identification of shared and tissue-specific signals to understand interactions between tissues. Overall, integrative approaches can provide a comprehensive view of disease mechanisms, highlight tissue-specific contributions, and facilitate the identification of potential therapeutic interventions.

Throughout this first chapter, we have two major questions :

- How does type 2 diabetes alter gene expression profiles across various tissues and how do these changes contribute to the pathogenesis and progression of the disease?
- How does the gene expression profile differ between relevant and non relevant tissues in individuals with type 2 diabetes compared to non-diabetic individuals, and what are the key genes and pathways involved?

GTEx data and construction of the model

In order to conduct a differential gene expression analysis (DGEA), we used expression and phenotyping data publicly available from the Genotype-Tissue Expression (GTEx) consortium (Lonsdale et al. 2013). GTEx was a large-scale research project initiated in the United States aimed to create a comprehensive catalogue of gene expression and the effects of genetic variation on gene expression across multiple human tissues from cadaveric donors. In this study we used the expression and phenotyping data from the v8 version, using the 49 tissues selected for cis eQTL mapping, as these had reasonable sample sizes (N > 70) with sufficient power to discover T2D genes.

The GTEx cohort included individuals between 20 and 70 years old (average age 53 years old) and with a BMI between 17 kg/m2 and 40 kg/m2 (average of 27.22 kg/m2). The distribution of men and women is for most of the tissues 2:1, except for sex-specific tissues (in total 653 males and 327 females). There were a total of 218 individuals with T2D in the cohort, with significantly (p = 2.1e-07) higher BMI than controls (Figure 10).

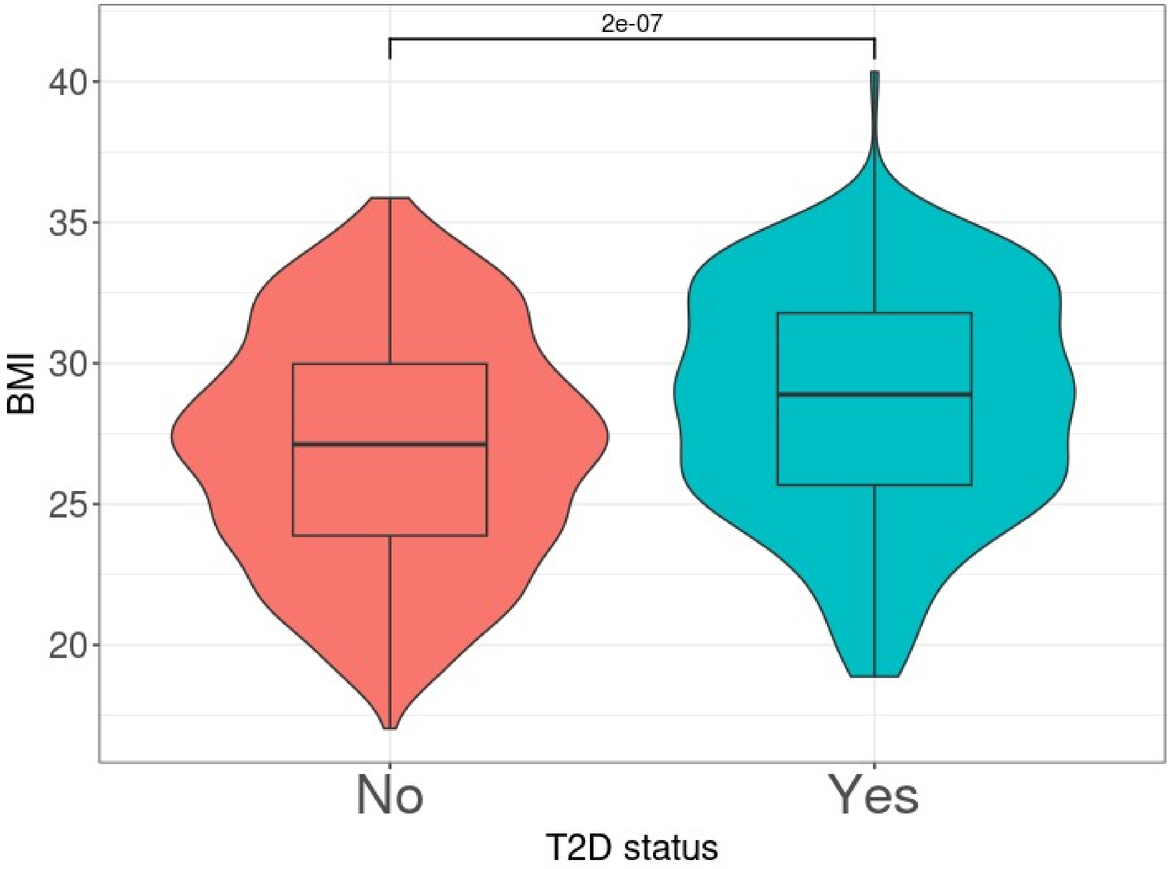


Figure 10: ***Distribution of BMI for individuals in GTEx by T2D status. BMI was significantly higher in individuals with T2D.*** *Mean BMI for controls = 26.9 and mean BMI for cases = 28.6.*

Main results

The GTEx study performed most of their analyses controlling for both biological and technical sources of variation in expression data using principal component analysis (PCA). However, PCs are known to capture important biological sources of information as well, so using this approach risks removing signals related to T2D. Therefore, we looked at the technical information provided by GTEx, related to sample collection, processing, and experimental procedures, to evaluate which of these explained a significant proportion of the expression variation. By considering these technical covariates, we also increase our statistical power to discover true associations as we reduce the noise in the data. Moreover, we also evaluated the effect of biological variables on genes expression in the context of T2D by looking at the effect of age, BMI and sex. First, we used a linear model to identify associated variables with the first 5 PCs of the expression data using a Wald test. In cases where multiple variables were used to capture one technical source of variation (for example, where multiple PCs were used to summarise genotype PCs or for categorical variables such as ethnicity, PCR and platform) we used a likelihood ratio test applied to nested models. To decide which covariates to include in the final DEGA regression model, we constructed a heuristic score for importance, based on the p value for association between the covariates, each of the 5 principal components, and the proportion of variance in expression explained by that covariate. This score was calculated for each covariate by summing the product of the -log10 p value with the regression coefficient from the five linear regressions between the PCs from gene expression and each of the covariates. Based on these scores, we decided to use age, sex, ethnicity and BMI as biological covariates, and RIN score, ischaemic time, autolysis, platform and pcr as experimental covariates (Figure 11).

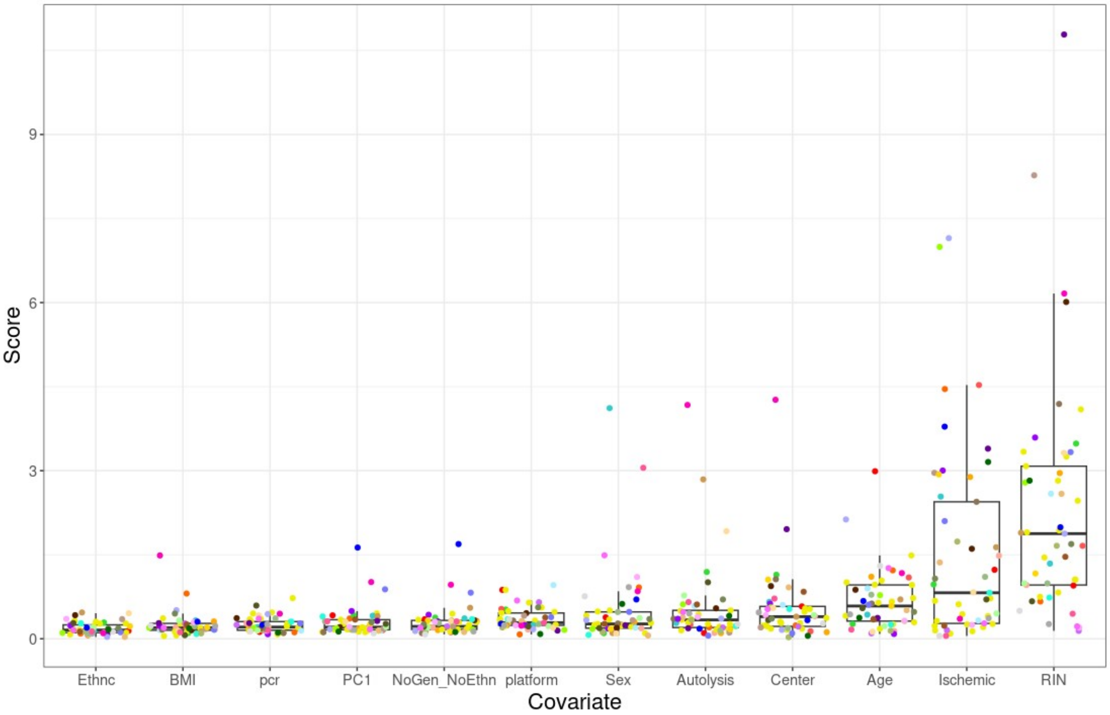


Figure 11: ***Scores calculated for each covariates to include in the regression model.*** *Each dot is a GTEx tissue. Variables are ranked by increasing mean score.*

Using a linear model controlling for technical variables, age, sex, ethnicity and BMI, we identified between 1 to 5174 differentially expressed genes (DEGs) for T2D across 49 tissues after multiple testing correction per tissue (FDR 5%) (Fig 12). Significant DEGs were found in 12 of the 49 tissues, but there was no particular relationship between the relevance of the tissue for T2D and the number of significant DEGs (Figure 13). The larger numbers of DEGs were found in the tibial nerve and the basal ganglia of the brain (respectively 5174 and 4873 genes). A recent study has highlighted the importance of this tissue for T2D (García-Pérez et al. 2022). T2D can lead to damage to the nerves, a condition known as diabetic neuropathy. Moreover, although the majority of these 12 tissues had a higher than average sample size (n > 310), there was no significant correlation between the sample size and the number of DEGs found (cf next section).

We also investigated DEGs for four other biological and technical factors, namely BMI, sex, age and ischaemic time (the time from death to tissue collection) for their high impact on T2D risk. We found more DEGs for these factors than we did for T2D mainly because GTEx is not a T2D cohort and the number of cases was not enough to detect more signals. With 15343 genes significantly associated with ischaemic time in lung tissue and 14319 genes significantly associated with sex in breast mammary tissue. About the number of tissues with significant DEGs, in comparison to the 12 tissues with DEGs for T2D, we observed a maximum of 47 tissues with genes significantly associated with age, 44 with sex, 34 with ischemic time and 31 with BMI. This gave us insights into the importance of these factors considering the number of DEGs discovered for them but also the importance of tissue relevance as we can observe different ranking of tissues in terms of discovery power depending on the trait we are looking at.

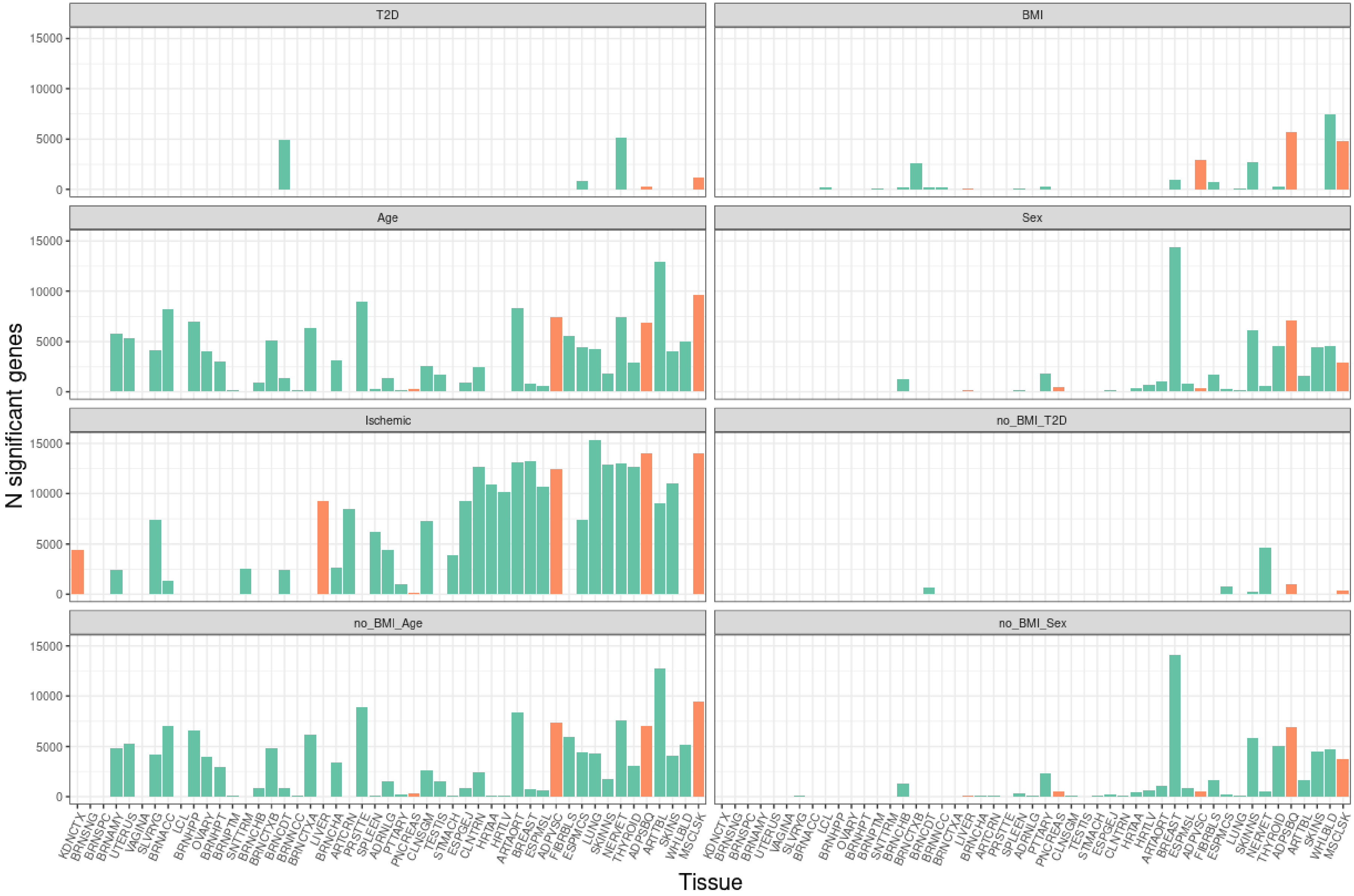


Figure 12: ***Number of significant genes for each tissue and each trait.*** *Colors are green for tissues non relevant for T2D and orange for relevant tissues.*

Correlation between discovery power and sample size

For all biological factors except T2D, the number of samples collected per tissue was an important determinant of the number of significant associations discovered. The linear relationship between sample size and power to discovered DEG was stronger for sex and ischaemic time (Figure 13). However, across tissues of similar samples sizes, such as the brain tissues (129 to 209), we would still observe variability in power to discover DEGs (from 0 to 8,176 significant DEGs across all traits). This linear relationship between sample size and number of significant associations was also weaker than the relationship with numbers of cis-eQTLs per tissue identified before (The GTEx Consortium 2020).

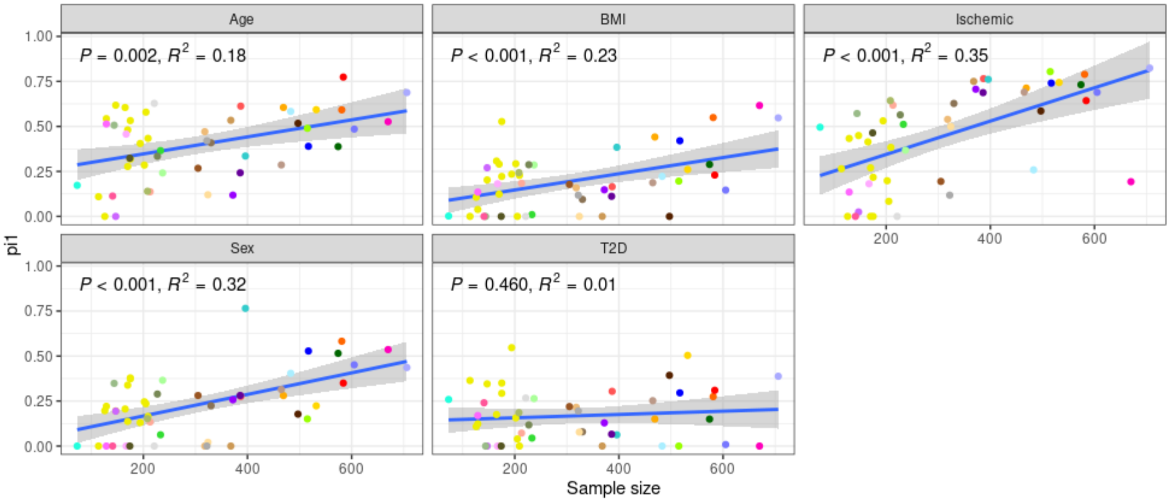


Figure 13: ***Proportion of significant associations against the sample size for traits without BMI correction.*** *Regression line in blue and coefficients above.*

We hypothesised that the lack of relationship between sample size and T2D DEGs may be driven by differences in the number of individuals with T2D for each tissue. However, a linear regression between number of DEG and the number of samples from individuals with T2D found no significant associations (p value = 0.702). Moreover, the association analysis identify brain caudate ganglia and tibial nerve were outliers (Figure 14), as both show high numbers of associations with T2D with low population of cases (respectively N = 47 and 105). This raises the possibility that many of these associations in the tissue could be driven by a few outliers. All in all, larger tissues identified more DEG per trait with smaller sample sizes showing high variability in the number of significant genes.

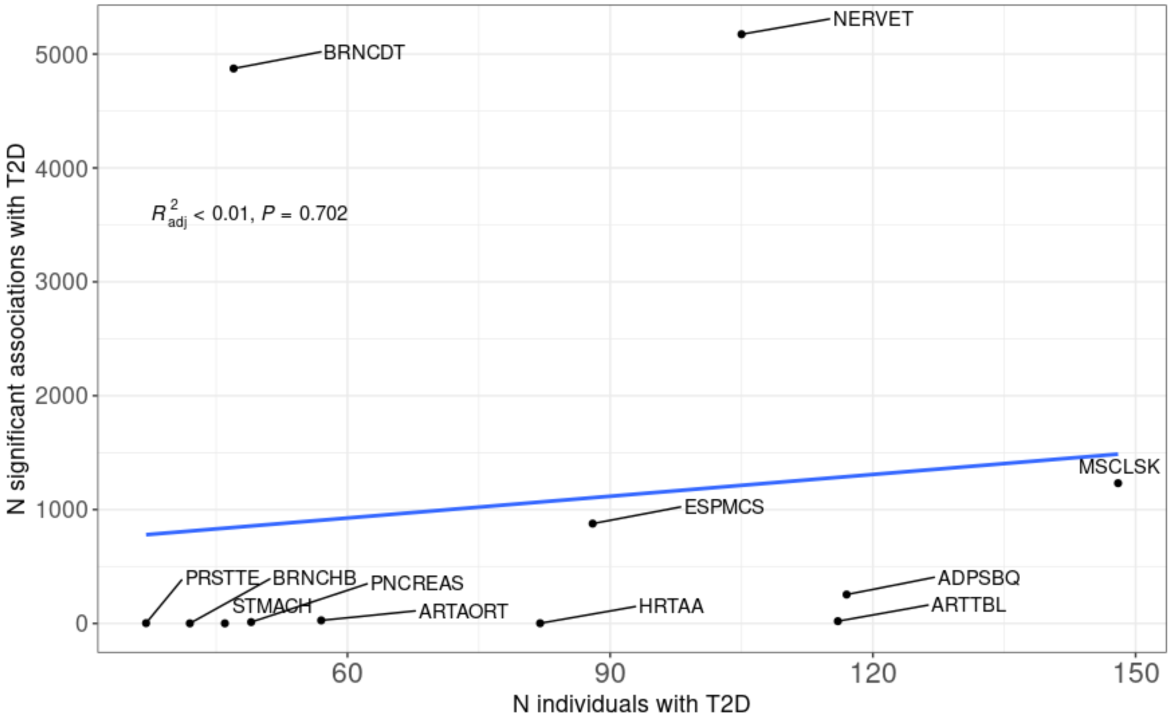


Figure 14: ***Correlation between number of individuals with T2D and number of significant associations.*** *Blue line represent the regression line with the statistics indicated on the graph.*

Degree of sharing between tissues

We next investigated the extent to which the same genes were repeatedly found to be differentially expressed in multiple tissues. Genes found to be differentially expressed in nerve tibial tissue were often found in one or more of 7 other tissues, with the highest degree of sharing between tibial nerve and adipose subcutaneous, aorta artery and heart atrial appendage (respectively 53%, 52% and 50% of significant genes in common) (Figure 15). Lower levels of sharing were observed for other tissues, with no clear pattern of sharing in related tissues. We also observed in some cases genes shared between multiple tissues but with different directions of effect. For example, *SLC13A3,* previously found to be involved in T2D, showed lower expression in the tibial nerve in individuals with T2D, but higher expression in subcutaneous adipose tissue.

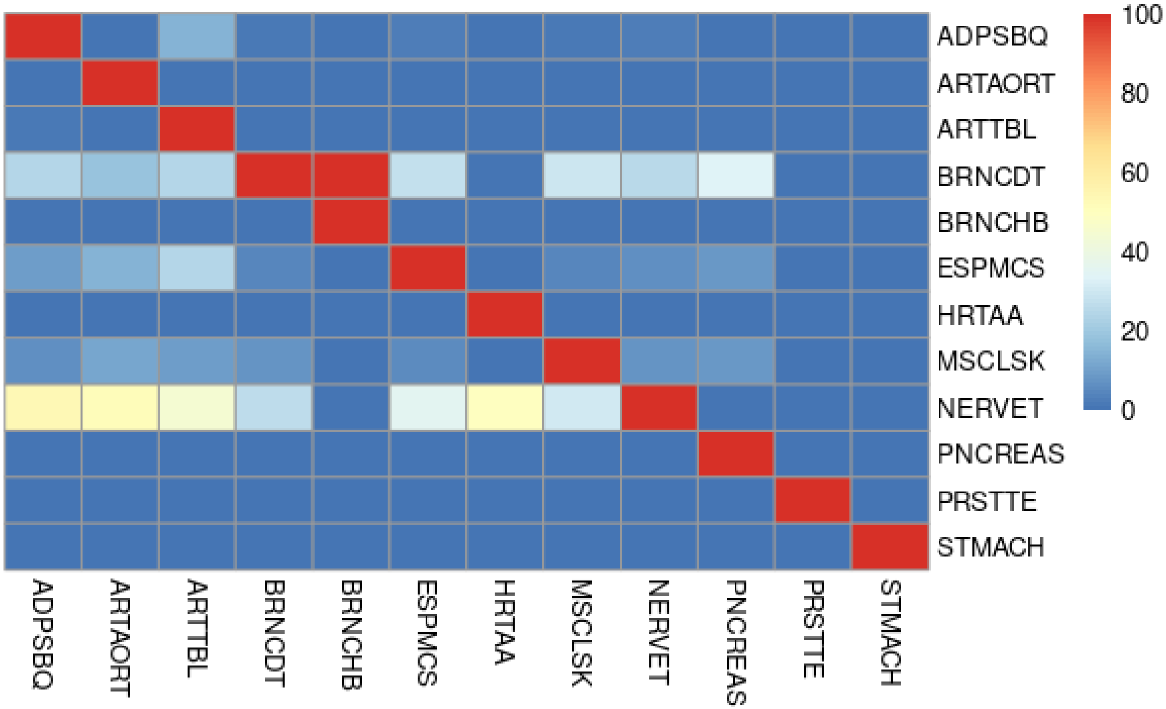


Figure 15: ***Heatmap of percentage of sharing between tissues with T2D signals.*** *From blue indicating no sharing to red 100% of significant genes shared.*

BMI effect on differential gene expression results

In the original model we included BMI as a biological covariate because obesity is known to be one of the major risk factors for the development of T2D. Controlling for BMI allows for the discovery of any T2D genes whose relationship is not mediated by obesity, just as controlling for age identify T2D genes whose relationship is independent of the age of the participants. Models not controlling for BMI, are therefore expected to identify more genes associated to T2D, as these would include T2D genes independent of obesity addition to those T2D genes associated to obesity. However, we found that by removing BMI from the model fewer genes were significantly associated with T2D in six tissues, while more genes were significantly associated in other seven other tissues (Figure 17). One of the larger changes was observed for the adipose subcutaneous tissue where 75% of DEGs were not significant after controlling for BMI (from 1015 genes to

254). We also observed an increase of 631% (666 to 4873) and 262% (340 to 1232) in the number of significant genes differentially expressed with T2D in brain caudate ganglia and skeletal muscle when controlling for BMI.

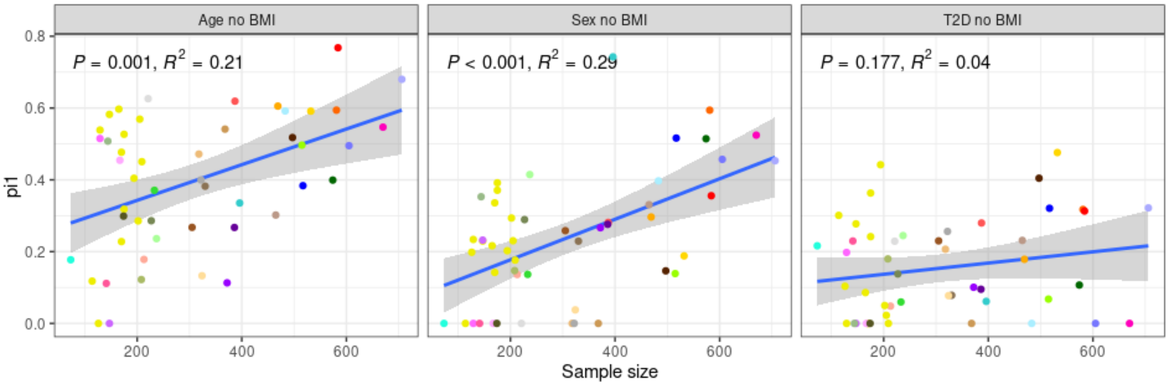


Figure 16: ***Proportion of significant associations against the sample size for traits without BMI correction.*** *Regression line in blue and coefficients above.*

Considering the relative importance of BMI as a factor in diseases related to these tissues (coronary heart disease, heart failure, T2D and various myopathies), controlling for this factor was expected to reduce the number of DEGs observed. Instead, we observed a significant increase of the number of genes supposedly acting independently of the BMI environment. In the case of skeletal muscle, although not well documented in the literature, there seems to be a consensus that muscle, as one of the main energy consumers of the body, can show a reverse relationship with diabetes aetiology, with higher muscle mass being protective against the disease.

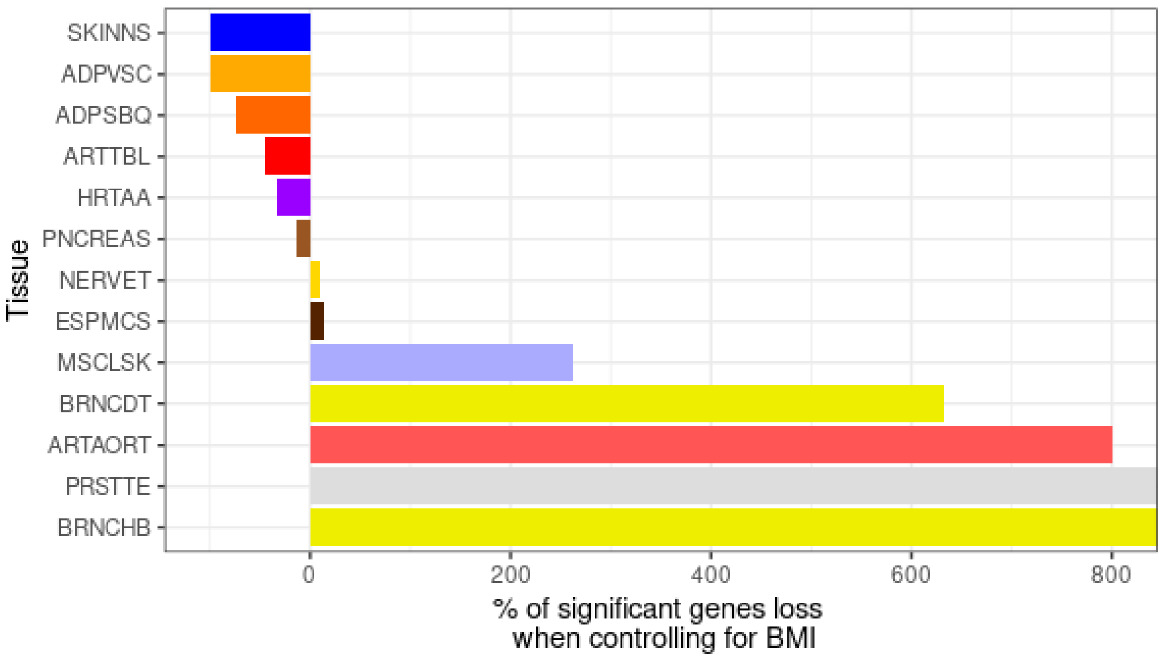


Figure 17: ***Representation of the percentage of total gene loss (or gain) for each tissue when controlling for BMI.***

For some tissues all DEGs were only identified when using one model. For example, for adipose visceral omentum, 16 DEGs were significantly associated with T2D only when not controlling for BMI effects. One of those significant genes was *SPX*, a gene coding for the spexin peptide, that have been previously associated with obesity and implicated in satiety and food intake (Behrooz et al. 2020). This suggests that T2D genes identified in these tissues were differentially expressed due to BMI. On the other hand, the *RHOG* gene was only significantly associated to T2D in the artery aorta when controlling for BMI. This gene is coding for a protein regulating insulin secretion pathway in the pancreatic islets, and suggest that genes identified in models controlling for BMI could be genes associated to beta-cell dysfunction. All in all, our results suggest that the influence of obesity-mediated T2D on gene expression was not uniform across tissues.

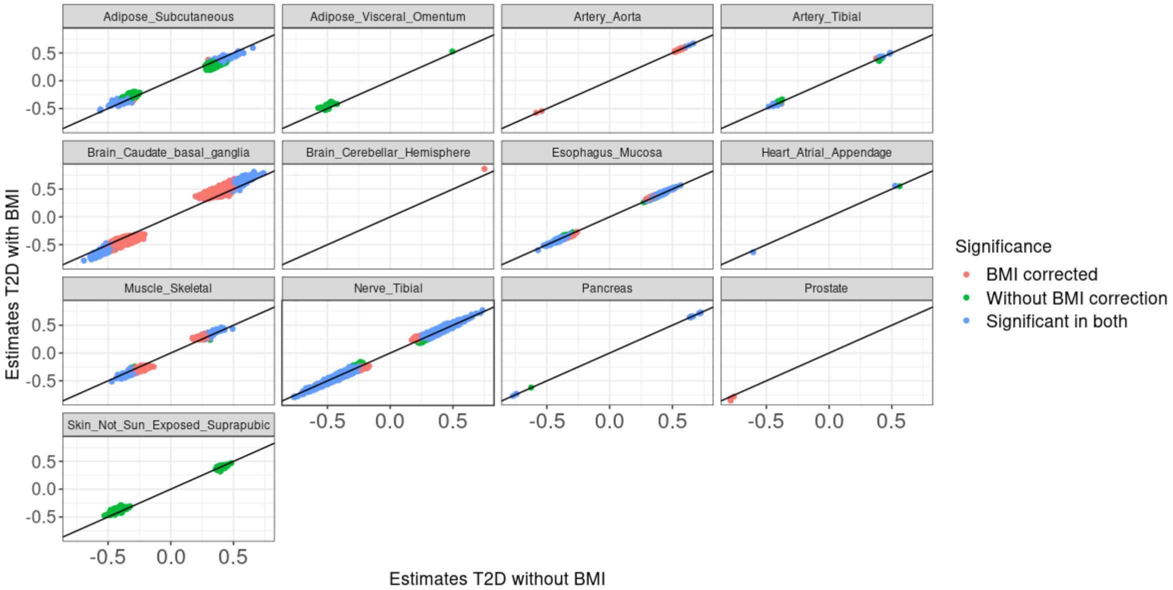


Figure 18: ***Detail of the difference in significant gene sets when controlling for BMI.*** *In red the genes significant when controlling for BMI, in green genes significant only when there is no BMI correction and in blue genes significant in both sets.*

Distribution of the variance explained by covariates

Given the large change observed when evaluating the role of BMI in the identification of T2D-genes, we investigated the proportion of variance explained by 5 of our biological and technical covariates, namely T2D status, BMI, sex, age and also ischaemic time as it is an important factor when studying post-mortem samples. In this way we could address the relative importance of these covariates for predicting expression, and assess whether this was dependent on the tissue being studied. In some cases, one biological covariate was more important than any of the others, for example the ischemic time was the most important technical predictor of expression in testis, while age was the most important biological predictor of expression in whole blood. However, these five variables explain a relatively small proportion of variance in expression across all tissues especially compared to heritability, the fraction explained by genetics which represents ~ 15-20%. Previous findings have shown that an unmeasured environment, including stochastic noise, is the most important contributor (Grundberg et al. 2012). In terms of overall relative importance, age and ischemic time are seen as more important than T2D status, sex and BMI. As expected, this matches the number of associations found with these covariates, ischaemic time showed the highest amount of significant associations (up to 15343 genes in lung), followed by sex and age (14319 in breast mammary tissue and 12897 genes in tibial artery respectively) identifying more DEGs than BMI (7437 genes in whole blood) and T2D status (5174 in tibial nerve).

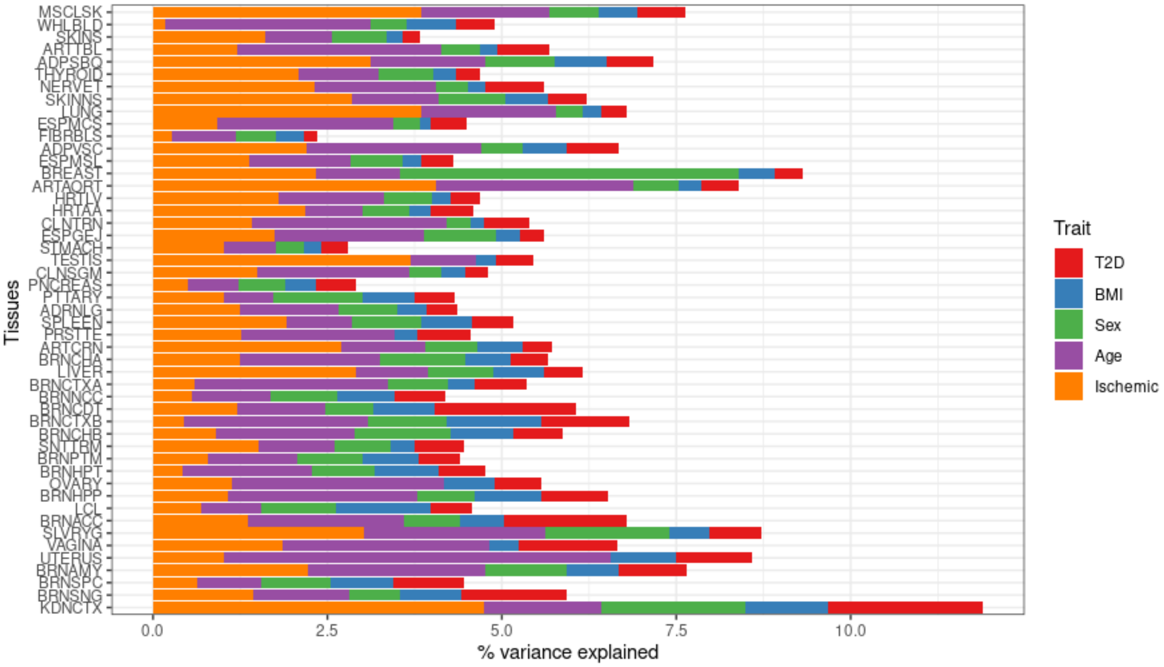


Figure 19: ***Percentage of variance explained for each tissue and for 5 traits : T2D, BMI, Sex, Age and ischemic time.*** *Tissues are ranked by sample size (lowest at the bottom).*

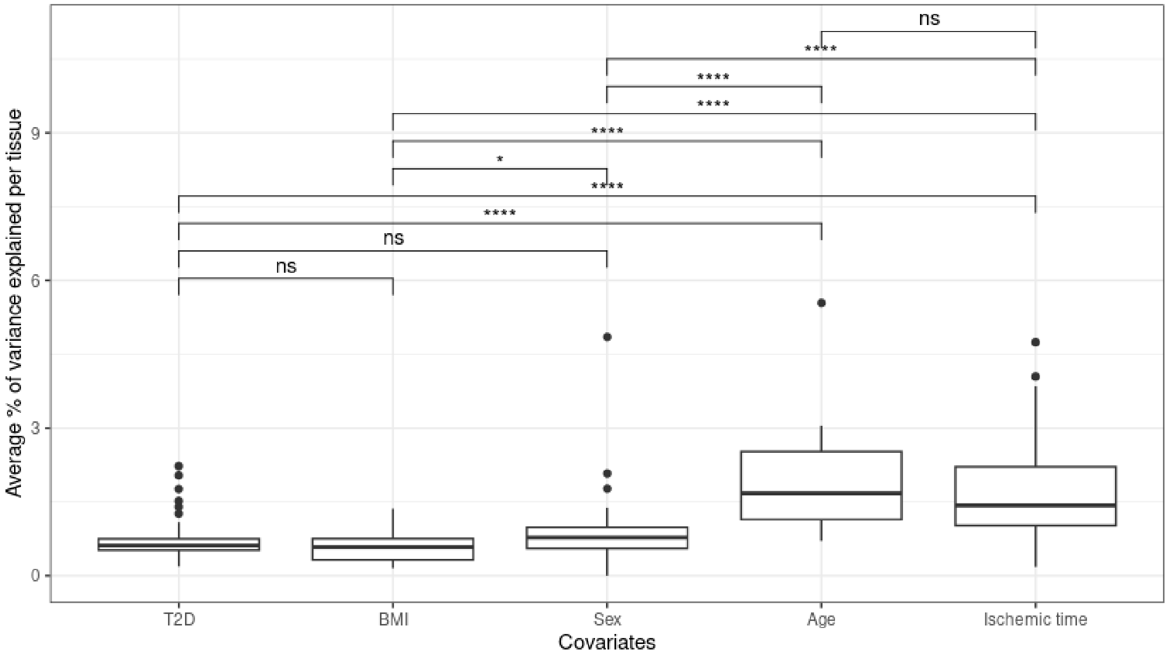


Figure 20: ***Boxplots comparing the average percentage of variance explained per tissue for each o the 5 studied traits.***

Although all tissues seem to display a similar distribution of proportion of variance explained by these variables, there are a few exceptions and outliers. For example, ischaemic time explains 27 times more variance in the kidney cortex compared to whole blood. The importance and impact of ischaemic time have been shown before in previous studies. The cascade of events resulting from the organism’s death and the response to stress such as hypoxia, cell death and autolysis is known to have an effect on the transcriptome.

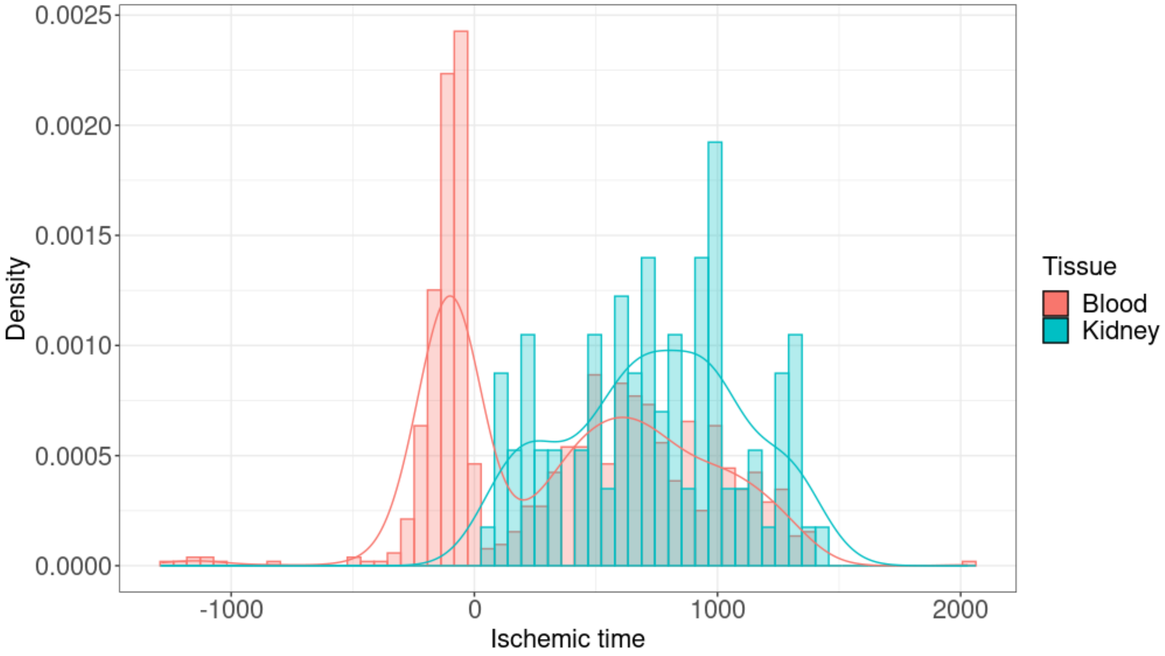


Figure 21: ***Proportion of ischemic time values in whole blood and kidney.***

All in all, we observed different proportions of variance explained by different variables. One of the driver of such as variation would be the cellular composition of tissues. GTEx generated bulk tissue expression profiles information, and tissues can differ greatly in the degree of homogeneity and exposure to environments. In particular, tissue cell composition is know to change with age which may explain why it was the more relevant variable in most tissues. However, the biological features we explore were limited as well, as other factors not measured such smoking status, cause of death and disease comorbidities, may explain more variance. For example, the fraction explained by genetics has been previous shown to explain a higher proportion of gene expression variance than age in multiple tissues (Viñuela et al. 2018).

Looking for insights into the causality of our signals

To explore whether the DEGs identified were drivers of disease, consequences of disease or consequences of co-morbidities of disease , we investigated the proportion of DEGs genes for which there was also genetic evidence of a causal effect on T2D risk. We used a curated set of 403 significant associations with T2D from a published GWAS study (Mahajan et al. 2018) and evaluated the enrichment of DEGs in a 1MB window around the significant GWAS hits with a chi squared test from two by two contingency table.

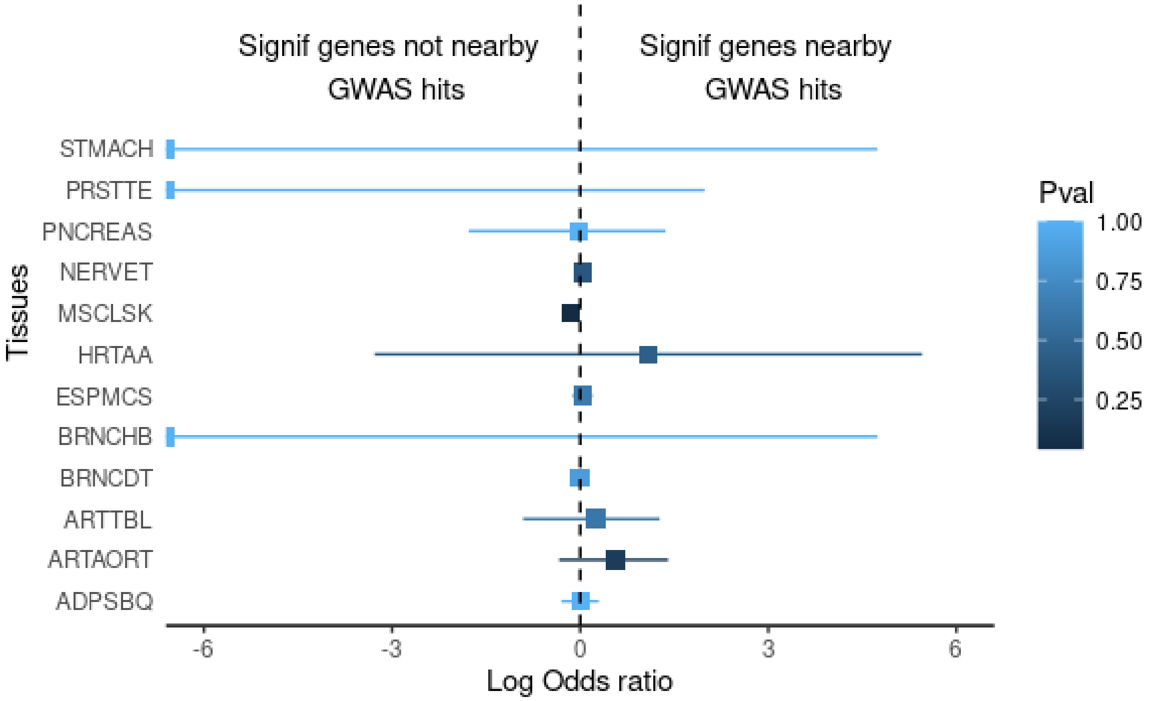


Figure 22: ***Odds ratio of genes found nearby a significant GWAS variant.***

We found no evidence of enrichment in any tissue (Figure 22). This would be due to low power because of the low number of GWAS hits and DEGs in most tissues, but others have also concluded that most DEGs are not causal for disease (Porcu et al. 2019). However, even if not causal for disease, DEGs would still be interesting from a treatment perspective, or to better understand how symptoms develop.

Associations for brain regions identified known T2D genes

The number of significant associations differed greatly across brain tissues, despite similar sample sizes. In some cases this was in agreement with previous experiments that have shown differences in blood flow in the ganglia region between patients with and without T2D (Jansen et al. 2016); this was also the brain region where we found most DEGs. This gene was detected in the caudate basal ganglia and the cerebellar hemisphere. A study on sex differences in brain connectivity, found a strong difference between men and women in the cerebellum area (Ingalhalikar et al. 2014), which was also one of the tissues where we discovered the largest numbers of DEGs for sex (N = 1281). These brain results demonstrate that even with small sample sizes, we would discover hundreds of DEGs, as long as the tissues were somehow relevant for the trait or disease (Figure 23).

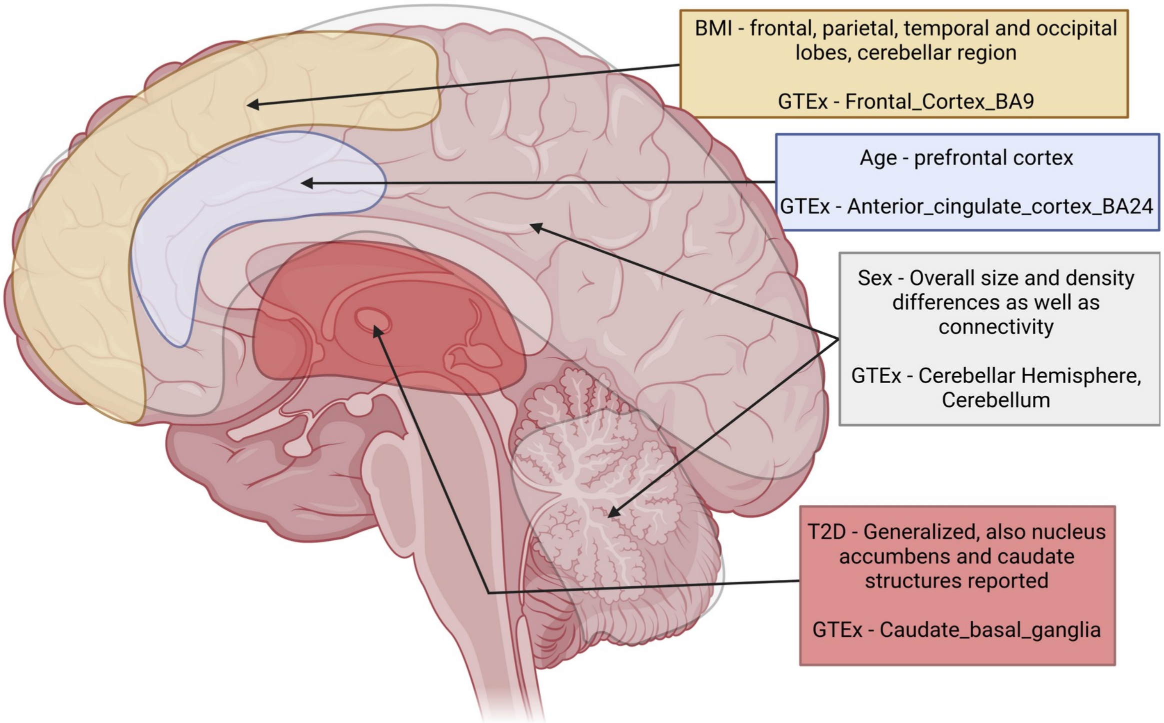

BMI - frontal, parietal, temporal and occipital lobes, cerebellar region
GTEx - Frontal_Cortex_BA9
Age - prefrontal cortex
GTEx - Anterior_cingulate_cortex_BA24
Sex - Overall size and density differences as well as connectivity
GTEx - Cerebellar Hemisphere, Cerebellum
T2D - Generalized, also nucleus accumbens and caudate structures reported
GTEx - Caudate_basal_ganglia

Figure 23: ***Comparison of regions of the brain found to be associated with a trait in the litterature vs in our results.***

The case of *INS*, the insulin gene

Since the *INS* gene, coding for the insulin protein, is an important gene in the molecular processes associated to T2D, we wanted to investigate changes in its expression across traits. Out of the 49 tissues evaluated, 42 expressed this genes, but only in the pancreas tissue was significantly associated with T2D (p value = 0.008, beta = - 0.41). The gene is only marginally expressed in all but pancreas tissues, since pancreatic islets are the only organ producing this molecule. Therefore, after multiple testing correction, none of these tissues were significant for any associations (Figure 24).

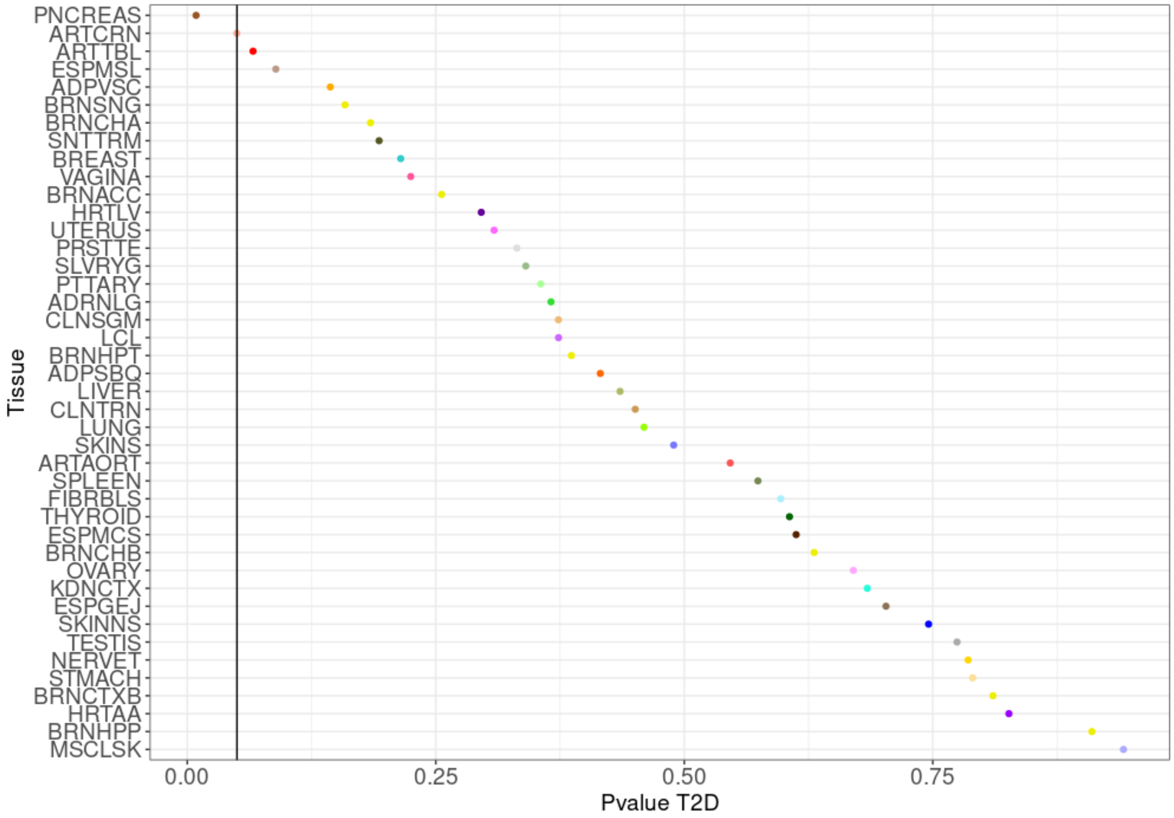


Figure 24: ***Distribution of pvalues for the INS DEG in all GTEx tissues.***

Gene set enrichment analysis

Finally, we performed a gene set enrichment analysis to look into the possible shared functions and pathways identified by the DEGs with T2D. Enrichment analyses evaluate if DEGs are overrepresented or enriched in specific functional categories or biological pathways than would be expected by chance. We used the GSEA (Gene Set Enrichment Analysis) tool (Mootha et al. 2003) as it uses a combination of methods including ranking of genes and permutations to make the results more robust. Rank-based approaches rank all genes based on their differential expression or other statistics (in our case, correlation p values) and have been shown to improve the overall pathway analysis results (Zyla et al. 2017). Permutations for statistical testing are non-parametric tests generating null distributions by randomising gene labels and then comparing actual test statistics to this null distribution to obtain a non-parametric p value.

We found 4921 unique pathways enriched across the DEG with T2D in the 49 tissues (qvalue < 0.01). There were no correlation between the number of pathways enriched and the total number of DEGs observed per tissue nor with the relevance of the tissue for T2D. However, pituitary was the tissue with the larger number of pathways enriched across its DEGs (1790), closely followed by stomach (1446), both tissues not directly related to the disease (Figure 25).

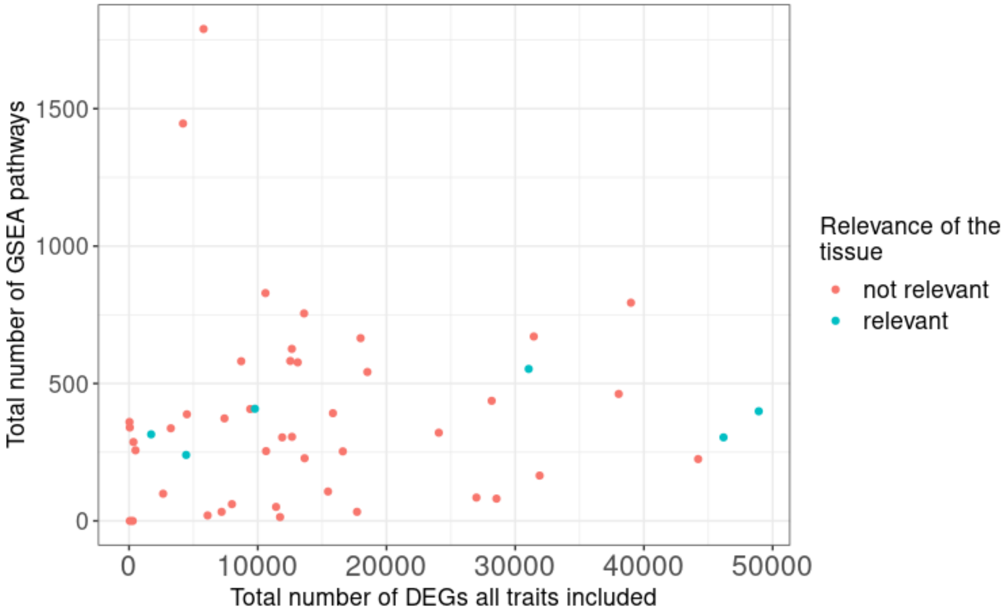


Figure 25: ***Number of total GSEA pathways against the total number of DEGs for each GTEx tissue.*** *Colors are for relevance of the tissues for T2D.*

Among the significantly enriched pathways, we observed a large number of mitochondrial related pathways, whether it is respiration or oxidative phosphorylation, were enriched across all tissues, with high enrichments scores (absolute NES > 3). These are key processes for cellular survival and energy consumption. A large number of these pathways were negatively enriched in skeletal muscle, a tissue heavily relying on cellular respiration to balance between exercise and normal state. To our surprise, we found no significant enrichment of T2D DEGs for the insulin related pathways in the pancreas despite being the *INS* genes DEG only in this tissue. We observed an enrichment on insulin receptors or responses to insulin pathways in some tissues such as prostate and vagina. Moreover, we saw enrichment of insulin secretion or regulation of insulin secretion pathways in brain regions, supporting a role of brain molecular processes in T2D (Figure 26).

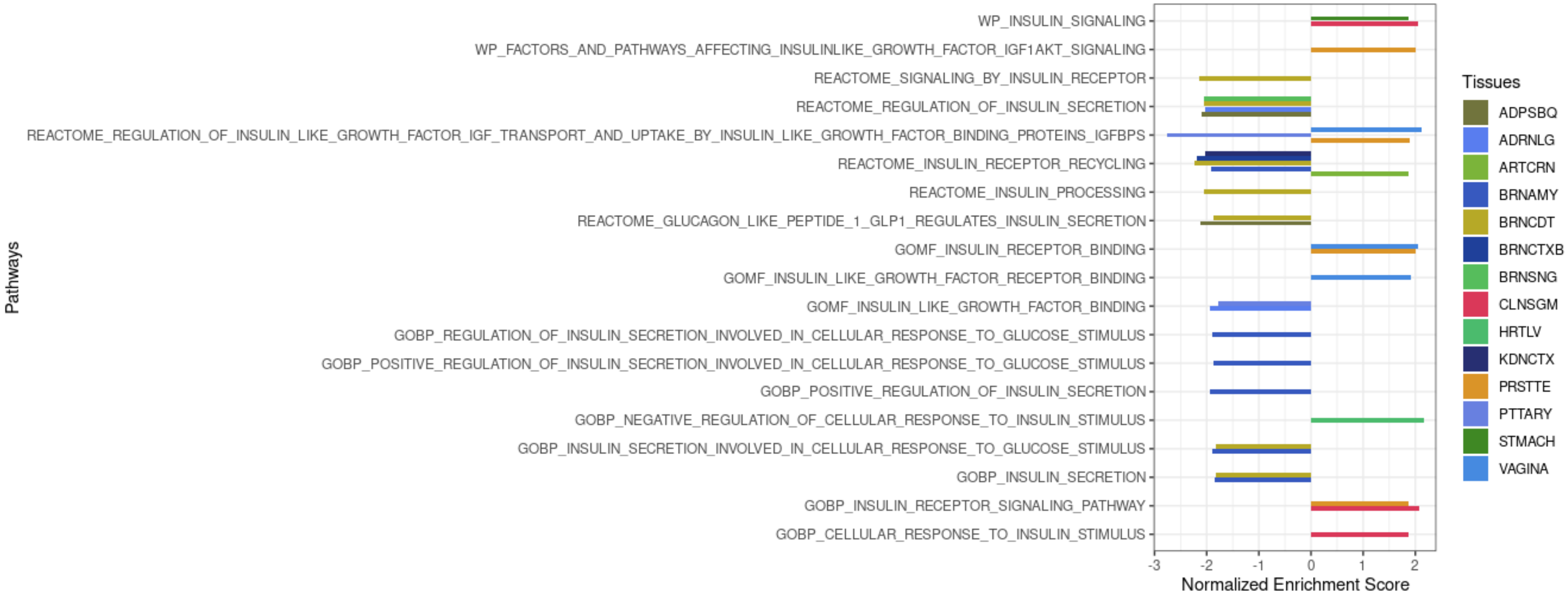


Figure 26: ***Insulin related significant pathways in GTEx tissues.***

Discussion

We conducted a DGEA on T2D using 49 tissues available from the GTEx consortium (Aguet et al. 2017). Given T2D is a disease involving a number of different tissues, this allowed us to investigate multiple processes that could lead to the disease. We identified between 1 and 5174 significant DEGs in 12 tissues including both relevant and non relevant tissues for the disease and a sample size between 175 and 581 individuals. However, we were surprised to find that neither the relevance of the tissue for the disease, nor the number of available samples in that tissue, influenced the number of genes we discovered. In comparison, the number of available samples was an important factor when looking for DEGs with other biological covariates, namely age, BMI, sex and also a technical variable such as ischaemic time. In general, these factors were much more influential on gene expression; 15,343 DEGs were discovered with ischaemic time in the lung for example. Relevance of tissue was also found to be an important factor for some of the variable. For example, the larger number of DEGs for sex were found in breast mammary tissue, which is known to have different cell compositions in men and women (Howard and Gusterson 2000).

A possible reason for the limited numbers of DEGs detected with T2D is that GTEx was not designed as a T2D study, and the number of people with diabetes was limited (from 42 to 148 , depending on tissue), limiting the power to discover DEGs. Moreover, T2D can be related to failures in beta cells production of insulin. These cells are contained in the pancreatic islets, but the corresponding tissue in GTEx is pancreas for which islets make up less than 2% of the cells. A study that used gene expression from pancreatic islets found that using these samples allowed them to explain more genetic associations with T2D and other related glycaemic traits than whole pancreas (Viñuela et al. 2020). In addition, GTEx samples also come almost exclusively from post-mortem cadaveric donors, the effect of this can be seen in the large numbers of associations discovered with conditions and as we have seen when looking at the results and analyses for ischaemic time.

While writing this chapter, a study very similar to this ours was published using the same data (García-Pérez et al. 2022). Through their work, similarly to what we found, they noticed that some demographic traits such as ancestry, age, sex and BMI have a significant impact on differences of gene expression between individuals. Although both results were not exactly the same (they used different tools but also a slightly different model including PEER factors as well as exonic rate), we both found interesting results regarding tissues such as tibial nerve displaying a high number of DEGs. Their analysis of variance explained by these traits was also similar to ours, showing a relative heterogeneity among different tissues and overall small numbers compared to heritability. One additional interesting point to mention is the absence of DEGs in the liver tissue similar to what we observed in our analysis. Both our models are including a large amount of covariates which could lead to overfitting scenarios and therefore a loss of significant true signals in some tissues with a model being too stringent. Among solutions for these issues, we can cite regularization methods such as L1 or L2 penalty to compensate overfitting (ridge or lasso method) or removing some of the least relevant parameters. Even though their work was more focused on splicing where we decided to discuss more causality and directionality of our DEGs, they added another layer of comprehension, especially on diabetic neuropathy and how this type of analyses can help investigate the multiple demographic and environmental contributors to gene expression and how it impacts our understanding of a disease.

One of the advantages of genetic studies is the promise that significant associations are causal in nature. Since T2D cannot mutate DNA, the direction of causality is from genetics towards disease. In contrast, DEGs are often assumed to be the consequence of disease (Porcu et al. 2021). In the case of T2D, the interpretation is even more complicated by the fact that the disease is often found alongside other conditions. These include obesity, cardiovascular disease, dementia and other traits. Each could have their own effect on expression, with a complicated causal network relationship between them. We aimed to understand the relationship between expression, obesity and T2D by looking at results with and without controlling for BMI and saw that when controlling for BMI, some tissues reduce the number of DEGs. This was driven by the fact that the effects of T2D and BMI on the expression of these genes were not independent. The opposite scenario was also displayed in tissues such as skeletal muscle or artery aorta meaning that these genes are independent from BMI and could be associated with other factors or comorbidities.

In order to understand if our findings were causal or consequences of T2D we also investigated the underlying genetic evidence of our signals. We found out that the majority of the DEGs were not overlapping with any gene associated to strong genetic variants from previous studies, therefore we shall consider changes in expression for these genes a consequences of the disease instead of causing factors. This is in line with the previous findings from Porcu et al, where the authors employed similar methods for DEGs and evaluate reverse causality on the link between gene expression and complex traits. The authors conclude that the trait often has more impact on the gene expression than gene expression on a trait, implying differentially expressed genes are not causal of the trait or disease.

However, although the majority of the signals might be a response to T2D, exploring and interpreting these associations could help us understand the various consequences and co-morbidities of T2D. In fact, knowing which genes are impacted by T2D could give us information about the disease progression but also the different symptoms. For example, identifying genes involved in glucose uptake with a reduced expression could let us learn how insulin resistance is impacting the glucose homeostasis.

Methods

The expression data from GTEx used for differential expression analysis was first filtered on protein coding and lincRNA using the corresponding GENCODE reference file (release v26) (Frankish et al. 2019). Then, as mentioned in the beginning of the results, multiple covariates where tested before including them in the regression model. First, principal component were calculated based on individual gene expression data and using the *prcomp* function to then extract the 5 first PCs. To test the importance of the covariates, we then combined these PCs with individual information on age, sex, BMI, ischaemic time, autolysis of the tissue, RIN score, centre, ethnicity as well as the first 5 genotyping PCs as another proxy for ethnicity. Then after extraction of the pvalue and variance explained by each of these covariates, we calculated a score based on the sum for each covariates of (-log10(pvalue) * variance explained).

The final model was as follow : *Exprs ~ T2D_status + BMI + Age + Sex + RIN_score + Autolysis + Center + Ischemic_time + Geno_PCs + pcr + platform* . We ran this model on each of the 49 GTEx tissue and extracted estimates, betas as well as calculated variance for T2D, age, sex, BMI, ischaemic time and the same without including BMI in the model. Pvalues were then adjusted using the *qvalue* function and extracting corresponding qvalues.

π1 values to represent the proportion of significant results were calculated using the *pi0est* function in R and subtracting π0 results to 1 for each of the tissues and traits tested.

The variance explained by each of the five traits was calculated from the regression statistics when running the differential gene expression analysis. In order to achieve this, we used this formula for each trait : variance = *anova(sum of squares) / sum(anova(sum of squares)) * 100*

To look at genes nearby GWAS hits we used the localisation information from GTEx data and significant SNPs from Mahajan et al 2018 with the bedtools tool to test for overlaps using the *intersect* function. Then, odds ratio were calculated by using the *odds.ratio* function and contingency table with the number of genes in common or not between the two sets.

The pathway analysis was done using the GSEA software (Subramanian et al. 2005). We run *GSEAPreranked* with every parameters by defaults and including database from KEGG, reactome, GO Terms and wikipathways as well as gene sets for each tissue. The results were then extracted and filtered to look at specific pathways using keywords.

Chapter 2 : Discovery of Type 2 diabetes genes using an accessible tissue

The limitations of the GWAS approach

Genome wide association studies (GWAS) in large cohorts have greatly expanded our knowledge on the genetic causes of complex diseases such as T2D, implicating hundreds of novel loci (Ali 2013; Visscher et al. 2012; Vujkovic et al. 2019). However, most of these genetic loci are located in non-coding regions of the genome. To understand the immediate consequences and function of non-coding variation, molecular studies have assayed gene expression in multiple human tissues and linked genetic variation to changes in gene expression. These studies provide a path for discovering disease causing genes and the tissues in which they are active by connecting GWAS loci to genes in a particular tissue (Viñuela et al. 2020; Alonso et al. 2021; Miguel-Escalada et al. 2019). Transcriptome wide association methods are a type of method using reference expression datasets to build predictive models of genetically regulated gene expression. These models are then applied to GWAS studies to produce cis-predicted expression phenotypes across thousands of genes, which can be tested for association with the disease. Where individual level data is unavailable, approximations exist using summary statistics. It has been argued that TWAS methods discover genes that are causal of the disease, because these are based solely on genotype data that predates the development of disease. However, this interpretation is complicated by issues such as pleiotropy (SNPs associated to the expression of several genes) and linkage contamination (correlated SNPs where one affects disease risk and the other expression)(Wainberg et al. 2019). This means that while the methodology may implicate a particular gene, the true causal drivers may be other variants and genes nearby. For this reason we will refer to TWAS significant associations as causal loci rather than causal genes.

The importance of reference tissues, sample size or relevance?

Reference datasets used to predict gene expression should be based on tissues that are directly relevant for the disease, to ensure that the tissue specific genetic effects that alter disease risk are captured(Sun et al. 2021). However, many disease relevant tissues are difficult to collect, limiting the sample size of such experiments and the ability of TWAS methods to produce accurate predictors of gene expression (Figure 27A). Studies with small sample size are best powered to discover expression quantitative trait loci (eQTLs) with the largest effect size, these are typically located in promoter regions (The GTEx Consortium 2020) and have cross tissue effects, reducing the identification of tissue specific genetic effects. For example, T2D, a complex disease characterized by increased levels of glucose in blood (WHO 1999), involves multiple tissues with physiological and genetic factors influencing its pathogenesis, such as insulin resistance, obesity and beta-cell dysfunction. For each mechanism of action, different tissues and organs play greater and lesser roles. Insulin resistance related T2D is often associated with molecular changes in skeletal muscle, adipose and liver, while beta-cell dysfunction is almost exclusively associated with the pancreatic islets, where beta-cells are localized (Galicia-Garcia et al. 2020). Many of these relevant tissues are difficult to collect pre-mortem, with consequences for our ability to discover the genetic and molecular processes involved in the development of T2D. On the other hand, gene expression studies with thousands of samples have shown that the expression of most genes is affected by multiple eQTLs and many of these effects are shared across multiple tissues (Aguet et al. 2017; Brown et al. 2023). Therefore, it is possible that larger molecular studies using samples which from tissues easy to collect such as blood, could identify disease relevant loci.

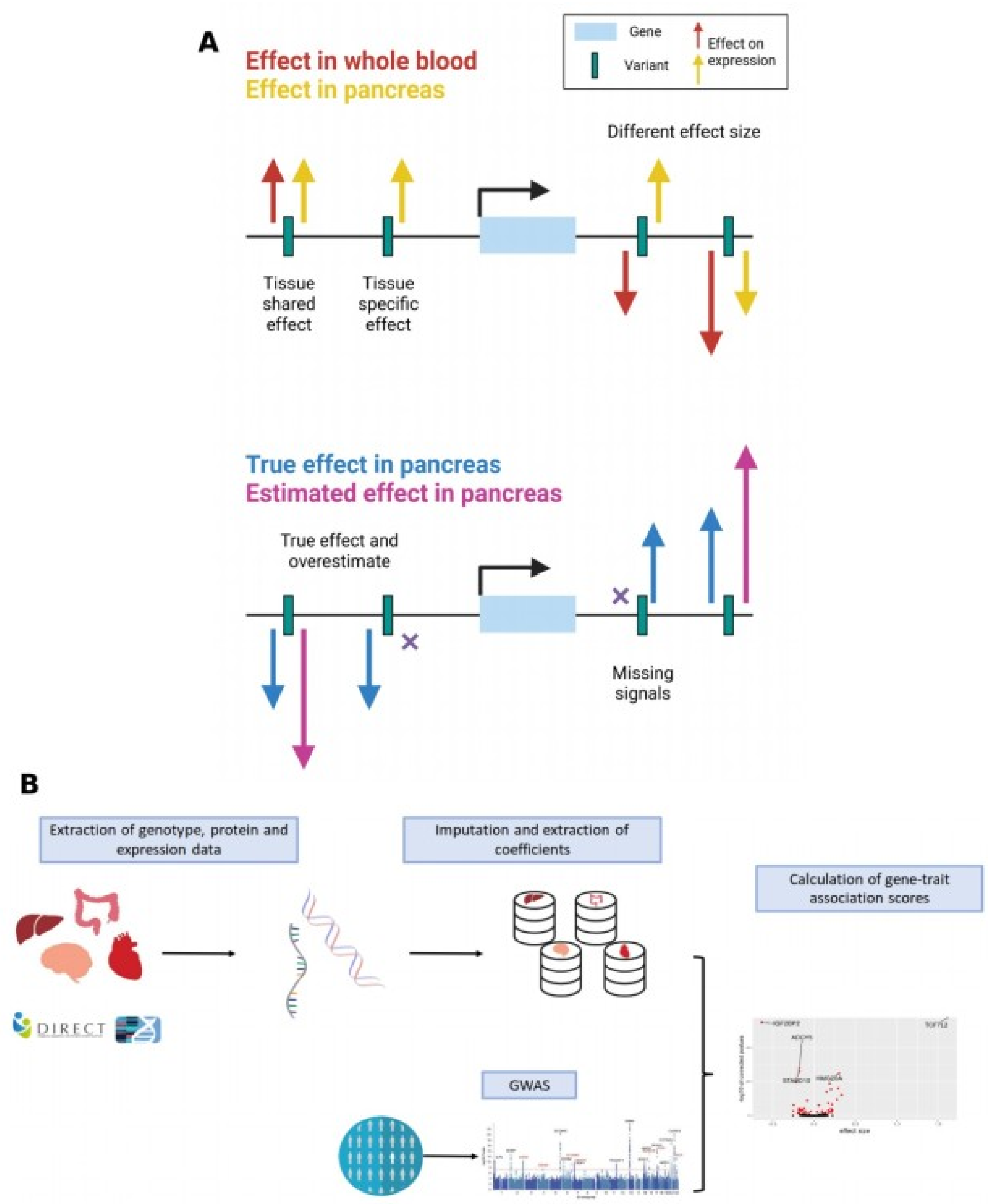

A
Effect in whole blood
Effect in pancreas
Gene
Variant
Effect on expression
Different effect size
Tissue shared effect
Tissue specific effect
True effect in pancreas
Estimated effect in pancreas
True effect and overestimate
Missing signals
B
Extraction of genotype, protein and expression data
Imputation and extraction of coefficients
Calculation of gene-trait association scores
DIRECT
GWAS

Figure 27: ***A. Schematic overview of the tradeoff between tissue specificity and sample size and their impact on detection of effects on expression.*** *The upper section shows the consequences of using different tissues on the effect captured with TWAS. The lower section shows the different scenarios when the sample size is small.* ***B. Study workflow.*** *We developed prediction models for genetically regulated gene expression and protein levels in multiple datasets with different sample sizes. Predictive models were then associated to T2D using GWAS summary statistics to identify differentially expressed genes with T2D.*

Aims of the chapter

Here, we aim to identify T2D relevant genes using TWAS methods in a large dataset from whole blood, an accessible tissue. We also assess the roles played by sample size and tissue relevance when identifying disease genes using TWAS methods. We compared the ability to find T2D related genes using a large whole blood expression dataset produced by the DIRECT consortium (n=3029) to the performance using relevant tissues but with smaller sample size (GTEx consortium, n < 706, and pancreatic islets from INSPIRE, n=420). We show that sample size is the chief determinant of discovery power for associations between gene expression and disease risk. Furthermore, we explored the effect of considering environmental factors such as BMI on discovering T2D loci and the use protein-QTL (pQTL) data to gain potential biological insights on the disease. Finally, using GWAS data from T2D subtypes, we identify novel genes implicated in Severe Insulin Deficient Diabetes.

Summary of the method

To identify T2D causal loci we used gene expression reference datasets produced by three consortia: DIRECT (whole blood, sample size 3,029), 49 tissues from GTEx (sample size 70 to 706) and InsPIRE (pancreatic islets, 420). These datasets were chosen to represent a full spectrum of modestly sized expression references generated from tissues that are directly relevant to T2D (pancreatic islets, subcutaneous and visceral adipose tissue, liver, skeletal muscle and kidney cortex) to a large dataset generated from a non-relevant tissue (whole blood, DIRECT) (Viñuela et al. 2020; The GTEx Consortium 2020; Koivula et al. 2014) (Methods). Transcriptome wide association studies construct models of gene expression from reference expression datasets and use these to predict expression in GWAS studies, which can then be used to test for association with disease. In this study, expression was modelled using conditionally independent cis-eQTLs released by each consortium and then associated with T2D using GWAS summary statistics from the DIAGRAM consortium and the MetaXcan method (Figure 27B)(Barbeira et al. 2016). By adopting this approach the results were directly comparable across studies, as all studies used the same pipeline to generate conditionally independent cis eQTLs.

Main results

Using a TWAS method, we identified 404 T2D associated loci using the whole blood DIRECT data (n=3,029 samples). This is the largest number of significant loci discovered across the reference panels, where 108 loci identified using the pancreatic islet INSPIRE dataset (n=420) and from 3 to 299 loci depending on tissue for GTEx (n<706) (1,568 in total; Figure 28A – C). We found a strong linear correlation between the sample size of the dataset and the number of significant loci discovered (Spearman correlation 0.95, p=7.62e^{-27}, Figure 28B). However, the DIRECT dataset did not follow the linear trend, identifying fewer loci than would be predicted (Figure 28B). This could be due to limitations specific to the DIRECT data, such as lower sequencing depth and multi-centre sample collection, or due to a saturation of signals to discover disease loci (Koivula et al. 2014). We also tested for an effect of tissue relevance, using a linear model excluding DIRECT of discovered loci against sample size and actually found that fewer loci were discovered using relevant tissues than would be expected based on sample size (estimate = -29.27, p-value = 0.019). In summary, we found a linear relationship between sample size and discovery of disease relevant causal loci with little influence of the suspected relevance of the tissue for the disease.

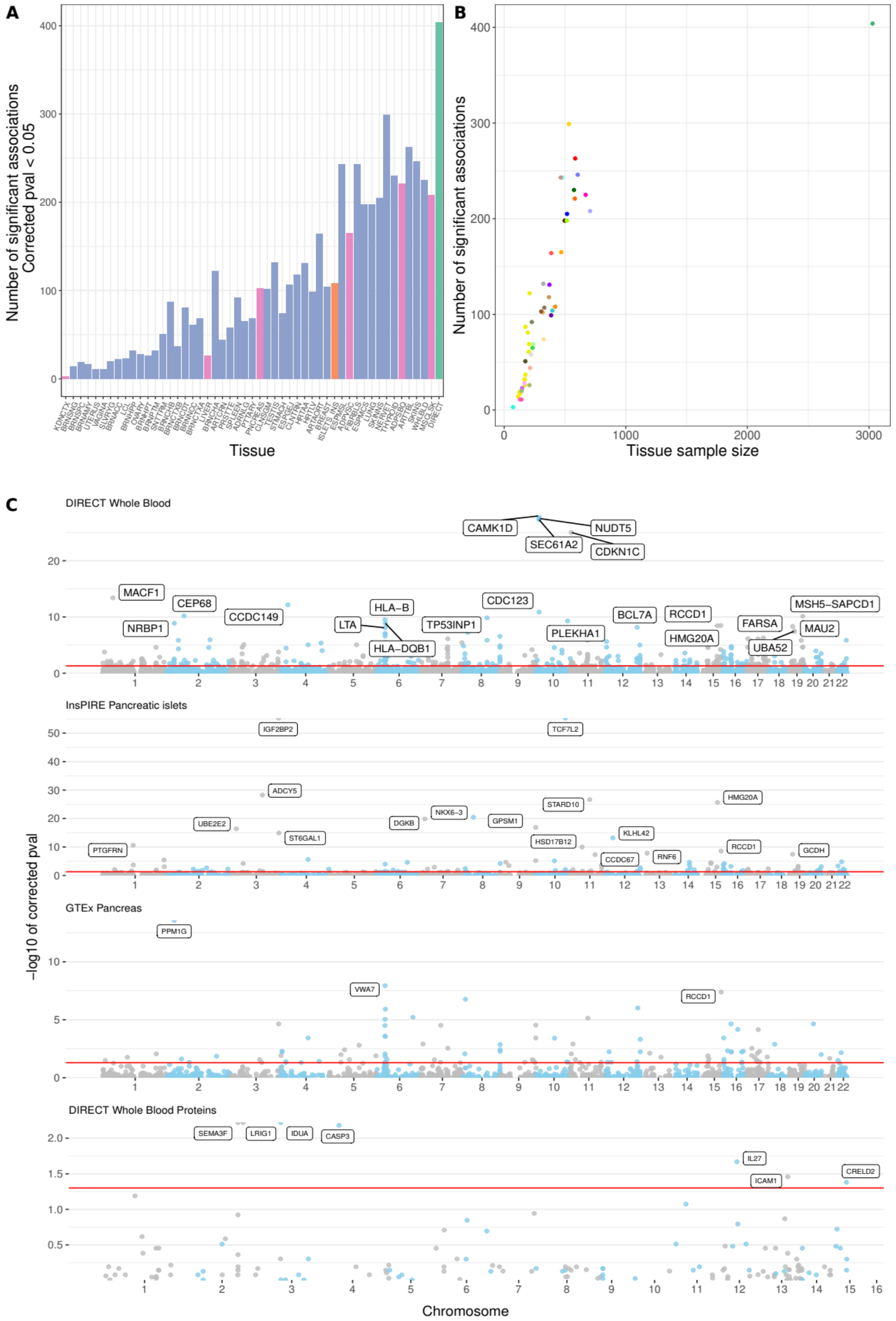


Figure 28: ***A. Distribution of the number of significant associations obtained with MetaXcan for each tissue tested.*** *Tissues are ranked by sample size and colored by datasets or disease relevance, green is DIRECT, orange is InsPIRE, pink and purple GTEx with pink showing relevance for T2D.* ***B. Scatter plot showing relationship between sample size and number of significant associations.*** *Colors use the GTEx official color palette plus DIRECT in green and InsPIRE in orange.* ***C. MetaXcan results for three tissues and DIRECT proteins in a Manhattan style.*** *The red line indicates the significance threshold at –log10 (0.05). The labelled genes or proteins, indicate highly significant associations (p < 5e-8 for genes and p< 0.05 for proteins).*

Tissue specific and shared associations

Large reference datasets identified more disease loci than smaller datasets, but the proportion of already known T2D loci was larger for relevant tissues (figure 29A). In DIRECT we identified 88 out of 404 (21.8%) of significant loci that were previously reported as related to T2D (Piñero et al. 2020). Across all GTEx tissues there were 251 known T2D loci out of a total 1,568 (16%, from 0 to 56 per tissue, Methods). In comparison, 30 out of 108 (28%) of the islets associated loci had previously been associated with T2D. For example, *UBA52* and *NAGLU* were among the DIRECT reported genes, which have previously observed to be over expressed in pancreas of individuals with T2D (Sjöstedt et al. 2020). On the other hand, only using pancreatic islets data were we able to identify an association between T2D and *TCF7L2*, a gene located near one of the strongest known T2D GWAS signals (del Bosque-Plata et al. 2021). Overall, this study identified 1,515 potentially novel T2D genes across all tissues.

We next investigated the degree to which the reported genes were discovered in multiple tissues. Comparing the genes reported in significant loci across datasets, we observed that more than half of all significant genes were only discovered in one tissue: 196 were identified using only DIRECT (49% of all DIRECT significant associations) while 51 were found using only InsPIRE (47%). Across tissues, we observed that similar tissues did not show any patterns of sharing genes (Figure 29B). This is consistent with most associations being driven by cross tissue effects, with a significant “winner’s curse” contribution to the discovery set in each tissue (Figure 29B). Moreover, we found that disease relevant tissues such as islets, liver and muscle, display more unique well known T2D genes compared to non-relevant tissues (Figure 29C). However, the pancreatic islets dataset missed known T2D genes such as *CDKN1C,* involved in beta-cell proliferation. This gene was only discovered using the increased power of the DIRECT reference. Our results show that studies with larger sample sizes such as DIRECT and the combined data of GTEx tissues, were able to identify more potentially novel genes (respectively 316 and 1317) compared to the relevant pancreatic islets tissue (78). Moreover, we show that processes in non-relevant tissues may be informative of genetic effects relevant to disease, identifying novel disease candidate genes that would be missed due to sample size limitations.

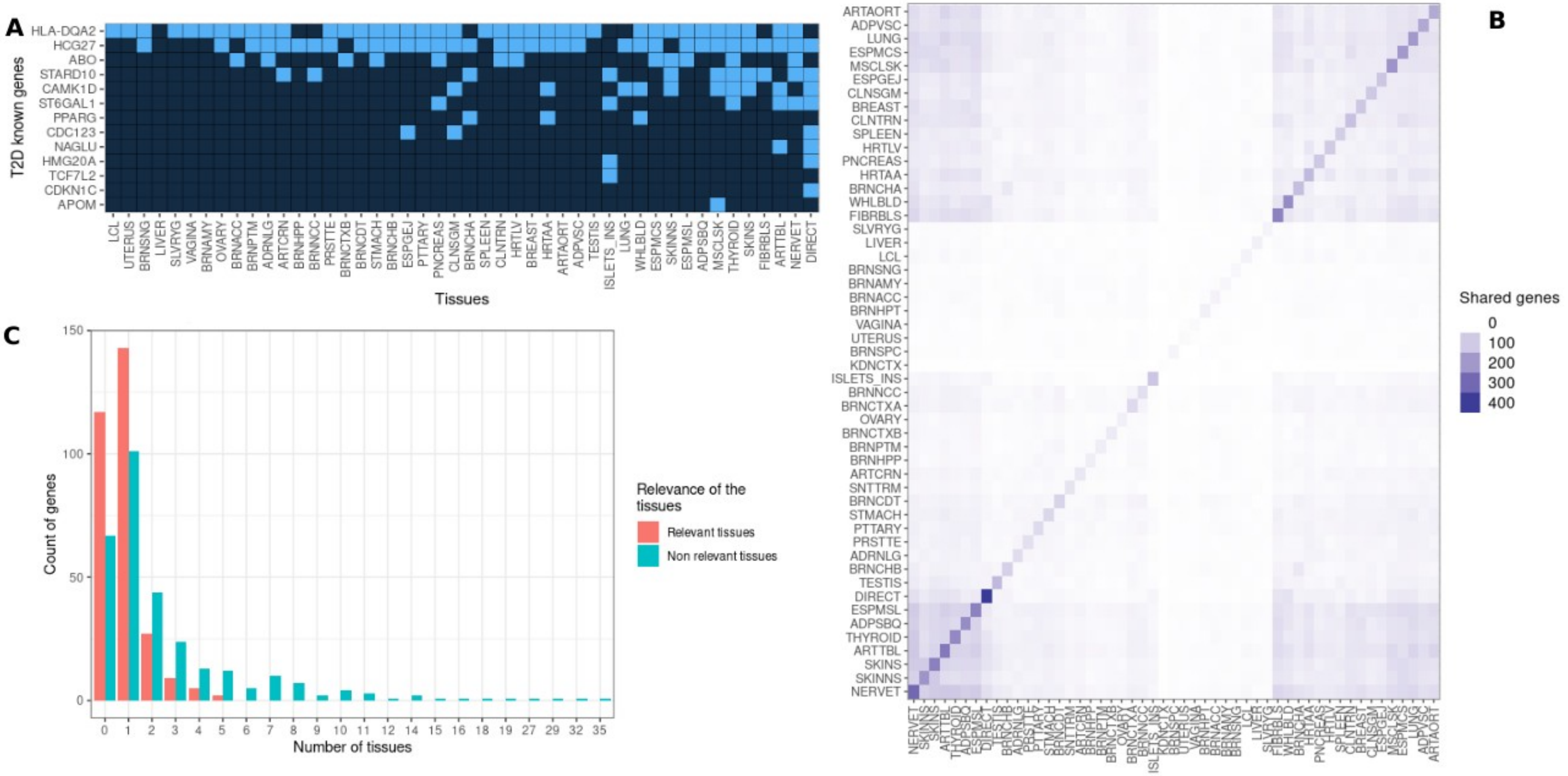


Figure 29: ***A. Heatmap of occurrences of 13 known T2D genes in all significant associations.*** *Light blue tiles designate the presence of the gene and dark blue the absence.* ***B. Heatmap of the number of shared significant association between each pair of tissues. C. Barplots representing the distribution of number of occurrences of significant known T2D genes in all tissues separated by the relevance of the tissue.*** *Red bars are for counts in relevant tissues and green bars are counts in non-relevant tissues.*

Looking for true causal effects

To better understand the nature of the SNPs underlying significant associations, we looked at the relationship between the causal loci and two sets of significant GWAS signals, to assess how well particular reference panels corroborated the loci implicated by these GWAS studies and to investigate whether loci could be driven by individual, highly significant signals. We used the 492 GWAS significant loci reported by Mahajan et al. 2018 (the GWAS used in the TWAS analysis), and the 186 signals reported by Spracklen et al. 2020, which used an East Asian population different from our main results. We found that DIRECT associated loci were more likely to be within 1MB of a significant GWAS signal than all loci from each of the GTEx tissues (Figure 30A). However, compared to islets loci, the DIRECT loci were less likely to be found near the GWAS signals of both studies ($OR_{Mahajan}$ = 1.16 and $OR_{Spracklen}$ = 1.22). This enrichment of DIRECT and INSPIRE loci relative to GTEx may be a result of loss of information when lifting over GWAS summary statistics to GRCh38 genome build, necessary to combine these statistics with GTEx reference panels. When considering GTEx tissues alone, the relevance of the tissue for T2D did not predict enrichment, with tissues such as adipose subcutaneous and pancreas depleted relative to the median tissue (Figure 30A). There was also no correlation between sample size and the enrichment of significant genes around GWAS loci (Figure 30B).

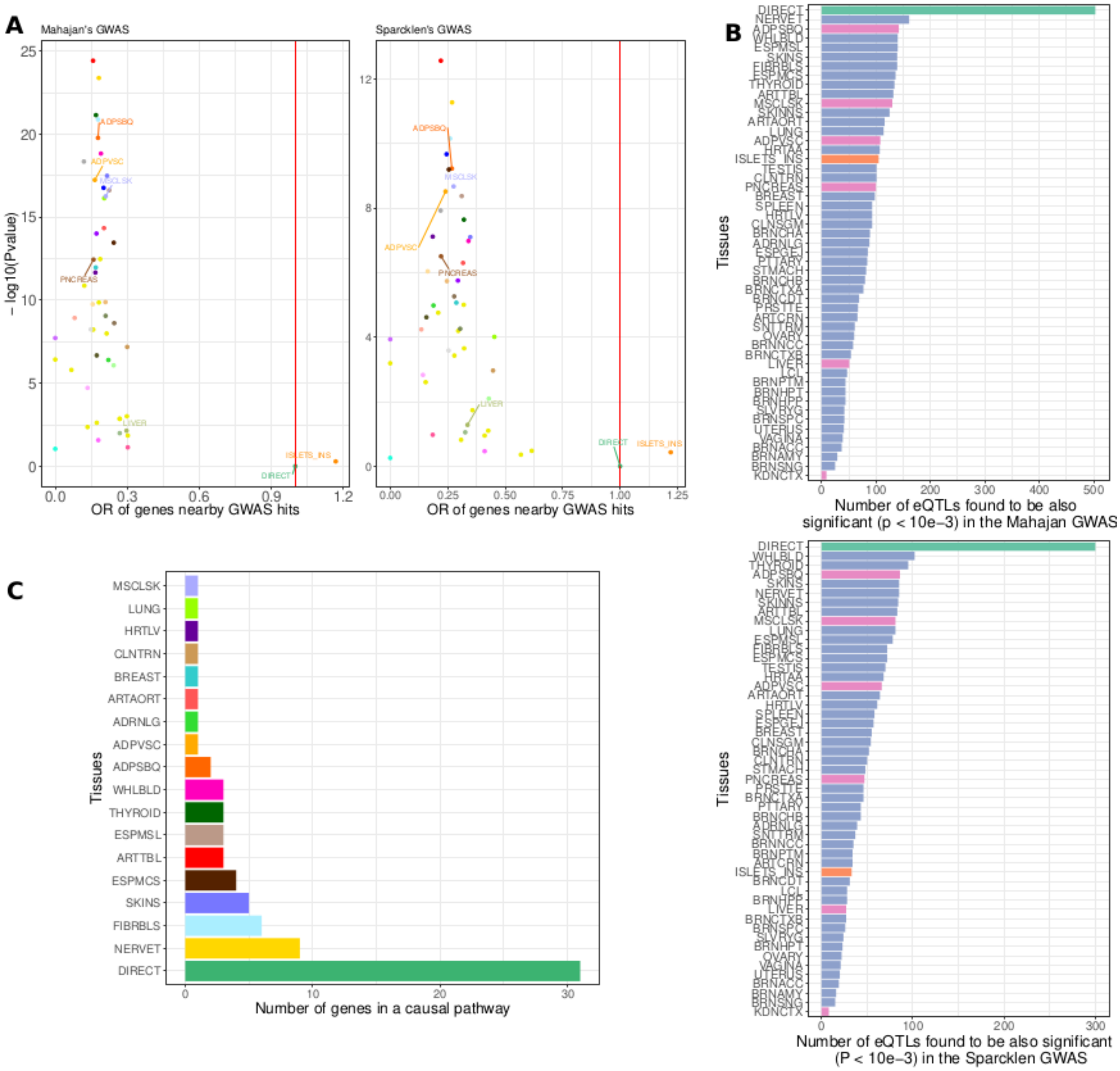


Figure 30: ***A. Proportion of significant genes located nearby significant GWAS hits from different studies.*** *Odds ratio have been calculated relative to DIRECT findings allowing to display on the right side of the red line tissues getting better results than DIRECT and on the left side of the line the opposite.* ***B. Barplot of tissues ranked relative to their number of eQTLs found to be significant in the two GWAS (P < 10e-3).*** *Colors are green for DIRECT, orange for InsPIRE, pink and purple GTEx with pink showing relevance for T2D.* ***C. MR analysis results.*** *We show the number of genes involved in causal pathways for each tissue tested.*

We hypothesized that the increase in the number of discovered loci by the DIRECT dataset, could be driven by loci involving multiple expression variants with moderate effects on T2D risk that are hard to find in smaller datasets. The DIRECT study originally reported more than 60,000 independent cis-eQTLs, while in this study we found that more DIRECT eQTL show suggestive association with T2D risk GWAS results: 503 eQTLs for DIRECT had P<0.001 in the Mahajan GWAS, compared to 105 eQTLs for islets (Figure 30B). To test more formally that more DIRECT loci were identified due to multiple eQTL effects on T2D risk, we applied multiple instrument Mendelian Randomization (MR). As proposed by the SMR and HEIDI methods (Y. Liu et al. 2021), we tested whether eQTL variants have consistent, non-zero effects on T2D risk. Using a Wald Ratio test, we looked for evidence of genetic effects affecting both traits, and applied a test of heterogeneity to evaluate the consistency of effects. In DIRECT, out of the 214 loci meeting the requirements for this MR methods, 31 were significant (corrected pvalue < 0.05 and a heterogeneity test pvalue > 0.05), compared to between 1 and 9 across the other reference datasets. We can notice that besides DIRECT, among the GTEx tissues, the top tissue with 9 genes validated through MR is tibial nerve, again demonstrating the importance of this tissue as observed in the first chapter. Therefore, we conclude that well powered reference datasets identified more loci with multiple variants with consistent effects on T2D risk than smaller reference panels (Figure 30C). These loci were supported by a larger number of SNPs located further away from the target genes, and therefore with moderate genetic effects, compared to SNPs used from smaller studies.

Proteomics and T2D

While genetic effects on disease risk are expected to be mediated by transcription at some point, for many diseases protein levels have been shown to be more informative biomarkers(Crutchfield et al. 2016; J. J. Li, Bickel, and Biggin 2014). We investigated if TWAS methods could identify T2D mediating protein loci by constructing models based on *cis* plasma targeted protein-QTLs (pQTLs). Using DIRECT genetic data and pQTLs affecting 149 proteins, we identified 7 loci associated with T2D: SEMA3F, IDUA, LRIG1, CASP3, IL27, ICAM1 and CRELD2 (Figure 28C). Among these, LRIG1, ICAM1 and CASP3 were coded by known T2D genes: LRIG1 is part of a protein family which is involved in lipid homeostasis and has been associated with T2D (Herdenberg et al. 2021), ICAM1 has been shown to be involved in diabetic retinopathy (Gu et al. 2013) and expression of *CASP3* has been associated with islet apoptosis (Liadis et al. 2005; Tomita 2010). Of the reported proteins, only LRIG1 and ICAM1 reported significant associations also with the genetically predicted gene expression for the coding gene, and for LRIG1 the reported associations showed opposite direction of effect ($Zscore_{expression}$ = -3.43, $pvalue_{expression}$ = 0.013, $Zscore_{protein}$ = 3.85, $pvalue_{protein}$ = 0.006). Despite the smaller number of loci identified with protein data, the proportion of associated loci *vs.* tested loci was approximately the same for proteins as it was for genes (4.6% compared to 4.7%). This suggests that both assays had similar power to discover mediating factors, though the genome-wide ability of RNA-seq to quantify mRNA means that more loci were tested and therefore discovered. We looked to see if for significantly associated proteins we were also more likely to find a significant association between disease and the corresponding gene, and we observe a large odds ratio for enrichment, though this was not significant due to the small numbers of proteins associated (odds ratio = 9.56, Fisher's Exact test pvalue = 0.51). This large odds ratio is consistent with other studies, that have found that genetic effects on proteins replicate on expression (Brown et al. 2023).

Looking for gene with BMI mediated effects

Previous studies have shown that processes occurring in the pancreatic islets are more likely related to T2D driven by beta-cell dysfunction, while obesity related T2D or insulin resistance would be linked to expression patterns in adipose or muscle tissue (Viñuela et al. 2020; Dimas et al. 2014). However, T2D GWAS studies have shown that most loci do not change their effect size when controlling for BMI (Spracklen et al. 2020). To explore disease subtypes causal loci, we repeated our association analysis using T2D GWAS summary statistics which controlled for BMI. Our assumption was that controlling for BMI would reduce the number of loci we discovered that relate to insulin resistance, which is known to correlate to BMI, and boost our ability to discover loci relating to beta-cell dysfunction by removing environmental variation. We found a reduction of 732 genes discovered overall, from 1,818 unique genes to 1086 and a fall in the number of protein associations from 7 to 4 (Figure 31A). We found a subset of 46 significant genes in pancreatic islets results detected only in the subset adjusted for BMI. Among these we can cite for exemple the *IGF2BP2* gene representing the highest zscore in the adjusted set of result (zscore = -16.17, pvalue = 5.30e-56). This gene, although being known for participating in the increase of T2D risk by disrupting insulin secretion, is also related to other metabolic diseases such as obesity, hence the potential presence of this gene in this subset (N. Dai et al. 2015; Christiansen et al. 2009). The effective sample size of the GWAS study controlling for BMI was ~30% smaller than the GWAS not controlling for BMI, explaining partially this reduction. However, we observed that across tissues the reduction was independent of their sample size (Figure 31B). When looking at the number of loci significantly associated with T2D only when using the GWAS controlling for BMI, we also saw no difference across tissues. From this we conclude that accounting only for BMI in the identification of IR or beta-cell dysfunction genes does not provide enough information to segregate genes from specific pathways.

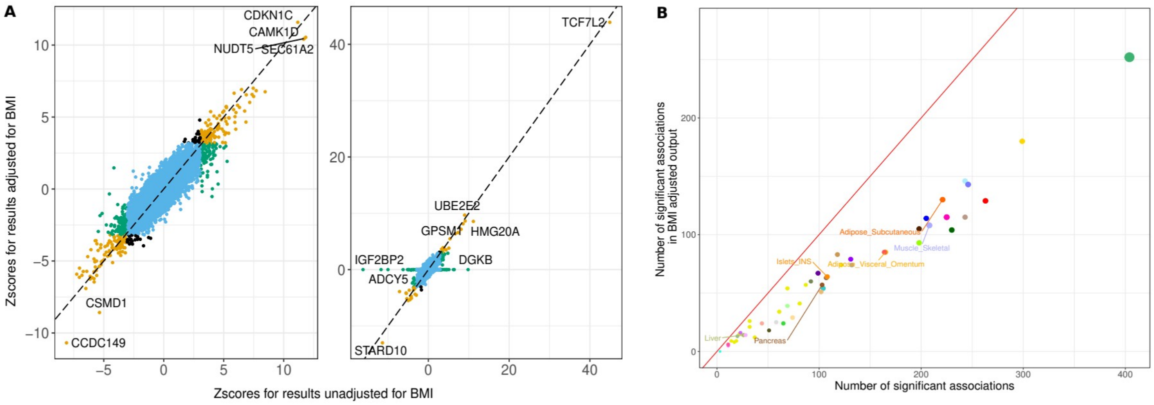


Figure 31: ***A. Comparison between results obtained with MetaXcan when adjusting for BMI or not.*** *Are shown, Zscores for DIRECT and InsPIRE. The colors are blue for non-significant associations in both outputs, orange for both, green for only unadjusted results and black for adjusted results only.* ***B. Comparison of the overall number of significant associations for each tissue when controlling for BMI.***

Genes specific to subtypes of T2D

T2D is a heterogeneous disease beyond the IR or beta cells dysfunction categories, involving different presentations; this influence choices in terms of study design and how cases should be defined. An inclusive definition of a case can help with recruitment, facilitating larger studies, while a more specific definition can allow the discovery of disease loci specific to particular aspects of disease. Recently, Ahlqvist et al. 2018 suggested that diabetes could be considered as five disease subtypes, namely severe autoimmune diabetes (SAID) also referred to as type 1 diabetes (T1D), severe insulin deficient (SIDD), severe insulin resistant diabetes (SIRD), mild obesity related (MOD) and mild age related diabetes (MARD). A GWAS study, found genetic signals specific to the particular five subtypes and not discovered using larger GWAS studies into T2D, despite the smaller sample size for the subtype GWAS (10,927 patients included compared to ~1 million)(Mansour Aly et al. 2021). Applying TWAS methods, we found that the size of the GWAS study also had an impact on the number of loci that we discover highlighting the importance not only of the size of the reference panel but also of the GWAS study used to calculate these associations scores. Similarly to the main analysis, DIRECT displayed the higher number of overall significant loci (42) compared to pancreatic islets resulting in 15 loci and between 3 and 37 for GTEx. When looking at the subtypes individually, we noticed that the SAID subtype TWAS uncovered the highest number of significant loci with 132 total unique genes, followed by MOD (17), SIDD (13), SIRD (11) and MARD (8). We replicated all significant loci reported in Mansour et al such *HLA*- loci in SAID, *TCF7L2* in SIDD, MOD and MARD and *ZNF503* in MOD. Moreover, we found that genes such as *TCF7L2* were only detected in disease related tissues such as pancreatic islets related to beta cell dysfunction mechanisms and associated here with SIDD. Among the new discovered loci we found *SIK3,* associated with SIRD using DIRECT blood, a gene previously identified in relation to insulin resistance (Uebi et al. 2012) or *ABO,* associated with MOD, a known T2D and obesity related genes (Siddiqui, Soni, and Khan 2019). DIRECT, as the largest dataset discovered more genes compared to the other tissues individually. However, for certain disease subtypes some tissues outperform the biggest reference dataset, such as SIDD, MOD and MARD where tibial nerve or pancreatic islets found more loci. Overall, the limited GWAS sample size drastically reduced the number of significant associations, with TWAS methods showing larger power to find relevant loci, even for disease subtypes.

Discussion

Our findings show that the relevance of the tissue does not play a large role in the numbers of genes discovered by TWAS methods, instead the sample size of the reference panels was the most important determinant. Not only did we find more genes using the largest available gene expression reference dataset, we also reported a substantial increase in associated genes by using a larger GWAS dataset for T2D than previously published (from 873 to 1,818 genes) (Vujkovic et al. 2019). However, we also saw some evidence that our power to discover causal genes may be beginning to become saturated at around the size of the DIRECT reference panel (~3000 samples). This sample size is feasible for less accessible tissues by combining datasets; for example recently a consortia has mapped eQTLs using a dataset of 2,970 brain cortex samples from individuals of European ancestry (de Klein et al. 2023), while others have combined datasets of pancreatic islets to obtain larger sample sizes (Alonso et al. 2021). It would be interesting to investigate if tissue relevance is more important when using reference datasets of this size, with increased ability to map tissue specific signals.

We built our predictive models using sets of curated independent eQTLs for DIRECT, GTEx and InsPIRE. Since, MetaXcan is calculating scores per gene based on overlaps between the models and the GWAS summary statistics, our upper detection limit will be the number of eQTLs present in each reference panel (DIRECT : 59972 , GTEx : 159236 for all tissues, InsPIRE : 4639). Therefore, we based our results on non-coding variations since eQTLs are in most cases located in non-coding regions of the genome such as enhancers or promoters. Where GWAS is focusing on coding variants by assigning these significant variants to nearby genes, TWAS is focused on finding and linking regulatory regions of the genome with changes in gene expression and downstream effects on the disease risk.

Comparing the performance of different omics (in this case transcriptomics and antibody based proteomics) to identify mediating molecular phenotypes, we observe that proportionally these assays have comparable power. However, the genome-wide nature of RNA-seq does mean that considerably more hypotheses can be tested using a single assay, while the power of proteomics assays may be boosted by the pre-selection of proteins directly relevant for disease. Though it was not significant due to small numbers, we observed a large odds ratio for enrichment when looking at expression and protein associations. This may be partly due to TWAS methods only considering genetic effects in cis, or local to the gene, and so excludes additional effects such as post transcriptional modifications including alternative splicing where different isoform of an mRNA can lead to different translations. This also suggests that approaches which look to combine eQTL and pQTL information when identifying gene targets could be reusing the same information (Mountjoy et al. 2021), obtained using the two different assays, which could lead to an overestimation of a gene’s suitability as a target. An alternative approach could focus on trans effects on proteins, less likely to be mediated by mRNA.

Type 2 diabetes is a complex disease involving different processes and tissues. Among these processes, insulin resistance is related to obesity and tissues such as skeletal muscle and adipose. By controlling for obesity/BMI we expected to remove genetic signals related to these specific tissues, but instead we observed a uniform decrease across all tissues, including pancreatic islets known to be related to beta-cell dysfunction. One reason could be that the original GWAS mainly included individuals of European ancestry, as it is known that the development of T2D in such cohorts is less driven by beta cell dysfunction than cohorts of other ethnicities (Siddiqui et al. 2022). A consequence of this could be that the study had greater power to discover insulin resistance related causes of T2D. However, this may also reflect the fact that beta-cell dysfunction is related to obesity as well, and in particular to the accumulation of liver fat (Inaishi and Saisho 2020). To produce a better list of genes understanding particular processes which drive the development of T2D, studying particular populations affected in greater proportion by these forms of T2D constitutes another approach, alongside expanding reference datasets to reliably map tissue specific effects in inaccessible tissues.

This study has discovered a large number of genes across a large number of tissues, in part due to the large sample size in both the original GWAS and the reference panels used. Because T2D is a highly heterogeneous disease, involving many tissues and many genes, this means that many of the implicated loci will have low effect size and be of unclear biological function. Studies of T2D have addressed this heterogeneity by constructing subtypes based on clinical phenotypes (Ahlqvist et al. 2018), and shown that these subtypes have different genetic associations (Mansour Aly et al. 2021). However, because of the necessity of collecting extra clinical phenotypes on these individuals, GWAS studies into subtypes are more limited in sample size, meaning our ability to find causal loci for these subtypes was highly reduced. However, we managed to find associations with each of the subtypes in various tissues and DIRECT, similarly to the main results, found the biggest number of significant loci. We also identified both already known but also potential novel loci for these subtypes. Indeed, in the case of insulin resistant diabetes (SIRD) we identify - *SIK3* – not previously reported in T2D, which could be further investigated downstream as well a *ABO* associated with MOD. This suggests that a strategy of using carefully defined disease phenotypes with expression data that prioritizes sample size over tissue of relevance can identify candidate loci for specific disease processes.

In this paper we have demonstrated that eQTL information from current bulk RNA-seq datasets are insufficient for answering questions on causal tissues for disease and tissue specific processes, a fact that others have also noted (Finucane et al. 2018). eQTL studies using single cell RNA-seq data from multiple individuals are beginning to be produced, these may have the potential to identify cell and tissue specific effects and thus inform on tissue specific disease loci (van der Wijst et al. 2018; Kang et al. 2018). However, these studies have lower power than studies in bulk tissue, and so this potential may not be realized in the near future (Cuomo et al. 2021). Concurrently, groups are producing omics data using tens of thousands of participants, with the potential to identify even more disease related loci, in particular in conjunction with smaller GWAS studies with carefully defined phenotypes. Discovery of causal loci and thus gene targets is increasingly driven by synthesizing multiple sources of evidence, functional, genetic and molecular, and we have shown here how studies in accessible but not directly disease relevant tissues can be a valuable source of such evidence.

Methods

eQTLs from three studies were used develop models to identify genetically predicted gene expression. The larger study DIRECT (Diabetes Research on Patient Stratification), with 3,029 samples, included 59,972 independent eQTLs from whole blood (https://doi.org/10.5281/zenodo.4475681). This cohort is composed of both pre-diabetics and newly diagnosed T2D patients with blood sample collected from venous blood. All characteristics and analyzed phenotypes are described in a previous paper (Koivula et al. 2014). The same study identify 1592 independent pQTLs that were used for protein models. For pancreatic islets models of expression, we used InsPIRE (Integrated Network for Systematic analysis of Pancreatic Islet RNA Expression, n =420), which identify 7741 gene level independent eQTLs after a eQTL analysis using fastQTL (https://zenodo.org/record/3408356). Sample collection, analyses and description of data produced by the consortium has been described in (Viñuela et al. 2020). For other tissues, GTEx (Genotype-Tissue Expression) v8 eQTLs were downloaded form the GTEx Portal (https://www.gtexportal.org/home/datasets), including 23,268 independent eQTLs from 49 tissues (73 to 706 samples). All methods from QTL studies to fine-mapping and functional analyses are described in the main paper for this study (Aguet et al. 2017).

GWAS summary statistics for T2D from the DIAGRAM study, adjusted and not adjusted for BMI were downloaded (https://diagram-consortium.org/downloads.html), selecting the dataset of European background and the meta-analysis of 32 GWAS with 74,124 cases and 824,006 controls (Mahajan et al. 2018). In addition, we used summary statistics from (Spracklen et al. 2020) which performed a meta-analysis gathering 23 studies in the 1000 Genomes phase 3 from the Asian Genetic Epidemiology Network (AGEN) consortium (https://blog.nus.edu.sg/agen/summary-statistics/t2d-2020/). Summary statistics were also lifted over to the GRCh build 38 to be analysed in the context of the GTEx reference panels.

The TWAS main analysis was performed using MetaXcan as described here: https://github.com/hakyimlab/MetaXcan. First, and using eQTL variants IDs and betas as weights we derived the MetaXcan models for gene expression prediction. The covariance between each pair of variants included in each gene-model was then calculated using R base *cov()* function and stored in matrices. Second, using as inputs the models, covariances and GWAS summary statistics we derived associations between genetically predicted gene expression and T2D. Pvalues for associations were corrected for multiple testing using the p.adjust() function in R and the Benjamini-Hochberg method as well as for the rest of the follow-up analyses.

In order to genes found nearby GWAS significant variants, we first selected the list of significant genes for each tissue and extracted the chromosome and positions for them. After doing the same for the GWAS significant variants, we applied the *intersect* command from the bedtools suite (Quinlan and Hall 2010) to overlap the two source files in order to filters genes found in a 1MB window around these selected GWAS variants. Then, in R, we used the base odds.ratio() function to get the enrichment of each tissue compared to DIRECT which was selected as the reference for this analysis.

To further investigate if genes identified were known T2D genes we used the DisGeNET platform (https://www.disgenet.org/home/). This online resource is a database of genes and variants associated to diseases publicly available which includes references from GWAS analyses, curated repositories and scientific literature.

We used the mr-egger function from the *MendelianRandomization* package in R (https://CRAN.R-project.org/package=MendelianRandomization) which is applying a MR method based on the Egger regression (Bowden, Davey Smith, and Burgess 2015). This method requires for each tested gene the betas and standard errors from the GWAS and the QTL analysis. We tested only significant genes with three or more variants included in the model and estimated MR pvalue and heterogeneity pvalue.

Chapter 3 : Gene expression changes after medication, effects and efficacy of metformin

T2D symptoms and treatments

Type 2 Diabetes (T2D) is a metabolic disorder that is characterised by elevated blood sugar levels. This condition, primarily rooted in insulin resistance and impaired glucose regulation, can have profound consequences on affected individuals if left untreated. The main consequence of T2D is the persistent elevation of blood glucose levels, as a result of the body's diminished ability to respond effectively to insulin, a hormone critical for glucose uptake. Over time, increasing high blood sugar levels inflict damage on various organ systems, leading to severe complications such as cardiovascular disease, neuropathy, retinopathy, and kidney dysfunction (Ekoru et al. 2019). These comorbidities represent a major health burden and they raise the necessity of early detection and intervention of T2D as its progression, if untreated, can have lethal consequences for patients.

Following diagnosis (usually by blood tests or an oral glucose tolerance test), the management of T2D typically involves a number of interventions, including lifestyle modifications, with exercise and dietary adjustments assuming pivotal roles. Regular physical activity not only enhances insulin sensitivity but also aids in weight management, a crucial factor in T2D control (Wilding 2014). Coupled with exercise, dietary changes focusing on balanced carbohydrate intake, fibre-rich foods, and controlled portions contribute to stabilising blood sugar levels. However, if T2D remains uncontrolled despite these primary measures, a secondary line of treatment comes into play. This phase often entails medication, and various drugs are available to regulate glucose metabolism. These medications span different classes, including oral antidiabetic drugs, insulin therapy, and newer innovations like glucagon-like peptide-1 (*GLP-1*) receptor agonists and sodium-glucose cotransporter-2 (*SGLT-2*) inhibitors. The second line of treatment is tailored to each patient's specific needs and may involve a combination of medications to achieve optimal glycemic control (Inzucchi et al. 2012; American Diabetes Association 2019).

Metformin, the most used medication for T2D

Metformin is a widely prescribed medication for T2D, which operates through multiple mechanisms to effectively manage blood glucose levels. Primarily, metformin enhances insulin sensitivity by promoting glucose uptake in muscle cells and adipose tissue (Zhou et al. 2001). It also restrains hepatic glucose production, a key factor contributing to elevated blood sugar in T2D patients. This is achieved by inhibiting gluconeogenesis, the process through which the liver produces glucose from non-carbohydrate sources (Madiraju et al. 2014). Metformin's influence on intestinal glucose absorption further contributes to its overall glucose-lowering effects (Bailey, Wilcock, and Scarpello 2008). Despite its established clinical efficacy, the precise molecular pathways by which metformin acts remain unknown. One prevailing theory centres on its activation of AMP-activated protein kinase (AMPK), an enzyme responsible for sensing cellular energy levels (Zhou et al. 2001). It is proposed that activation of AMPK by metformin triggers a cascade of events that regulate glucose metabolism and insulin sensitivity. However, the exact interplay between metformin, AMPK, and other potential mediators is the subject of ongoing investigation. A better understanding of how metformin works could result in more targeted prescribing practices, and also suggest new therapeutic compounds.

One tool to better understand the molecule processes triggered by the administering of metformin is to look at factors that affect the efficacy of the drug. The efficacy of metformin is known to vary greatly across individuals. Factors such as genetics, age, underlying metabolic conditions, and other medications contribute to this heterogeneity (Ilias, Rizzo, and Zabuliene 2022). Concentrating on genetics, GWAS studies have uncovered genetic regions that are associated with metformin efficacy (Zhou, Pedersen, et al. 2016). Understanding which molecules are affected by these regions can identify the molecules that interact with the drug. Frequently with pharmacogenetics, this has been a simple process: many pharmacogenetic variants are in coding regions of genes with known function. For example, hundreds of pharmacogenetic coding variants have been identified in the *CYP2D6* gene, which alter the enzyme's ability to metabolise a wide range of drugs (Zhou et al. 2008). However, recent studies into metformin efficacy have implicated non-coding regulatory variants, these are harder to interpret as there is usually uncertainty on which gene is affected (Xie, Hanson, and Sinha 2019).

Metformin is also accompanied by a spectrum of side effects that can influence patients' adherence to treatment. Among the most serious side effects are cardiovascular effects, gastrointestinal disturbances, including nausea, vomiting, and diarrhoea, which can lead some individuals to discontinue the medication (Zhang et al. 2020). While many of these side effects can be managed, they have an impact on the ability of metformin to control glucose levels. One way to study how these side-effects are produced would be to look at the molecular consequences of administering metformin in a differential gene expression analysis. Unlike the analysis presented in chapter one, where we looked at differential expression based on phenotypes inherent to the sample population, here we look at differential expression in relation to an intervention, the administering of metformin. By looking at the downstream consequences of this intervention, it could be possible to identify the genes that are responsible for the adverse effects.

Study aims

This chapter will look to understand the molecular consequences of metformin by taking these two approaches. Firstly, using genetic data on metformin efficacy, we will attempt to learn the genes which are associated with this. We will concentrate on looking at understanding regulatory variants which affect efficacy, and take two approaches, one based on only genetic information and one that combines genetic and transcriptomic information to link genes to drug efficacy. After this, we will look at genes that are differentially expressed under medication. These are more difficult to interpret, the consequences we observe could be how the drug functions to control glucose, or they could be separate from the therapeutic action, possibly driving side effects.

Comparison of three approaches : MAGMA, TWAS and DGEA

To do this we will use data from GTEx, DIRECT and K. Zhou, Yee, et al. 2016, a GWAS study into the efficacy of metformin. This study was an extension of a previous GWAS study on metformin using the GoDARTS cohort (n = 1372 participants of European ancestry) (Zhou et al. 2011). Efficacy of metformin was estimated based on the percentage reduction in HbA1c levels over a period of 18 months after the medication was started. They found a SNP, rs8192675, near the *SLC2A2* gene which was associated with reduction in HbA1c, and showed that this variant was an eQTL of *SLC2A2* in the liver. After replication within other cohorts (UKPDS and MetGen), the estimate of effect size for this gene was a reduction of 0.17% in HbA1c per allele [$p=6.6\times10{-}14$]).

To link genes to metformin efficacy based on genetic information we will use Multi-marker Analysis of GenoMic Annotation (MAGMA), a tool to identify associated genes based on GWAS summary statistics (Leeuw et al. 2015). MAGMA links genes to GWAS phenotypes based purely on genetic information. Summary statistics from GWAS studies are mapped to genes by defining a genomic window around the gene (35kb upstream and 10 kb downstream for our analysis). Then it aggregates all the SNPs within each gene to calculate a p value based on the GWAS summary statistics in the region. This p value is then converted into an equivalent z score, which represents the strength of associations between gene and the phenotype. Correcting for LD structure is also important to avoid under- or overestimations of *P* values. To solve that, MAGMA is using an external LD reference panel to calculate correlation matrices between genes to be included in the main model to correct for this effect. A subsequent step of MAGMA is to use these z scores to estimate enrichment for sets of multiple genes, which can be based on pathway information, tissue specific patterns of expression, or other information.

In contrast with MAGMA, TWAS combines both transcriptomic information with genetic information to identify these genes. This methodology is particularly relevant in the context of drug efficacy, offering insights into the genes that determine how individuals respond to medications. TWAS looks for evidence that the same genetic variants that affect medication efficacy also affect gene expression in consistent ways. In doing so it gives information about the direction of effect of the associations, proposing whether higher or lower quantities of the gene would produce a greater or smaller drug response. Namba et al. 2022 used TWAS to study multiple common and rare diseases and their medication categories and showed that this approach identified drug targets of asthma. By pinpointing specific genes whose expression patterns influence drug responses, TWAS offers a promising avenue for advancing precision medicine and optimising therapeutic interventions tailored to individual patients.

Finally, differential gene expression analysis (DGEA) provides another tool to better understand the genes which are involved in treatment effects and side effects. However, a differential gene expression analysis of medication has distinct differences with that of disease explored in Chapter 1. For disease aetiology and progression, the study is by definition observational in nature. Across the cohort, individuals will differ by disease state, differences in expression can be related to this state. But this means that these studies suffer from confounding, where it is difficult to tell if the change in expression is due to disease or another factor correlated with disease (Assimon 2021). In contrast, when looking at the impact of medications, the researcher has the ability to intervene on the population, meaning that with the correct study design causality can be inferred. Randomised controlled trials and crossover trials have revolutionised medicine, by providing researchers the tools to conclusively decide if a medication is effective at treating disease (Spieth et al. 2016). In a standard randomised controlled trial participants are randomised at the beginning of the trial to either receive the medication or a placebo. The randomisation is to ensure that other factors that could affect treatment outcome are as likely to be assigned to the control group as the treatment group. The disadvantage of randomised controlled trials is the expense: studies have to be designed and run to test a single hypothesis, limiting the ability of the study to answer other research questions. In this chapter, the differential expression analysis will be based on an observational study. Individuals were followed up over a period of time, but the decision to start treatment was made on clinical grounds. This means that we will need to take care when interpreting the results: as treatment was started in response to a worsening of disease, either the disease or the medication could be responsible for changes.

Summary of the results

To summarise, this chapter presents three analyses aimed at understanding the molecular processes involved in metformin action. Firstly, using GWAS summary statistics, we will look at genes which are associated with metformin efficacy. Secondly, we will investigate the same question but using methods which combine GWAS summary statistics with transcriptomic information. The advantage of this approach is that we can use it to learn about the tissue of action and the direction of effect of the associated genes: whether an over-expresssion or under-expression of the gene is associated with an increase in efficacy. Finally, we present the results of a longitudinal analysis of differential gene expression in response to medication, to learn which genes are affected by the introduction of medication.

MAGMA results, looking for metformin efficacy genes

We used MAGMA and GWAS summary statistics from K. Zhou, Yee, et al. 2016 to identify genes with strong genetic signals for association with metformin efficacy in the nearby region. We found 893 significant genes (FDR < 0.05, Figure 32) associated with metformin efficacy. We compared this list of genes with two different sets from two studies, one based on the original GWAS conclusions, the other based on a text mining approach to identify metformin related genes.

Surprisingly, the set of significant genes did not include the genes implicated by the original GWAS study, or a follow up study from a few years later (*ATM* or *SLC2A2*). In the original study, among the variants significantly associated with metformin efficacy, none of the corresponding genes reached the significance threshold in the MAGMA results. The second gene set was from Dawed et al. 2017 where they used a text-mining based method to prioritise and assign scores to genes related to the pharmacokinetics and pharmacodynamics of metformin. However, none of these genes were significant in our analysis. One of the possible explanations for this outcome could be that we defined a window around the gene to capture regulatory effects on metformin efficacy; it is possible that the size of the window dilutes the effect of the promoter variants reported in this paper, meaning that these genes are no longer found significant. It is also possible that methods based on text mining around pharmacogenetic variants will focus on protein coding changes to enzymes, which impact the ability to metabolise the drug, and will miss regulatory effects we focus on here.

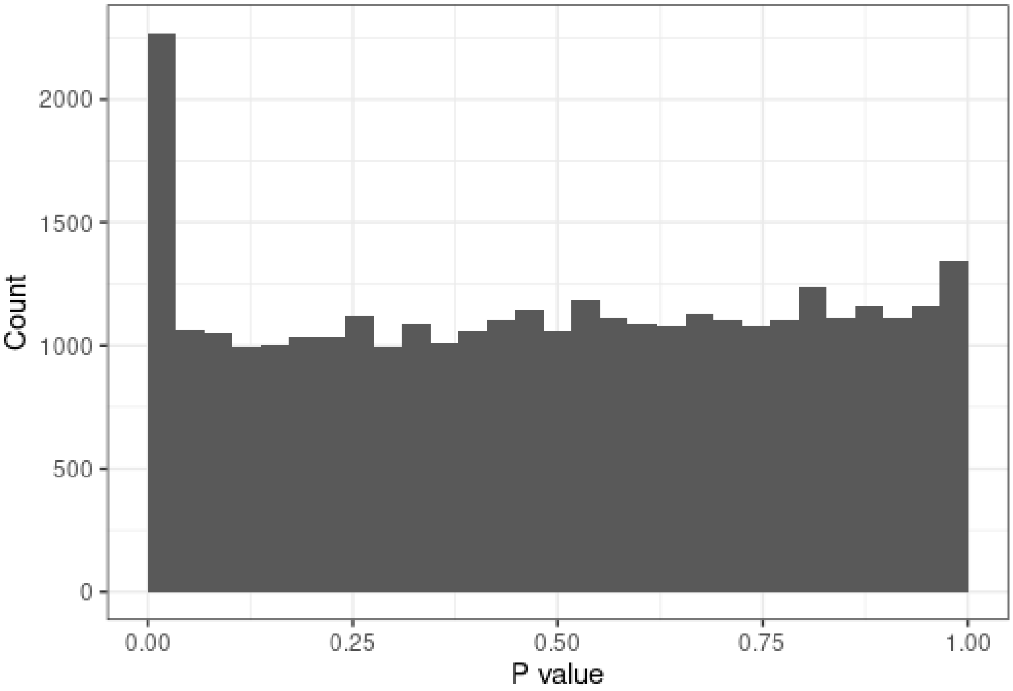


Figure 32: ***P value as calculated by MAGMA for association between 33053 genes and metformin efficacy.*** *The peak on the left represents the 893 genes that were found to be significantly associated (FDR=0.05).*

To understand better the genes that were implicated in this analysis, we looked into the tissues in which they were expressed. Finucane et al. 2018 proposed that genes specifically expressed in particular tissues were more likely to be involved in disease development, and that by studying this specific expression it was possible to infer the relevant tissue for the disease. Using the GTEx data, and defining a gene to be specifically expressed if it was in the top 10% for the strength of evidence of differential expression compared to all other tissues in GTEx, we found the associated genes to be enriched in 6 tissues : prostate, colon, oesophagus as well as brain basal ganglia, amygdala and anterior cortex (Figure 33). These tissues do not include those commonly thought to be involved in metformin action, such as the liver and the gut.

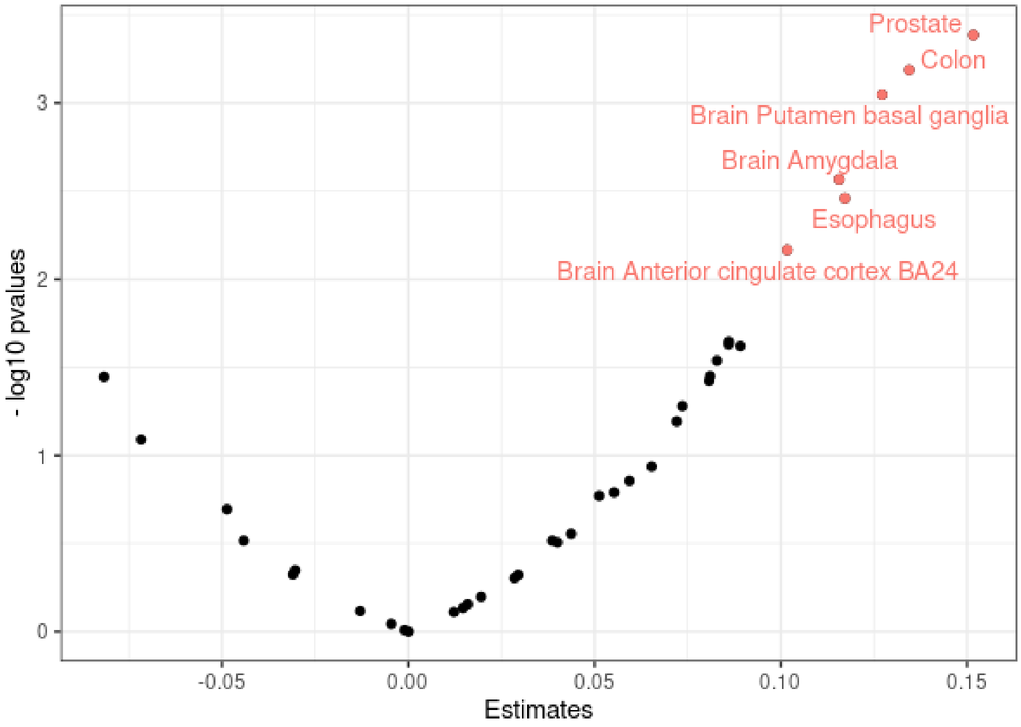


Figure 33: ***Distribution of estimates of enrichment vs -log10 p values for GTEx tissues enriched for metformin efficacy genes using MAGMA.*** *The significant tissues are coloured in red.*

TWAS results, sample size still playing an important role

Next, we applied the TWAS methodology presented in the previous chapter to the same set of summary statistics analysed with MAGMA. In this way we are looking for genes where there is evidence that genetic effects on metformin efficacy match with genetic effects on expression. Using the models created based on DIRECT and GTEx, we identified 43 significant genes (FDR < 0.05). Among these genes, 36 were found using DIRECT and 7 with GTEx (2 in breast mammary tissue and 1 in uterus, salivary gland, sigmoid colon, brain hippocampus and coronary artery, Figure 34). As was seen in the previous chapter, there was no relation between the relevance of the tissue and the number of gene associations discovered. Similarly with chapter 2, the associations were heavily driven by the sample size, with the vast majority found using the DIRECT reference panel.

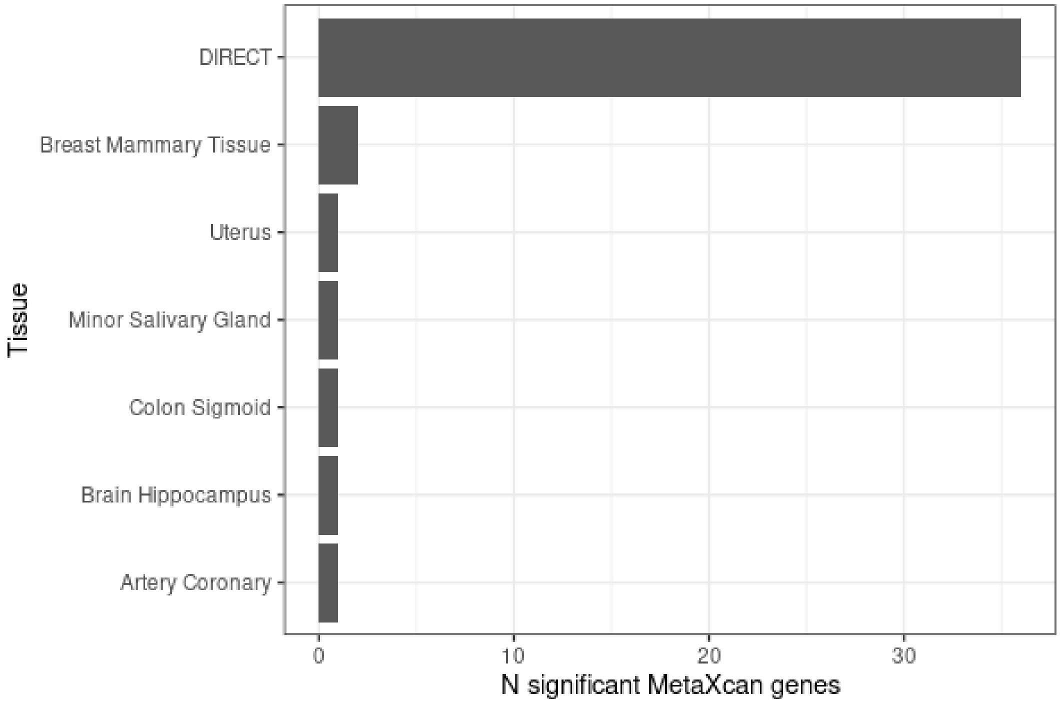


Figure 34: ***Number of significant gene associations with metformin efficacy found using MetaXcan in the tissues of GTEx and DIRECT.***

In the previous chapter, we produced a list of genes with a putative causal role in the development of diabetes. We hypothesised that metformin could target some of the genes, and that these genes could also be involved in metformin efficacy. Overall, three out of the 43 genes were also associated with T2D. Two of them (*BCL7A* and *PDGFD*) had the same direction of effect and one (*MTMR11*) opposite direction of effect (over-expression was associated with an increase in metformin efficacy and a decrease in the risk of T2D).

DGEA results, direct effects of metformin on gene expression

Finally, we ran a differential gene expression analysis using longitudinal context to study the effect of metformin on gene expression. The model we used is described in the methods and is based on comparing expression before and after the initiation of metformin treatment, controlling for age, BMI, sex, and technical covariates. Though we found no DEGs at a standard FDR threshold of 0.05, at a weaker FDR of 0.2, 356 genes were differentially expressed (Figure 35).

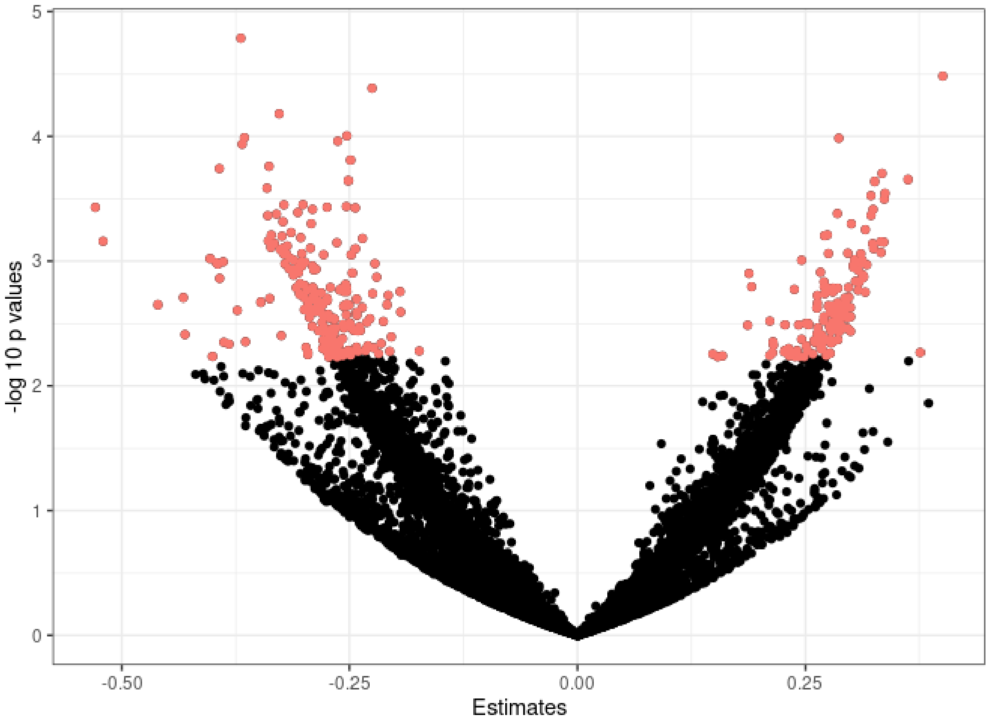


Figure 35: ***Volcano plot of regression estimates vs -log10 pvalues from the differential gene expression analysis.*** *Genes in red are significant DEGs at FDR 20%.*

Of these 356 associated genes, 4 had previously been reported as showing differences in protein levels in response to metformin: *IDO1* and *TAGLN* had been reported in Giusti et al. 2022 and *IL5RA* and *MZB1* in Zhong et al. 2021. In each of these 4 cases, after metformin intake we saw significant reduction of the gene expression between the two timepoints, similar to the reduction previously reported for protein levels (Figure 36). Overall, 144 were overexpressed in response to metformin and 212 were underexpressed. In addition to results observed in the literature on metformin effects, we also detected changes in genes known to be related to metformin such as the Sirtuin 3 (*SIRT3*), for which we observed a decrease in expression (Figure 37).

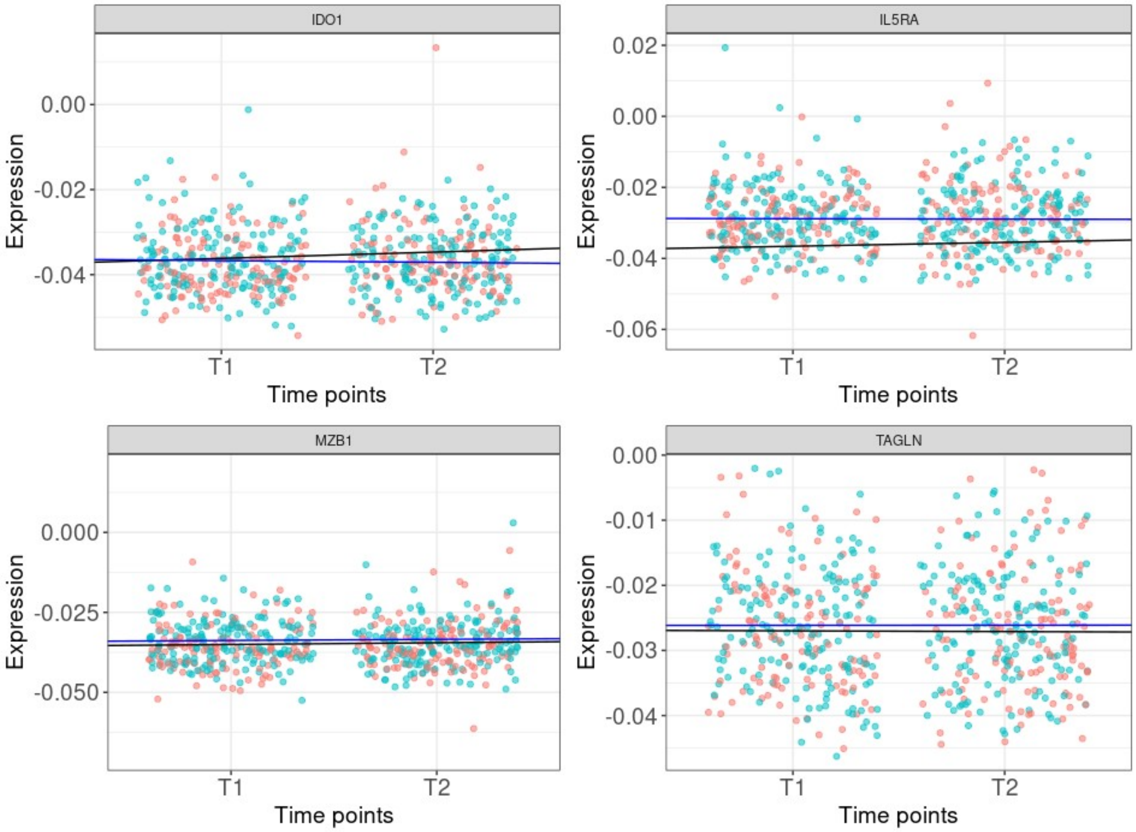


Figure 36: ***Comparison of expression between two timepoints for the 4 genes replicated in the literature.*** *Data is separated between individuals taking metformin (blue) or not (red). Black line is the regression line for controls and blue line for individuals on metformin. Differential expression statistics : IDO1 : Estimate -0.26 Pvalue 0.156 ; IL5RA : Estimate -0.23 Pvalue 0.166 ; MZB1 Estimate -0.25 Pvalue 0.168 ; TAGLN Estimate -0.24 Pvalue 0.168*

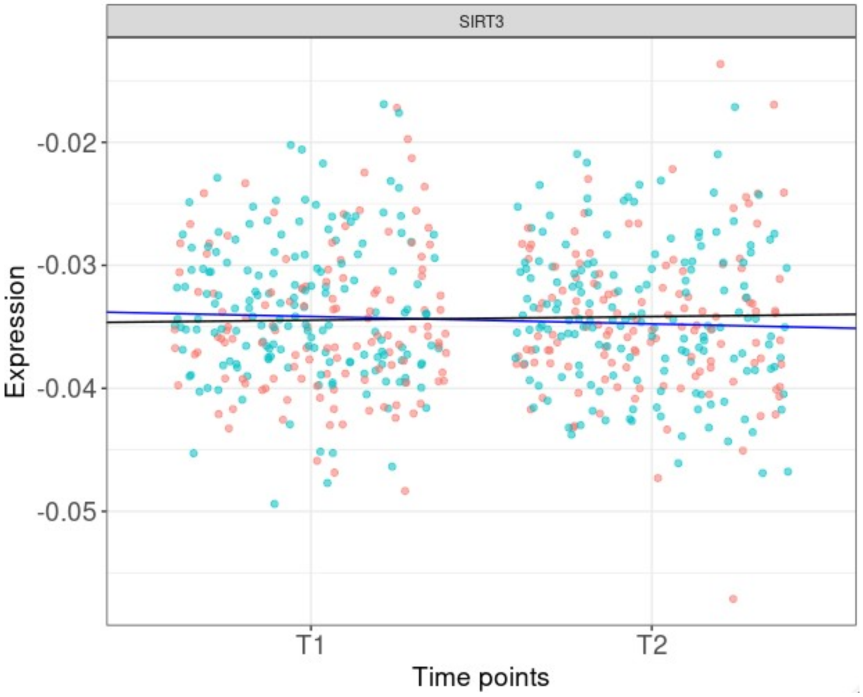


Figure 37: ***Expression of SIRT3 for individuals on metformin and controls between the two timepoints.*** *Colours are the same as in previous figure. The black line is the regression line for controls and the blue line is for individuals on metformin. Estimate -0.34 Pvalue 0.15*

We also looked at the overlaps between these DEGs and our results in the previous chapter on causal genes associated with T2D, as genes that cause T2D are potential targets of treatments for the disease. Using the largest study in the same tissue, we found 14 genes in common, with a total of 20 genes differentially expressed and associated with T2D across all tissues (Figure 38). When investigating direction of effect, we found that in 14 out of 20 genes the two analyses produced opposite directions of effect. If overexpression increased risk of disease, then initiation of metformin would reduce expression, consistent with these genes being involved in the processes by which metformin treats disease severity. However, this result was not significant (p = 0.057, binomial test that same and opposite directions of effect were equally likely), possibly due to the small numbers of genes significant in both analyses.

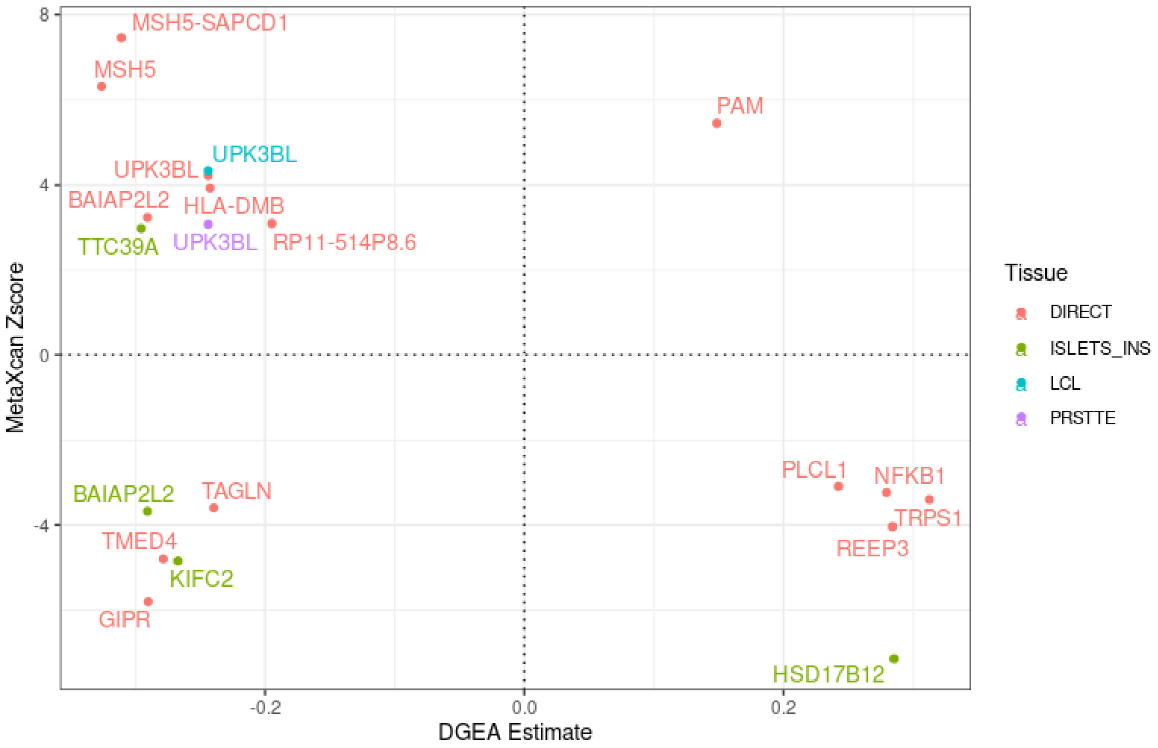


Figure 38: ***Effect sizes (beta estimates for DGEA and zscore for MetaXcan) of the genes in common between DGEA results and previous chapter's associations with T2D.*** *Colours are for tissues.*

Comparison between the three methods

When comparing the overlaps between the three methods, we found 3 genes in common between DGEA and MetaXcan results (*CD37, ZNF600* and *FNDC3A*), 1 between MAGMA and MetaXcan(*SIK3*) and 5 between MAGMA and DGEA (*ALOX15, SIRT3, SLC22A18, STON1-GTF2A1L, ZNF513*)(Figure 39). In none of these cases was this a significant enrichment (odds ratios/p values of 3.05/0.08, 0.83/1, 0.61/0.36 respectively). While the lack of overlap of DGEA genes with the other two methods can be explained by the fact that they are testing different hypotheses, it is more surprising to observe the lack of correspondence between the MAGMA and MetaXcan analysis (though there is some correlation between p values for association, Figure 40). Not only are both of these approaches searching for genes associated with metformin efficacy, but they use the same underlying GWAS data. However, MetaXcan has the further requirement that GWAS effects are mediated by regulatory processes: the lack of overlap suggests that the significant GWAS hits in this case do not have regulatory effects.

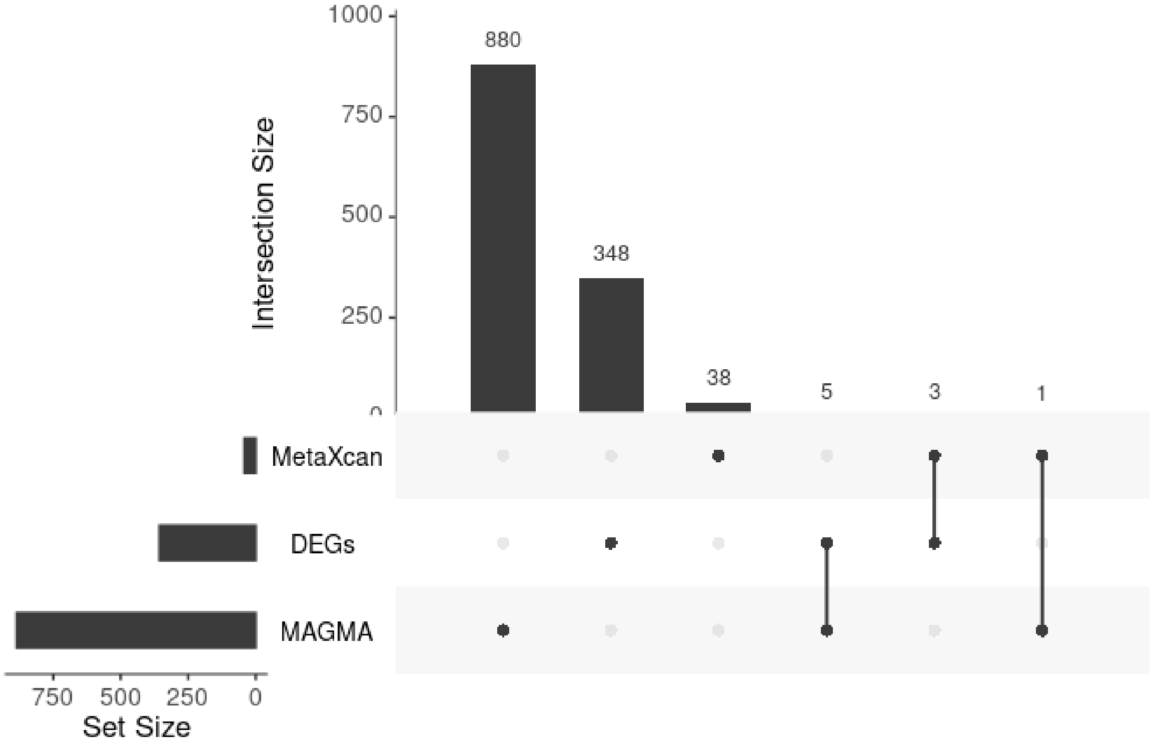


Figure 39: ***Upset plot of the number of significant genes and their overlap in each of the three analyses on metformin.***

Comparing the direction of effect of differential expression with the direction of increased metformin efficacy, we observe that for the three significant genes for the two methods, the direction is the same for one of them (*ZNF600*) This would promote a virtuous circle: we observe in these cases that metformin alters gene expression in ways that are beneficial to its action. Across all DEGs this virtuous circle is more often the case than otherwise (159/386 genes share the same direction of effect in the two analyses) but this is not significant when testing the hypothesis that concordance is random using a one-sided binomial test (P=0.13).

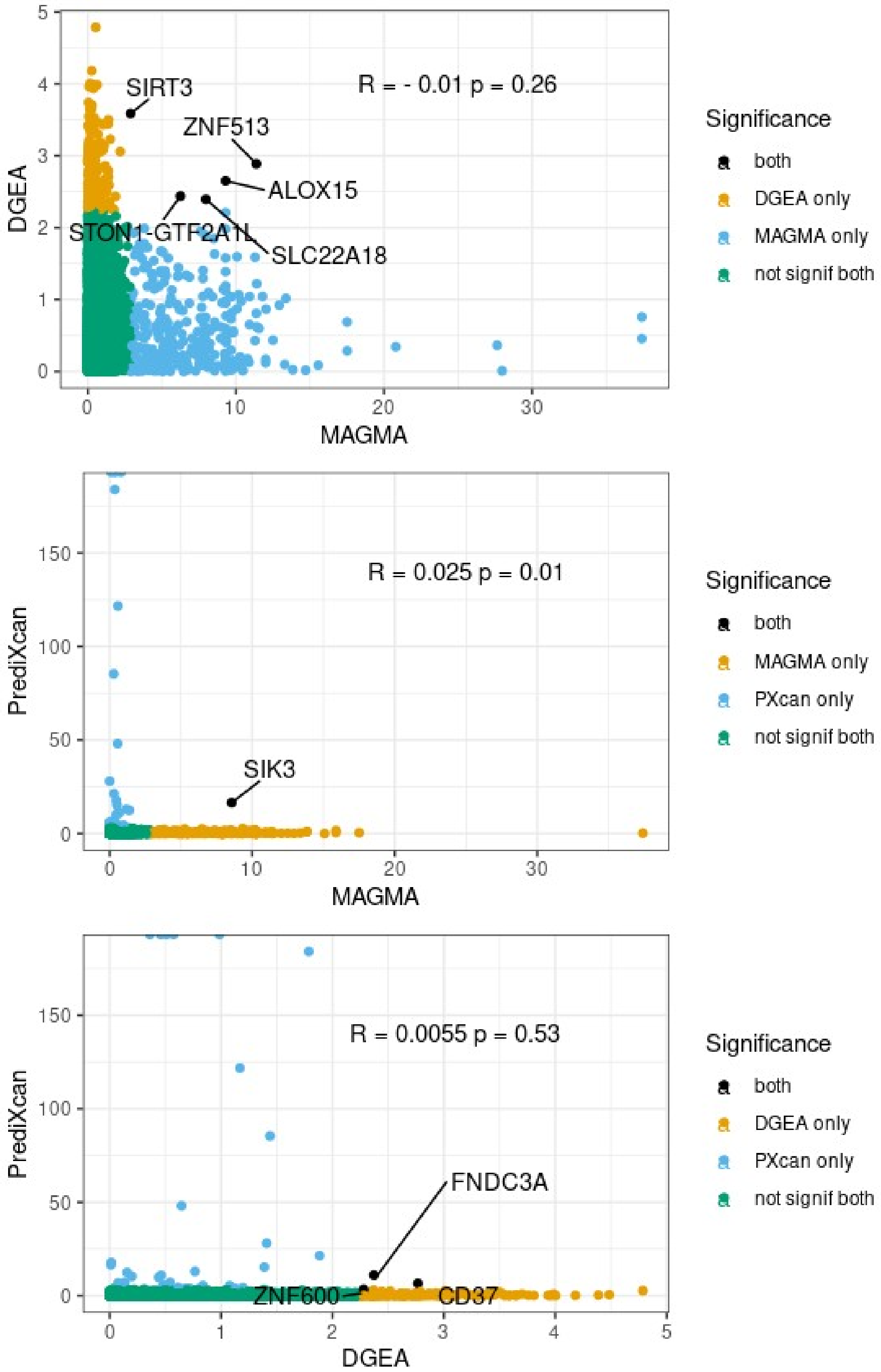


Figure 40: ***Comparison of -log10 raw p values for results in the three analyses composing this chapter.*** *Colours are red for genes significant in both sets, and green/blue for genes with significant adjusted p values only in one database. Spearman correlation rho and p values indicated.*

Discussion

We present in this chapter three analyses that investigate the effect of metformin at a molecular level. By using MAGMA and MetaXcan combined with a GWAS summary statistics for metformin, we have produced a list of genes implicated in affecting metformin efficacy. The DGEA instead looks at changes of gene expression in response to metformin, using longitudinal expression data and correcting for multiple factors. We found 893 significant genes with MAGMA, 43 significant genes with MetaXcan and 356 significant genes with DGEA.

The gene *SIRT3* was discovered by both the MAGMA and DGEA analyses. *SIRT3* has been linked to mitochondrial stress resistance and to longevity based on energy efficiency on low calorie diets (Dhillon et al. 2022). Previous publications have also discussed additional effects of metformin on patient's health and showed that not only it had therapeutic benefits on diabetes but also on ageing and longevity (S. Chen et al. 2022), and *SIRT3* is a possible explanation for these effects (H. Li and Cai 2023).

The MAGMA analysis identified the gene *SIK3* to be associated with metformin efficacy. *SIK3* codes for a salt-inducible kinase, which regulates AMP-kinases, one of the major actors in energy homeostasis. Studies have shown a decreased gene expression for this gene in obese and insulin-resistant individuals and metformin is also known to be an AMPK activator, providing a potential path for enhanced of glucose uptake through *SIK3* regulation (Säll et al. 2017; Uebi et al. 2012). One interesting point of discussion regarding MAGMA results is the significant enrichment of the two top tissues in the output, prostate and colon. Although not being tissues usually thought to be closely related to T2D such as pancreas or adipose, previous studies have shown interaction between metformin and risk of cancer in these regions of the body. These studies discussed a potential protective role of metformin on the risk to develop colorectal or prostate cancer (Ala 2022; Anwar et al. 2014). The fact that these two tissues are highlighted in our study shows not only the complexity of the mechanisms of action for metformin but also the potential for these two tissues in GTEx to be used to study more in depth effect of metformin on these tissue-related diseases such as cancers.

A number of the genes identified by the DGEA have previously been linked to T2D, and could be potential targets of metformin for its effect on the regulation of glucose homeostasis. For example, previous studies have found *LIPC* to be associated with regulation of cholesterol levels and an increased risk of diabetes (Guerra-García et al. 2021). A number of the differentially expressed genes are also involved in insulin signalling pathways, including *CD36, CPT1A,* and *GIPR,* key in insulin resistance and glucose absorption processes (Moon et al. 2020; Ren et al. 2021; Gasbjerg et al. 2018). Finally, ATP and NADP related genes are important for energy management, we found *MAPK6, ATP13A3, NAGK* with the potential to be involved in glucose intake and regulation (Loza-Valdes et al. 2022; ten Klooster et al. 2018). All of these genes are potential targets for the drug.

The DGEA was based on an observational study; individuals were placed on the drug due to a deterioration of their condition. This makes it more difficult to infer that the changes in expression are due to the medication, and not due to the disease, that we address by controlling for disease variables available to us. However, as individuals are not randomised to treatment, we can also expect them to be on other medication, and interactions between medications are not considered here. Drug-drug interactions are known to present a significant challenge when assessing the efficacy and consequences of a particular drug (Palleria et al. 2013). It would be interesting to compare our results with randomised studies into metformin, and we are in the process of reaching out to do so. Finally, drug response can be to particular populations as well (Burroughs, Maxey, and Levy 2002). The DIRECT study is restricted to Europeans and we do not know how these results will generalise to other populations.

Of our three analyses, MAGMA was clearly the most powerful, discovering the highest number of associations. However, this approach does have drawbacks, it produces only a link between the gene and the outcome, with no information on what the relationship is. This means that it misses directionality, and we cannot infer whether increasing or decreasing expression of the gene would have a beneficial effect on outcome. Tissue specific information is also unavailable for individual genes, and though global information on relevant tissue can be estimated using reference datasets, in our case the results of this were difficult to interpret. Pleiotropy and linkage contamination are also issues for this approach, genes may be implicated because they are merely in linkage disequilibrium with causal genes.

In contrast, MetaXcan combines GWAS information with information on gene expression to provide information both on directionality and, in theory, the relevant tissue. However, as we saw in the previous chapter, many genetic effects on gene expression are shared among tissues, meaning that this approach will frequently implicate the wrong tissue, especially if the wrong tissue has a greater sample size and is thus better powered. Finally, differential gene expression analyses can reveal genes directly affected by medication, but it is difficult to understand what the consequences of this intervention are, in terms of drug action and side effects. However, we did observe opposite effects of metformin on a number of genes linked to the development of T2D, suggesting that metformin targets many different genes or pathways in treating T2D.

One important point of discussion when comparing results from different methods such as MAGMA and TWAS based on eQTLs is the notion of gene boundaries. This will define the observational window for the analysis. For example, for MAGMA, the boundaries were 35 kbp upstream and 10kbp downstream the gene TSS where for eQTLs, the window was 1MB upstream and downstream of the TSS. This already changes drastically the ability to detect regulatory elements for a specific gene. For MAGMA, we will be able to detect mainly intragenic enhancers or promoters for example. With the much wider frame, although still considered in *cis*- so in a proximal distance, eQTLs are able to capture in addition to intragenic effects, intergenic and more distal regulatory elements of gene expression. Modulating the definition of these gene boundaries could also change the output of individuals methods as well as the overlap between results.

This chapter has focused on gene expression when trying to better understand the molecular changes in response to metformin. We have combined data from a number of different study designs: GWAS studies, linking genetic information to patient data, reference transcriptomics studies, where the transcriptome was collected from subjects without a particular disease, and a longitudinal omics study, where samples are collected as the disease progresses. But these represent a small fraction of the work on better understanding T2D and its treatment, which includes studies such as DIRECT, InsPIRE, MAGIC, PROVALID and many others (Alonso et al. 2021; Viñuela et al. 2019; C.-T. Liu et al. 2019; Eder et al. 2018). By adding further layers of curated clinical information to expanding multi-omics data we have the potential to better understand the factors influencing the success of metformin and other drugs on managing and treating T2D.

Methods

For the MAGMA analysis, summary GWAS statistics were provided from the K. Zhou, Yee, et al. 2016 study. The location of the transcription start site (TSS) of all annotated genes was extracted from the gencode version 19 annotation file and a window for each gene was defined as 35kp upstream and 10kbp downstream of the TSS. The dbSNP database of genetic variants was used to assign rs IDs to all variants with available summary statistics and the MAGMA pipeline used these IDs to annotate SNPs to the gene regions and calculate gene based P values. To perform enrichment analysis, genes were identified as specific to a particular tissue by performing a differential expression analysis, using a t test to compare expression in that tissue to all others in GTEx. Genes in the top 10% in terms of significance of this test were defined as tissue specific. The MAGMA pipeline used this annotation to calculate tissue based enrichment scores. Multiple testing was corrected for using *p.adjust* and a Benjamini Hochberg method.

We used the gene expression predictive models generated from chapter 2 combined with the GWAS summary statistics on metformin efficacy mentioned earlier and tested for association using the MetaXcan software(Barbeira et al. 2016). These models are based on independent eQTLs reported by the two consortia, using reported effect sizes as weights in the gene expression models. In order to correct for LD between SNPs in the GWAS summary statistics, we used a covariance matrix that was based on reference genotype data from GTEx v8 and DIRECT. Similarly to chapter 2, the GWAS summary statistics was lifted over from build 37 of the human genome to build 38 using the liftOver tool ('The UCSC Genome Browser Database: Update 2006 - PubMed', n.d.) to match the GTEx predictive models variant IDs. P values from the raw results were then adjusted using the *p.adjust* function and Benjamini Hochberg method.

The DIRECT study recruited 789 individuals who had recently received a diagnosis of T2D. Study participants made three visits to the clinic, at the beginning of the study, at 18 months and at 36 months. At each visit a blood sample was taken. This blood sample was sequenced to provide quantifications of gene expression as described in (Brown et al. 2023). Data at each time point was mapped onto a normal distribution independently, so that the data from each time point has the same distribution. We identified 151 individuals with T2D who were put on metformin between visits, either between the first and the second time point or the second and the third time point. For controls, we used 454 individuals not on metformin, and randomly chose a pair of timepoints in the same proportion to the time points available for starting metformin. We also extracted clinical and technical variables including sex, age, HbA1c levels, BMI, centre where the data was sampled, RIN score of the sample, insert size and GC mean of the sequencing. The model we then used for the linear regression was as follows : *exprs T2 ~ exprs T1 + medication + age T2 + BMI T2 + HbA1c T2 + sex + center + RIN score + GC mean + insert size* . After extraction of estimates and p values for each gene, we adjusted the p values for multiple testing using the *qvalue* package (https://github.com/StoreyLab/qvalue).

Discussion

Summary of the thesis

Type 2 diabetes is a complex disease, involving many different risk factors which all interact with each other. Two main processes are responsible for the development of the disease, beta cells dysfunction and insulin resistance, and both are influenced by multiple environmental and genetic risk factors. The aim of this thesis has been to explore the molecular factors that drive the development of disease, the molecular consequences of the disease, and the molecular processes involved in treatment.

In the first chapter, we used gene expression data produced by GTEx from 49 tissues to look for genes that were differentially expressed according to T2D status. We found that we had limited power to discover these differentially expressed genes, likely since GTEx was not a diabetes centred study. This is complicated by the fact that T2D has many comorbidities that need to be taken into account, and information on these for GTEx participants is not comprehensive. But we were able to show differential expression with a number of biological covariates, such as sex, age, and BMI, demonstrating environmental factors can influence gene expression.

In chapter two we focused on how genetic information from GWAS studies can be combined with reference transcriptome data to identify genes likely to cause disease. We showed that the major determinant of our ability to find disease genes was the size of the reference panel: using molecular data from directly relevant tissues such as pancreatic islets brought no increase in power but was necessary to discover some signals such as *TCF7L2.* We also showed that more targeted GWAS studies, with more homogeneous definitions of T2D subtypes can be useful in providing causal genes with a more obvious biological interpretation.

Finally, we looked at the effect of metformin treatment using three different approaches: using genetic evidence on factors that influence drug efficacy, combining this evidence with reference gene expression datasets to infer information on relevant tissues and effect directions, and by looking at genes which changed expression in response to metformin. Surprisingly, all three approaches produced different gene lists.

All three methods used were quite different not only in their design but also regarding the questions they were trying to answer. MAGMA and TWAS were looking at metformin efficacy where DGEA was focused only on difference in gene expression through time after medication therefore asking two separate questions. In addition to this, within the methods focused on metformin efficacy, MAGMA and TWAS are also relatively distinct. One is only gathering GWAS statistics to derive zscores from mean pvalues, the other is combining genotype and transcriptomic data with GWAS summary statistics to calculate gene – trait associations as well as giving information about directionality. Finally, the choice of the data, as always, is important as the low number of individuals selected for the DGEA in DIRECT could also explain the small numbers and the lack of overlap with the other results. The filtering step was relatively strict since we wanted individuals with two distinct point in time following each other and including a metformin intake in between these two points. Despite all this, the differential gene analysis produced two pieces of evidence that metformin targets several different genes and pathways. This analysis revealed a number of genes in different T2D related pathways showed differential expression is response to the treatment. We also observed that for the genes linked to T2D in the previous chapter, the application of metformin tended to have the opposite effect to the one that would provoke disease development, suggesting a protective effect.

Main areas of discussion

One of the difficulties we encountered when studying T2D was the wide range of confounders and comorbidities. For example, in chapter two we faced the decision on whether or not to control for BMI when searching for genes causal for T2D. Our concern was that controlling for BMI would remove signals where BMI was on the causal pathway: we did find fewer genes but believe that this is because the underlying GWAS studies had fewer participants. Accounting for confounders was particularly difficult in the differential expression analysis of chapter one. While information was available on other factors such as BMI, age and sex, this was less true for disease comorbidities such as cardiovascular disease, neuropathy and retinopathy. We discovered a large number of genes to be differentially expressed in the tibial nerve, we do not know if these individuals suffered from neuropathy.

A focus of this thesis has been about which tissues are necessary to study T2D. We saw in chapter two the benefits of a well powered study, even if the tissue is not directly relevant for the disease : increasing sample size increases our ability to find gene associations using TWAS and also means more genetic instruments are available for multiple instrument Mendelian randomisation. When studying complex diseases such as T2D we expect multiple tissues to be involved in disease development, perhaps meaning that picking the “correct” relevant tissue is less important. For subtypes of T2D, which may be more linked to a particular tissue or mechanism, it may be more important to collect the correct tissue. We could also see from chapter three, even if the tissue being studied is not the tissue the medication acts in, we still have the ability to discover differentially expressed genes.

T2D is a result of a complex interaction between a number of genetics, biological and environmental factors. Even with a simple hypothesis, for example identifying genetic variants associated with T2D, it is important to consider which biological factors should be considered in our analysis, and how they should be accounted for. We saw that correcting for BMI for example removed some signals, therefore we can hypothesise how this relates to specific biological processes such as insulin resistance and beta cell dysfunction. These considerations can also apply to other factors, such as age, ethnicity, and lifestyle behaviours such as smoking, alcohol and exercise. This also relates to how we view T2D, as one disease or as many, with particular risk factors applying to particular subtypes.

The concept of causality is also one which runs through this thesis. For all of our analyses we can ask the question whether the effects we observe are causes or consequences of the disease. Because transcriptome wide studies are based on genetic associations, and the associations on expression are near to the gene, there is an assumption that causality runs from variant to gene to disease. However, there is still the potential for finding non causal genes, because of factors such as horizontal pleiotropy and gene co-regulation. This means that best practice is to use methods such as colocalization or mendelian randomization to validate these results.

In contrast, when considering differential gene expression, in most cases researchers have assumed that differentially expressed genes are a consequence of disease and not a cause (Porcu et al. 2021). This is what we observed in the first chapter with a large amount of signals detected in tibial nerve. Yes this tissue was among the largest sample size in GTEx but we also know that neuropathy is one of the main co-morbidities of T2D with almost 50% of individuals with T2D developing diabetic neuropathy (Feldman et al. 2019). Although still rather unknown, it is hypothesized that high blood sugar levels can cause peripheral nerve damage and result in sensory disturbance. This is a good example of consequence of T2D captured with differential expression approaches. In addition to this, we need to consider also study design. When considering an intervention, such as application of a medicine within a randomised controlled trial, differentially expressed genes change due to the medication, implying causality. In chapter three we use data from an observational study. But if we believe that the medication is the major driver of change to the system, and confounders such as disease state are likely to have more modest effects, then we can propose the changes do result from the medication.

Our approach for medication looked at both genes which affected drug efficacy, and genes which were affected by the medication. We expected to see some agreement across these two sets of genes. When we observe that a gene is differentially expressed after treatment, we can ask the question, is this gene responsible for the positive changes in the patient? Answering this question can help us address issues of secondary effects and optimise existing medication, focusing on factors bringing improvements.

Potential for future studies

In this thesis, we have focused on genomics and transcriptomics. However, a complete understanding of the disease needs to model the full causal path from variant to disease, encompassing interactions between expressed genes and also introducing other molecular phenotypes such as proteomics and metabolomics. Recent studies have started to investigate T2D using these supplemental molecular phenotypes in T2D (Z.-Z. Chen and Gerszten 2020; Gou et al. 2022). Evolving technologies will also produce better quantifications of these phenotypes as well (X. Dai and Shen 2022). But we are already seeing the benefits of widening the types of molecular phenotypes we use. For example, Zaghlool et al. 2022 have identified clusters of proteins and metabolites specific to each of the five subtypes of T2D, raising the possibility of using these as new subtype biomarkers.

On the topic of subtypes, we used in this thesis information about subtypes of T2D to look at genes associated with specific subtypes. Another approach which also works towards the understanding of GWAS signals could be GWAS clustering in order to link genetic variants to pathways specific to mechanisms of the disease such as insulin resistance or beta-cells dysfunction (Kim et al. 2023; Udler et al. 2018). This approach uses a selected set of curated genome-wide significant loci associated with T2D and information about diabetes related traits such as glycaemic or insulin traits. Then using soft clustering methods such as Bayesian non-negative matrix factorisation (bNMF) or k-means, we can aggregate clusters of loci implicated in specific mechanisms (obesity related pathways, insulin production or secretion pathways, etc…). This adds another layer of information about subtypes of T2D in addition to clusters based on clinical data such as the one used in this thesis, clusters based on genetic information could help refine the diagnosis of patients with T2D or even optimize medication approaches.

The studies we used in this thesis were mainly or exclusively of European ancestry. But it is known that T2D presents very differently in other populations, at a younger age and with a lower BMI in South Asian populations for example (Siddiqui et al. 2022). The focus on European populations limits our ability to understand the subtypes of the disease which present at younger ages and at lower BMI, more common on non-European populations. However, there is a concerted effort at the moment to expand studies into other populations, both genetic studies of T2D (Mahajan et al. 2022; Suzuki et al. 2023), and molecular studies as well (Vujkovic et al. 2020).

Another factor in all the studies we used was that they were all based on bulk tissue samples. We saw some evidence that this dilutes our ability to discover effects that are specific to particular parts of the organ or particular cell types. For example, an association with *TCF7L2* was found using pancreatic islets but missed using the whole pancreas data from GTEx. T2D involves many cell types, including immune related cells for the inflammation and stress response, beta cells in the production of insulin and myocytes and adipocytes involved in the insulin sensitivity via GLUT4 receptors. Studying these cell types using single cell technologies could bring another layer of information about the different causes and consequences of T2D.

Finally, most of the analyses presented in this thesis have looked for main effects, and ignored interactions. But one topic where interactions can be crucial is in the field of polysubstance studies of medication. When studying the effect of metformin, we used a single drug model, only focusing on metformin. In reality, patients with T2D are often on combinations of drugs, which treat not only T2D but also other comorbidities such as hypertension and cardiovascular disease. It would be interesting to know how these medications interact, and what the consequences are on molecular traits, but this may need studies far larger than those currently available.

References

Aguet, François, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, et al. 2017. ‘Genetic Effects on Gene Expression across Human Tissues’. *Nature* 550 (7675): 204–13. https://doi.org/10.1038/nature24277.

Ahlqvist, Emma, Rashmi B. Prasad, and Leif Groop. 2020. ‘Subtypes of Type 2 Diabetes Determined From Clinical Parameters’. *Diabetes* 69 (10): 2086–93. https://doi.org/10.2337/dbi20-0001.

Ahlqvist, Emma, Petter Storm, Annemari Käräjämäki, Mats Martinell, Mozhgan Dorkhan, Annelie Carlsson, Petter Vikman, et al. 2018. ‘Novel Subgroups of Adult-Onset Diabetes and Their Association with Outcomes: A Data-Driven Cluster Analysis of Six Variables’. *The Lancet Diabetes & Endocrinology* 6 (5): 361–69. https://doi.org/10.1016/S2213-8587(18)30051-2.

Ala, Moein. 2022. ‘The Emerging Role of Metformin in the Prevention and Treatment of Colorectal Cancer: A Game Changer for the Management of Colorectal Cancer’. *Current Diabetes Reviews* 18 (8): e051121197762. https://doi.org/10.2174/1573399818666211105125129.

Ali, Omar. 2013. ‘Genetics of Type 2 Diabetes’. *World Journal of Diabetes* 4 (4): 114–23. https://doi.org/10.4239/wjd.v4.i4.114.

Alonso, Lorena, Anthony Piron, Ignasi Morán, Marta Guindo-Martínez, Sílvia Bonàs-Guarch, Goutham Atla, Irene Miguel-Escalada, et al. 2021. ‘TIGER: The Gene Expression Regulatory Variation Landscape of Human Pancreatic Islets’. *Cell Reports* 37 (2): 109807. https://doi.org/10.1016/j.celrep.2021.109807.

American Diabetes Association. 2019. ‘9. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes—2020’. *Diabetes Care* 43 (Supplement_1): S98–110. https://doi.org/10.2337/dc20-S009.

Anwar, M. Akhtar, Wassim Abou Kheir, Stephanie Eid, Joanna Fares, Xiaoqi Liu, Ali H. Eid, and Assaad A. Eid. 2014. ‘Colorectal and Prostate Cancer Risk in Diabetes: Metformin, an Actor behind the Scene’. *Journal of Cancer* 5 (9): 736–44. https://doi.org/10.7150/jca.9726.

Assimon, Magdalene M. 2021. ‘Confounding in Observational Studies Evaluating the Safety and Effectiveness of Medical Treatments’. *Kidney360* 2 (7): 1156–59. https://doi.org/10.34067/KID.0007022020.

Bailey, C. J., C. Wilcock, and J. H. B. Scarpello. 2008. ‘Metformin and the Intestine’. *Diabetologia* 51 (8): 1552–53. https://doi.org/10.1007/s00125-008-1053-5.

Baldi, Pierre, and G. Wesley Hatfield. 2002. *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511541773.

Barbeira, Alvaro, Kaanan P. Shah, Jason M. Torres, Heather E. Wheeler, Eric S. Torstenson, Todd Edwards, Tzintzuni Garcia, et al. 2016. ‘MetaXcan: Summary Statistics Based Gene-Level Association Method Infers Accurate PrediXcan Results’. *bioRxiv*, March, 045260. https://doi.org/10.1101/045260.

Barroso, Inês, Jian’an Luan, Rita P. S Middelberg, Anne-Helen Harding, Paul W Franks, Rupert W Jakes, David Clayton, Alan J Schafer, Stephen O’Rahilly, and Nicholas J Wareham. 2003. ‘Candidate Gene Association Study in Type 2 Diabetes Indicates a

Role for Genes Involved in β-Cell Function as Well as Insulin Action'. *PLoS Biology* 1 (1): e20. https://doi.org/10.1371/journal.pbio.0000020.

Behrooz, Maryam, Elnaz Vaghef-Mehrabany, Vahid Maleki, Samira Pourmoradian, Zahra Fathifar, and Alireza Ostadrahimi. 2020. 'Spexin Status in Relation to Obesity and Its Related Comorbidities: A Systematic Review'. *Journal of Diabetes and Metabolic Disorders* 19 (2): 1943–57. https://doi.org/10.1007/s40200-020-00636-8.

Birney, Ewan, John A. Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R. Gingeras, Elliott H. Margulies, Zhiping Weng, et al. 2007. 'Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project'. *Nature* 447 (7146): 799–816. https://doi.org/10.1038/nature05874.

Bommer, Christian, Vera Sagalova, Esther Heesemann, Jennifer Manne-Goehler, Rifat Atun, Till Bärnighausen, Justine Davies, and Sebastian Vollmer. 2018. 'Global Economic Burden of Diabetes in Adults: Projections From 2015 to 2030'. *Diabetes Care* 41 (5): 963–70. https://doi.org/10.2337/dc17-1962.

Bosque-Plata, Laura del, Eduardo Martínez-Martínez, Miguel Ángel Espinoza-Camacho, and Claudia Gragnoli. 2021. 'The Role of TCF7L2 in Type 2 Diabetes'. *Diabetes* 70 (6): 1220–28. https://doi.org/10.2337/db20-0573.

Bowden, Jack, George Davey Smith, and Stephen Burgess. 2015. 'Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression'. *International Journal of Epidemiology* 44 (2): 512–25. https://doi.org/10.1093/ije/dyv080.

Brown, Juan J. Fernandez-Tajes, Mun-gwan Hong, Caroline A. Brorsson, Robert W. Koivula, David Davtian, Théo Dupuis, et al. 2023. 'Genetic Analysis of Blood Molecular Phenotypes Reveals Common Properties in the Regulatory Networks Affecting Complex Traits'. *Nature Communications* 14 (1): 5062. https://doi.org/10.1038/s41467-023-40569-3.

Brown, Ana Viñuela, Olivier Delaneau, Tim D. Spector, Kerrin S. Small, and Emmanouil T. Dermitzakis. 2017. 'Predicting Causal Variants Affecting Expression by Using Whole-Genome Sequencing and RNA-Seq from Multiple Human Tissues'. *Nature Genetics* 49 (12): 1747–51. https://doi.org/10.1038/ng.3979.

Burroughs, Valentine J., Randall W. Maxey, and Richard A. Levy. 2002. 'Racial and Ethnic Differences in Response to Medicines: Towards Individualized Pharmaceutical Treatment.' *Journal of the National Medical Association* 94 (10 Suppl): 1–26.

Bush, William S., and Jason H. Moore. 2012. 'Chapter 11: Genome-Wide Association Studies'. *PLoS Computational Biology* 8 (12). https://doi.org/10.1371/journal.pcbi.1002822.

Cai, Dongsheng, Minsheng Yuan, Daniel F Frantz, Peter A Melendez, Lone Hansen, Jongsoon Lee, and Steven E Shoelson. 2005. 'Local and Systemic Insulin Resistance Resulting from Hepatic Activation of IKK-β and NF-κB'. *Nature Medicine* 11 (2): 183–90. https://doi.org/10.1038/nm1166.

Cai, Lina, Eleanor Wheeler, Nicola D. Kerrison, Jian'an Luan, Panos Deloukas, Paul W. Franks, Pilar Amiano, et al. 2020. 'Genome-Wide Association Analysis of Type 2 Diabetes in the EPIC-InterAct Study'. *Scientific Data* 7 (November): 393. https://doi.org/10.1038/s41597-020-00716-7.

Cerf, Marlon E. 2013. 'Beta Cell Dysfunction and Insulin Resistance'. *Frontiers in Endocrinology* 4 (March): 37. https://doi.org/10.3389/fendo.2013.00037.

Chen, Sheng, Donghao Gan, Sixiong Lin, Yiming Zhong, Mingjue Chen, Xuenong Zou, Zengwu Shao, and Guozhi Xiao. 2022. 'Metformin in Aging and Aging-Related

Diseases: Clinical Applications and Relevant Mechanisms’. *Theranostics* 12 (6): 2722–40. https://doi.org/10.7150/thno.71360.

Chen, Zsu-Zsu, and Robert E. Gerszten. 2020. ‘Metabolomics and Proteomics in Type 2 Diabetes’. *Circulation Research* 126 (11): 1613–27. https://doi.org/10.1161/CIRCRESAHA.120.315898.

Christensen, Ashley A., and Maureen Gannon. 2019. ‘The Beta Cell in Type 2 Diabetes’. *Current Diabetes Reports* 19 (9): 81. https://doi.org/10.1007/s11892-019-1196-4.

Christiansen, Jan, Astrid M. Kolte, Thomas v O. Hansen, and Finn C. Nielsen. 2009. ‘IGF2 mRNA-Binding Protein 2: Biological Function and Putative Role in Type 2 Diabetes’. *Journal of Molecular Endocrinology* 43 (5): 187–95. https://doi.org/10.1677/JME-09-0016.

Clish, Clary B. 2015. ‘Metabolomics: An Emerging but Powerful Tool for Precision Medicine’. *Cold Spring Harbor Molecular Case Studies* 1 (1): a000588. https://doi.org/10.1101/mcs.a000588.

Collins, Francis S., and Leslie Fink. 1995. ‘The Human Genome Project’. *Alcohol Health and Research World* 19 (3): 190–95.

Crutchfield, Christopher A., Stefani N. Thomas, Lori J. Sokoll, and Daniel W. Chan. 2016. ‘Advances in Mass Spectrometry-Based Clinical Biomarker Discovery’. *Clinical Proteomics* 13 (1): 1. https://doi.org/10.1186/s12014-015-9102-9.

Cuomo, Anna S. E., Giordano Alvari, Christina B. Azodi, Davis J. McCarthy, and Marc Jan Bonder. 2021. ‘Optimizing Expression Quantitative Trait Locus Mapping Workflows for Single-Cell Studies’. *Genome Biology* 22 (June): 188. https://doi.org/10.1186/s13059-021-02407-x.

Czech, Michael P. 2017. ‘Insulin Action and Resistance in Obesity and Type 2 Diabetes’. *Nature Medicine* 23 (7): 804–14. https://doi.org/10.1038/nm.4350.

Da Silva Xavier, Gabriela. 2018. ‘The Cells of the Islets of Langerhans’. *Journal of Clinical Medicine* 7 (3): 54. https://doi.org/10.3390/jcm7030054.

Dai, Ning, Liping Zhao, Diedra Wrighting, Dana Krämer, Amit Majithia, Yanqun Wang, Valentin Cracan, et al. 2015. ‘IGF2BP2/IMP2-Deficient Mice Resist Obesity through Enhanced Translation of Ucp1 mRNA and Other mRNAs Encoding Mitochondrial Proteins’. *Cell Metabolism* 21 (4): 609–21. https://doi.org/10.1016/j.cmet.2015.03.006.

Dai, Xiaofeng, and Li Shen. 2022. ‘Advances and Trends in Omics Technology Development’. *Frontiers in Medicine* 9. https://www.frontiersin.org/articles/10.3389/fmed.2022.911861.

Dawed, Adem Y., Ashfaq Ali, Kaixin Zhou, Ewan R. Pearson, and Paul W. Franks. 2017. ‘Evidence-Based Prioritisation and Enrichment of Genes Interacting with Metformin in Type 2 Diabetes’. *Diabetologia* 60 (11): 2231–39. https://doi.org/10.1007/s00125-017-4404-2.

Debard, C., M. Laville, V. Berbe, E. Loizon, C. Guillet, B. Morio-Liondore, Y. Boirie, and H. Vidal. 2004. ‘Expression of Key Genes of Fatty Acid Oxidation, Including Adiponectin Receptors, in Skeletal Muscle of Type 2 Diabetic Patients’. *Diabetologia* 47 (5): 917–25. https://doi.org/10.1007/s00125-004-1394-7.

DeFronzo, R. A., J. D. Tobin, and R. Andres. 1979. ‘Glucose Clamp Technique: A Method for Quantifying Insulin Secretion and Resistance’. *The American Journal of Physiology* 237 (3): E214-223. https://doi.org/10.1152/ajpendo.1979.237.3.E214.

Delaneau, Olivier, Halit Ongen, Andrew A. Brown, Alexandre Fort, Nikolaos I. Panousis, and Emmanouil T. Dermitzakis. 2017. ‘A Complete Tool Set for Molecular QTL

Discovery and Analysis’. *Nature Communications* 8 (1): 15452. https://doi.org/10.1038/ncomms15452.

Dhillon, Rashpal S., Yiming (Amy) Qin, Paul R. van Ginkel, Vivian X. Fu, James M. Vann, Alexis J. Lawton, Cara L. Green, et al. 2022. ‘SIRT3 Deficiency Decreases Oxidative Metabolism Capacity but Increases Lifespan in Male Mice under Caloric Restriction’. *Aging Cell* 21 (12): e13721. https://doi.org/10.1111/acel.13721.

Dimas, Antigone S., Vasiliki Lagou, Adam Barker, Joshua W. Knowles, Reedik Mägi, Marie-France Hivert, Andrea Benazzo, et al. 2014. ‘Impact of Type 2 Diabetes Susceptibility Variants on Quantitative Glycemic Traits Reveals Mechanistic Heterogeneity’. *Diabetes* 63 (6): 2158–71. https://doi.org/10.2337/db13-0949.

Dudbridge, Frank, and Arief Gusnanto. 2008. ‘Estimation of Significance Thresholds for Genomewide Association Scans’. *Genetic Epidemiology* 32 (3): 227–34. https://doi.org/10.1002/gepi.20297.

Duggirala, R, J Blangero, L Almasy, T D Dyer, K L Williams, R J Leach, P O’Connell, and M P Stern. 1999. ‘Linkage of Type 2 Diabetes Mellitus and of Age at Onset to a Genetic Location on Chromosome 10q in Mexican Americans.’ *American Journal of Human Genetics* 64 (4): 1127–40.

Eder, Susanne, Johannes Leierer, Julia Kerschbaum, Laszlo Rosivall, Andrzej Wiecek, Dick de Zeeuw, Patrick B. Mark, et al. 2018. ‘A Prospective Cohort Study in Patients with Type 2 Diabetes Mellitus for Validation of Biomarkers (PROVALID) - Study Design and Baseline Characteristics’. *Kidney & Blood Pressure Research* 43 (1): 181–90. https://doi.org/10.1159/000487500.

Ekoru, Kenneth, Ayo Doumatey, Amy R. Bentley, Guanjie Chen, Jie Zhou, Daniel Shriner, Olufemi Fasanmade, et al. 2019. ‘Type 2 Diabetes Complications and Comorbidity in Sub-Saharan Africans’. *EClinicalMedicine* 16 (November): 30–41. https://doi.org/10.1016/j.eclinm.2019.09.001.

Feldman, Eva L., Brian C. Callaghan, Rodica Pop-Busui, Douglas W. Zochodne, Douglas E. Wright, David L. Bennett, Vera Bril, James W. Russell, and Vijay Viswanathan. 2019. ‘Diabetic Neuropathy’. *Nature Reviews Disease Primers* 5 (1): 1–18. https://doi.org/10.1038/s41572-019-0092-1.

Fernandez-Zapico, Martin E., Jennifer C. van Velkinburgh, Ruth Gutiérrez-Aguilar, Bernadette Neve, Philippe Froguel, Raul Urrutia, and Roland Stein. 2009. ‘MODY7 Gene, KLF11, Is a Novel P300-Dependent Regulator of Pdx-1 (MODY4) Transcription in Pancreatic Islet β Cells *’. *Journal of Biological Chemistry* 284 (52): 36482–90. https://doi.org/10.1074/jbc.M109.028852.

Finucane, Hilary K., Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, et al. 2018. ‘Heritability Enrichment of Specifically Expressed Genes Identifies Disease-Relevant Tissues and Cell Types’. *Nature Genetics* 50 (4): 621–29. https://doi.org/10.1038/s41588-018-0081-4.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. ‘GENCODE Reference Annotation for the Human and Mouse Genomes’. *Nucleic Acids Research* 47 (D1): D766–73. https://doi.org/10.1093/nar/gky955.

Galicia-Garcia, Unai, Asier Benito-Vicente, Shifa Jebari, Asier Larrea-Sebal, Haziq Siddiqi, Kepa B. Uribe, Helena Ostolaza, and César Martín. 2020. ‘Pathophysiology of Type 2 Diabetes Mellitus’. *International Journal of Molecular Sciences* 21 (17): 6275. https://doi.org/10.3390/ijms21176275.

Gamazon, Eric R., Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, et al. 2015. ‘A Gene-Based

Association Method for Mapping Traits Using Reference Transcriptome Data'. *Nature Genetics* 47 (9): 1091–98. https://doi.org/10.1038/ng.3367.

García-Pérez, Raquel, Jose Miguel Ramirez, Aida Ripoll-Cladellas, Ruben Chazarra-Gil, Winona Oliveros, Oleksandra Soldatkina, Mattia Bosio, et al. 2022. 'The Landscape of Expression and Alternative Splicing Variation across Human Traits'. *Cell Genomics* 3 (1): 100244. https://doi.org/10.1016/j.xgen.2022.100244.

Gasbjerg, Lærke Smidt, Maria Buur Nordskov Gabe, Bolette Hartmann, Mikkel Bring Christensen, Filip Krag Knop, Jens Juul Holst, and Mette Marie Rosenkilde. 2018. 'Glucose-Dependent Insulinotropic Polypeptide (GIP) Receptor Antagonists as Anti-Diabetic Agents'. *Peptides* 100 (February): 173–81. https://doi.org/10.1016/j.peptides.2017.11.021.

Gayon, Jean. 2016. 'From Mendel to Epigenetics: History of Genetics'. *Comptes Rendus Biologies*, Trajectories of genetics, 150 years after Mendel / Trajectoire de la génétique, 150 après Mendel Guest Editors / Rédacteurs en chef invités : Bernard Dujon, Georges Pelletier, 339 (7): 225–30. https://doi.org/10.1016/j.crvi.2016.05.009.

Giusti, Laura, Marta Tesi, Federica Ciregia, Lorella Marselli, Lorenzo Zallocco, Mara Suleiman, Carmela De Luca, et al. 2022. 'The Protective Action of Metformin against Pro-Inflammatory Cytokine-Induced Human Islet Cell Damage and the Mechanisms Involved'. *Cells* 11 (15): 2465. https://doi.org/10.3390/cells11152465.

Gou, Wanglong, Liang Yue, Xin-Yi Tang, Yan-Yan Wu, Xue Cai, Menglei Shuai, Zelei Miao, et al. 2022. 'Circulating Proteome and Progression of Type 2 Diabetes'. *The Journal of Clinical Endocrinology and Metabolism* 107 (6): 1616–25. https://doi.org/10.1210/clinem/dgac098.

Grant, Struan F. A., Gudmar Thorleifsson, Inga Reynisdottir, Rafn Benediktsson, Andrei Manolescu, Jesus Sainz, Agnar Helgason, et al. 2006. 'Variant of Transcription Factor 7-like 2 (TCF7L2) Gene Confers Risk of Type 2 Diabetes'. *Nature Genetics* 38 (3): 320–23. https://doi.org/10.1038/ng1732.

Grundberg, Elin, Kerrin S. Small, Åsa K. Hedman, Alexandra C. Nica, Alfonso Buil, Sarah Keildson, Jordana T. Bell, et al. 2012. 'Mapping Cis- and Trans-Regulatory Effects across Multiple Tissues in Twins'. *Nature Genetics* 44 (10): 1084–89. https://doi.org/10.1038/ng.2394.

GTEx Consortium. 2015. 'Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans'. *Science (New York, N.Y.)* 348 (6235): 648–60. https://doi.org/10.1126/science.1262110.

Gu, Harvest F., Jun Ma, Karolin T. Gu, and Kerstin Brismar. 2013. 'Association of Intercellular Adhesion Molecule 1 (ICAM1) with Diabetes and Diabetic Nephropathy'. *Frontiers in Endocrinology* 3 (January): 179. https://doi.org/10.3389/fendo.2012.00179.

Guerra-García, M. T., H. Moreno-Macías, A. Ochoa-Guzmán, M. L. Ordoñez-Sánchez, R. Rodríguez-Guillen, P. Vázquez-Cárdenas, V. M. Ortíz-Ortega, M. Peimbert-Torres, C. A. Aguilar-Salinas, and M. T. Tusié-Luna. 2021. 'The -514C>T Polymorphism in the LIPC Gene Modifies Type 2 Diabetes Risk through Modulation of HDL-Cholesterol Levels in Mexicans'. *Journal of Endocrinological Investigation* 44 (3): 557–65. https://doi.org/10.1007/s40618-020-01346-x.

Halban, Philippe A., Kenneth S. Polonsky, Donald W. Bowden, Meredith A. Hawkins, Charlotte Ling, Kieren J. Mather, Alvin C. Powers, Christopher J. Rhodes, Lori Sussel, and Gordon C. Weir. 2014. 'β-Cell Failure in Type 2 Diabetes: Postulated

Mechanisms and Prospects for Prevention and Treatment'. *Diabetes Care* 37 (6): 1751–58. https://doi.org/10.2337/dc14-0396.

Herdenberg, Carl, Pascal M. Mutie, Ola Billing, Ahmad Abdullah, Rona J. Strawbridge, Ingrid Dahlman, Simon Tuck, et al. 2021. 'LRIG Proteins Regulate Lipid Metabolism via BMP Signaling and Affect the Risk of Type 2 Diabetes'. *Communications Biology* 4 (1): 1–15. https://doi.org/10.1038/s42003-020-01613-w.

Howard, B. A., and B. A. Gusterson. 2000. 'Human Breast Development'. *Journal of Mammary Gland Biology and Neoplasia* 5 (2): 119–37. https://doi.org/10.1023/a:1026487120779.

Ilias, Ioannis, Manfredi Rizzo, and Lina Zabuliene. 2022. 'Metformin: Sex/Gender Differences in Its Uses and Effects—Narrative Review'. *Medicina* 58 (3): 430. https://doi.org/10.3390/medicina58030430.

Inaishi, Jun, and Yoshifumi Saisho. 2020. 'Beta-Cell Mass in Obesity and Type 2 Diabetes, and Its Relation to Pancreas Fat: A Mini-Review'. *Nutrients* 12 (12): 3846. https://doi.org/10.3390/nu12123846.

Ingalhalikar, Madhura, Alex Smith, Drew Parker, Theodore D. Satterthwaite, Mark A. Elliott, Kosha Ruparel, Hakon Hakonarson, Raquel E. Gur, Ruben C. Gur, and Ragini Verma. 2014. 'Sex Differences in the Structural Connectome of the Human Brain'. *Proceedings of the National Academy of Sciences* 111 (2): 823–28. https://doi.org/10.1073/pnas.1316909110.

Inzucchi, Silvio E., Richard M. Bergenstal, John B. Buse, Michaela Diamant, Ele Ferrannini, Michael Nauck, Anne L. Peters, Apostolos Tsapas, Richard Wender, and David R. Matthews. 2012. 'Management of Hyperglycemia in Type 2 Diabetes: A Patient-Centered Approach'. *Diabetes Care* 35 (6): 1364–79. https://doi.org/10.2337/dc12-0413.

Jannot, Anne-Sophie, Georg Ehret, and Thomas Perneger. 2015. 'P < 5 × 10−8 Has Emerged as a Standard of Statistical Significance for Genome-Wide Association Studies'. *Journal of Clinical Epidemiology* 68 (4): 460–65. https://doi.org/10.1016/j.jclinepi.2015.01.001.

Jansen, Jacobus F. A., Frank C. G. van Bussel, Harm J. van de Haar, Matthias J. P. van Osch, Paul A. M. Hofman, Martin P. J. van Boxtel, Robert J. van Oostenbrugge, et al. 2016. 'Cerebral Blood Flow, Blood Supply, and Cognition in Type 2 Diabetes Mellitus'. *Scientific Reports* 6 (December): 160003. https://doi.org/10.1038/s41598-016-0003-6.

Kang, Hyun Min, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, et al. 2018. 'Multiplexed Droplet Single-Cell RNA-Sequencing Using Natural Genetic Variation'. *Nature Biotechnology* 36 (1): 89–94. https://doi.org/10.1038/nbt.4042.

Kasela, Silva, François Aguet, Sarah Kim-Hellmuth, Brielin C. Brown, Daniel C. Nachun, Russell P. Tracy, Peter Durda, et al. 2023. 'Interaction Molecular QTL Mapping Discovers Cellular and Environmental Modifiers of Genetic Regulatory Effects'. bioRxiv. https://doi.org/10.1101/2023.06.26.546528.

Khan, Moien Abdul Basith, Muhammad Jawad Hashim, Jeffrey Kwan King, Romona Devi Govender, Halla Mustafa, and Juma Al Kaabi. 2020. 'Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends'. *Journal of Epidemiology and Global Health* 10 (1): 107–11. https://doi.org/10.2991/jegh.k.191028.001.

Khetan, Shubham, Romy Kursawe, Ahrim Youn, Nathan Lawlor, Alexandria Jillette, Eladio J. Marquez, Duygu Ucar, and Michael L. Stitzel. 2018. 'Type 2 Diabetes–

Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets’. *Diabetes* 67 (11): 2466. https://doi.org/10.2337/db18-0393.

Kim, Hyunkyung, Kenneth E. Westerman, Kirk Smith, Joshua Chiou, Joanne B. Cole, Timothy Majarian, Marcin von Grotthuss, et al. 2023. ‘High-Throughput Genetic Clustering of Type 2 Diabetes Loci Reveals Heterogeneous Mechanistic Pathways of Metabolic Disease’. *Diabetologia* 66 (3): 495–507. https://doi.org/10.1007/s00125-022-05848-6.

Klein, Niek de, Ellen A. Tsai, Martijn Vochteloo, Denis Baird, Yunfeng Huang, Chia-Yen Chen, Sipko van Dam, et al. 2023. ‘Brain Expression Quantitative Trait Locus and Network Analyses Reveal Downstream Effects and Putative Drivers for Brain-Related Diseases’. *Nature Genetics*, February, 1–12. https://doi.org/10.1038/s41588-023-01300-6.

Klooster, Jean Paul ten, Alexandros Sotiriou, Sjef Boeren, Stefan Vaessen, Jacques Vervoort, and Raymond Pieters. 2018. ‘Type 2 Diabetes-Related Proteins Derived from an in Vitro Model of Inflamed Fat Tissue’. *Archives of Biochemistry and Biophysics* 644 (April): 81–92. https://doi.org/10.1016/j.abb.2018.03.003.

Koivula, Robert W., Alison Heggie, Anna Barnett, Henna Cederberg, Tue H. Hansen, Anitra D. Koopman, Martin Ridderstråle, et al. 2014. ‘Discovery of Biomarkers for Glycaemic Deterioration before and after the Onset of Type 2 Diabetes: Rationale and Design of the Epidemiological Studies within the IMI DIRECT Consortium’. *Diabetologia* 57 (6): 1132–42. https://doi.org/10.1007/s00125-014-3216-x.

Kwon, Jennifer M. 2000. ‘The Candidate Gene Approach’ 24 (3).

Lackey, Denise E., and Jerrold M. Olefsky. 2016. ‘Regulation of Metabolism by the Innate Immune System’. *Nature Reviews Endocrinology* 12 (1): 15–28. https://doi.org/10.1038/nrendo.2015.189.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. ‘Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts’. *Genome Biology* 15 (2): R29. https://doi.org/10.1186/gb-2014-15-2-r29.

Leeuw, Christiaan A. de, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. 2015. ‘MAGMA: Generalized Gene-Set Analysis of GWAS Data’. *PLOS Computational Biology* 11 (4): e1004219. https://doi.org/10.1371/journal.pcbi.1004219.

Leighton, Emma, Christopher AR Sainsbury, and Gregory C. Jones. 2017. ‘A Practical Review of C-Peptide Testing in Diabetes’. *Diabetes Therapy* 8 (3): 475–87. https://doi.org/10.1007/s13300-017-0265-4.

Li, Hongyan, and Zhiyou Cai. 2023. ‘SIRT3 Regulates Mitochondrial Biogenesis in Aging-Related Diseases’. *Journal of Biomedical Research* 37 (2): 77–88. https://doi.org/10.7555/JBR.36.20220078.

Li, Jingyi Jessica, Peter J. Bickel, and Mark D. Biggin. 2014. ‘System Wide Analyses Have Underestimated Protein Abundances and the Importance of Transcription in Mammals’. *PeerJ* 2 (February): e270. https://doi.org/10.7717/peerj.270.

Li, Yumei, Xinzhou Ge, Fanglue Peng, Wei Li, and Jingyi Jessica Li. 2022. ‘Exaggerated False Positives by Popular Differential Expression Methods When Analyzing Human Population Samples’. *Genome Biology* 23 (1): 79. https://doi.org/10.1186/s13059-022-02648-4.

Liadis, Nicole, Kiichi Murakami, Mohamed Eweida, Alisha R. Elford, Laura Sheu, Herbert Y. Gaisano, Razqallah Hakem, Pamela S. Ohashi, and Minna Woo. 2005. ‘Caspase-3-Dependent β-Cell Apoptosis in the Initiation of Autoimmune Diabetes Mellitus’. *Molecular and Cellular Biology* 25 (9): 3620–29. https://doi.org/10.1128/MCB.25.9.3620-3629.2005.

Lin, Xiling, Yufeng Xu, Xiaowen Pan, Jingya Xu, Yue Ding, Xue Sun, Xiaoxiao Song, Yuezhong Ren, and Peng-Fei Shan. 2020. ‘Global, Regional, and National Burden and Trend of Diabetes in 195 Countries and Territories: An Analysis from 1990 to 2025’. *Scientific Reports* 10 (1): 14790. https://doi.org/10.1038/s41598-020-71908-9.

Liu, Ching-Ti, Jordi Merino, Denis Rybin, Daniel DiCorpo, Kelly S. Benke, Jennifer L. Bragg-Gresham, Mickaël Canouil, et al. 2019. ‘Genome-Wide Association Study of Change in Fasting Glucose over Time in 13,807 Non-Diabetic European Ancestry Individuals’. *Scientific Reports* 9 (1): 9439. https://doi.org/10.1038/s41598-019-45823-7.

Liu, Shiyi, Zitao Wang, Ronghui Zhu, Feiyan Wang, Yanxiang Cheng, and Yeqiang Liu. 2021. ‘Three Differential Expression Analysis Methods for RNA Sequencing: Limma, EdgeR, DESeq2’. *Journal of Visualized Experiments: JoVE*, no. 175 (September). https://doi.org/10.3791/62528.

Liu, Yaozhong, Biao Li, Yingxu Ma, Yunying Huang, Feifan Ouyang, and Qiming Liu. 2021. ‘Mendelian Randomization Integrating GWAS, eQTL, and mQTL Data Identified Genes Pleiotropically Associated With Atrial Fibrillation’. *Frontiers in Cardiovascular Medicine* 8. https://www.frontiersin.org/article/10.3389/fcvm.2021.745757.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. ‘The Genotype-Tissue Expression (GTEx) Project’. *Nature Genetics* 45 (6): 580–85. https://doi.org/10.1038/ng.2653.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. ‘Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2’. *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Loza-Valdes, Angel, Rabih El-Merahbi, Toufic Kassouf, Agnieszka Demczuk, Saskia Reuter, Jonathan Trujillo Viera, Till Karwen, et al. 2022. ‘Targeting ERK3/MK5 Complex for Treatment of Obesity and Diabetes’. *Biochemical and Biophysical Research Communications* 612 (July): 119–25. https://doi.org/10.1016/j.bbrc.2022.04.070.

Madiraju, Anila K., Derek M. Erion, Yasmeen Rahimi, Xian-Man Zhang, Demetrios Braddock, Ronald A. Albright, Brett J. Prigaro, et al. 2014. ‘Metformin Suppresses Gluconeogenesis by Inhibiting Mitochondrial Glycerophosphate Dehydrogenase’. *Nature* 510 (7506): 542–46. https://doi.org/10.1038/nature13270.

Mahajan, Anubha, Cassandra N. Spracklen, Weihua Zhang, Maggie C. Y. Ng, Lauren E. Petty, Hidetoshi Kitajima, Grace Z. Yu, et al. 2022. ‘Multi-Ancestry Genetic Study of Type 2 Diabetes Highlights the Power of Diverse Populations for Discovery and Translation’. *Nature Genetics* 54 (5): 560–72. https://doi.org/10.1038/s41588-022-01058-3.

Mahajan, Anubha, Daniel Taliun, Matthias Thurner, Neil R. Robertson, Jason M. Torres, N. William Rayner, Anthony J. Payne, et al. 2018. ‘Fine-Mapping Type 2 Diabetes Loci to Single-Variant Resolution Using High-Density Imputation and Islet-Specific Epigenome Maps’. *Nature Genetics* 50 (11): 1505–13. https://doi.org/10.1038/s41588-018-0241-6.

Mancuso, Nicholas, Simon Gayther, Alexander Gusev, Wei Zheng, Kathryn L. Penney, Zsofia Kote-Jarai, Rosalind Eeles, Matthew Freedman, Christopher Haiman, and Bogdan Pasaniuc. 2018. ‘Large-Scale Transcriptome-Wide Association Study Identifies New Prostate Cancer Risk Regions’. *Nature Communications* 9 (1): 4079. https://doi.org/10.1038/s41467-018-06302-1.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. ‘Finding the Missing Heritability of Complex Diseases’. *Nature* 461 (7265): 747–53. https://doi.org/10.1038/nature08494.

Mansour Aly, Dina, Om Prakash Dwivedi, Rashmi B. Prasad, Annemari Käräjämäki, Rebecka Hjort, Manonanthini Thangam, Mikael Åkerlund, et al. 2021. ‘Genome-Wide Association Analyses Highlight Etiological Differences Underlying Newly Defined Subtypes of Diabetes’. *Nature Genetics* 53 (11): 1534–42. https://doi.org/10.1038/s41588-021-00948-2.

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. ‘Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation’. *Nucleic Acids Research* 40 (10): 4288–97. https://doi.org/10.1093/nar/gks042.

Mendel, Gregor. n.d. ‘EXPERIMENTS IN PLANT HYBRIDIZATION (1865)’.

Miguel-Escalada, Irene, Silvia Bonàs-Guarch, Inês Cebola, Joan Ponsa-Cobas, Julen Mendieta-Esteban, Goutham Atla, Biola M. Javierre, et al. 2019. ‘Human Pancreatic Islet Three-Dimensional Chromatin Architecture Provides Insights into the Genetics of Type 2 Diabetes’. *Nature Genetics* 51 (7): 1137–48. https://doi.org/10.1038/s41588-019-0457-0.

Moon, Jun Sung, Udayakumar Karunakaran, Elumalai Suma, Seung Min Chung, and Kyu Chang Won. 2020. ‘The Role of CD36 in Type 2 Diabetes Mellitus: β-Cell Dysfunction and Beyond’. *Diabetes & Metabolism Journal* 44 (2): 222–33. https://doi.org/10.4093/dmj.2020.0053.

Mootha, Vamsi K., Cecilia M. Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, et al. 2003. ‘PGC-1α-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes’. *Nature Genetics* 34 (3): 267–73. https://doi.org/10.1038/ng1180.

Mountjoy, Edward, Ellen M. Schmidt, Miguel Carmona, Jeremy Schwartzentruber, Gareth Peat, Alfredo Miranda, Luca Fumis, et al. 2021. ‘An Open Approach to Systematically Prioritize Causal Variants and Genes at All Published Human GWAS Trait-Associated Loci’. *Nature Genetics* 53 (11): 1527–33. https://doi.org/10.1038/s41588-021-00945-5.

Namba, Shinichi, Takahiro Konuma, Kuan-Han Wu, Wei Zhou, and Yukinori Okada. 2022. ‘A Practical Guideline of Genomics-Driven Drug Discovery in the Era of Global Biobank Meta-Analysis’. *Cell Genomics* 2 (10): 100190. https://doi.org/10.1016/j.xgen.2022.100190.

Nauck, Michael A., Daniel R. Quast, Jakob Wefers, and Juris J. Meier. 2020. ‘GLP-1 Receptor Agonists in the Treatment of Type 2 Diabetes – State-of-the-Art’. *Molecular Metabolism* 46 (October): 101102. https://doi.org/10.1016/j.molmet.2020.101102.

Ndungu, Anne, Anthony Payne, Jason M. Torres, Martijn van de Bunt, and Mark I. McCarthy. 2020. ‘A Multi-Tissue Transcriptome Analysis of Human Metabolites Guides Interpretability of Associations Based on Multi-SNP Models for Gene Expression’. *The American Journal of Human Genetics* 106 (2): 188–201. https://doi.org/10.1016/j.ajhg.2020.01.003.

Palleria, Caterina, Antonello Di Paolo, Chiara Giofrè, Chiara Caglioti, Giacomo Leuzzi, Antonio Siniscalchi, Giovambattista De Sarro, and Luca Gallelli. 2013. ‘Pharmacokinetic Drug-Drug Interaction and Their Implication in Clinical

Management'. *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences* 18 (7): 601–10.

Pearson, Taliesin, Jonathan A. D. Wattis, John R. King, Ian A. MacDonald, and Dawn J. Mazzatti. 2016. 'The Effects of Insulin Resistance on Individual Tissues: An Application of a Mathematical Model of Metabolism in Humans'. *Bulletin of Mathematical Biology* 78: 1189–1217. https://doi.org/10.1007/s11538-016-0181-1.

Piñero, Janet, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. 2020. 'The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update'. *Nucleic Acids Research* 48 (D1): D845–55. https://doi.org/10.1093/nar/gkz1021.

Porcu, Eleonora, Sina Rüeger, Kaido Lepik, Federico A. Santoni, Alexandre Reymond, and Zoltán Kutalik. 2019. 'Mendelian Randomization Integrating GWAS and eQTL Data Reveals Genetic Determinants of Complex and Clinical Traits'. *Nature Communications* 10 (1): 3300. https://doi.org/10.1038/s41467-019-10936-0.

Porcu, Eleonora, Marie C. Sadler, Kaido Lepik, Chiara Auwerx, Andrew R. Wood, Antoine Weihs, Maroun S. Bou Sleiman, et al. 2021. 'Differentially Expressed Genes Reflect Disease-Induced Rather than Disease-Causing Changes in the Transcriptome'. *Nature Communications* 12 (1): 5647. https://doi.org/10.1038/s41467-021-25805-y.

Quinlan, Aaron R., and Ira M. Hall. 2010. 'BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features'. *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Rahman, Md Saidur, Khandkar Shaharina Hossain, Sharnali Das, Sushmita Kundu, Elikanah Olusayo Adegoke, Md. Ataur Rahman, Md. Abdul Hannan, Md Jamal Uddin, and Myung-Geol Pang. 2021. 'Role of Insulin in Health and Disease: An Update'. *International Journal of Molecular Sciences* 22 (12): 6403. https://doi.org/10.3390/ijms22126403.

Ren, Qingwen, Mengzhu Guo, Feifei Yang, Tianbi Han, Wenqiong Du, Feng Zhao, Jinbo Li, et al. 2021. 'Association of CPT1A Gene Polymorphism with the Risk of Gestational Diabetes Mellitus: A Case-Control Study'. *Journal of Assisted Reproduction and Genetics* 38 (7): 1861–69. https://doi.org/10.1007/s10815-021-02143-y.

Roden, Michael, and Gerald I. Shulman. 2019. 'The Integrative Biology of Type 2 Diabetes'. *Nature* 576 (7785): 51–60. https://doi.org/10.1038/s41586-019-1797-8.

Rodriguez-Fontenla, Cristina, and Angel Carracedo. 2021. 'UTMOST, a Single and Cross-Tissue TWAS (Transcriptome Wide Association Study), Reveals New ASD (Autism Spectrum Disorder) Associated Genes'. *Translational Psychiatry* 11 (1): 1–11. https://doi.org/10.1038/s41398-021-01378-8.

Säll, Johanna, Annie M. L. Pettersson, Christel Björk, Emma Henriksson, Sebastian Wasserstrom, Wilhelm Linder, Yuedan Zhou, et al. 2017. 'Salt-Inducible Kinase 2 and -3 Are Downregulated in Adipose Tissue from Obese or Insulin-Resistant Individuals: Implications for Insulin Signalling and Glucose Uptake in Human Adipocytes'. *Diabetologia* 60 (2): 314–23. https://doi.org/10.1007/s00125-016-4141-y.

Schuster, Stephan C. 2008. 'Next-Generation Sequencing Transforms Today's Biology'. *Nature Methods* 5 (1): 16–18. https://doi.org/10.1038/nmeth1156.

Siddiqui, Ranjit Mohan Anjana, Adem Y. Dawed, Cyrielle Martoeau, Sundararajan Srinivasan, Jebarani Saravanan, Sathish K. Madanagopal, et al. 2022. 'Young-Onset Diabetes in Asian Indians Is Associated with Lower Measured and Genetically

Determined Beta Cell Function’. *Diabetologia* 65 (6): 973–83. https://doi.org/10.1007/s00125-022-05671-z.

Siddiqui, Ajay Soni, and Sarfaraz Alam Khan. 2019. ‘Association of ABO Blood Types and Novel Obesity Markers in Healthy Adolescents’. *Journal of Education and Health Promotion* 8 (August): 153. https://doi.org/10.4103/jehp.jehp_462_18.

Sjöstedt, Evelina, Wen Zhong, Linn Fagerberg, Max Karlsson, Nicholas Mitsios, Csaba Adori, Per Oksvold, et al. 2020. ‘An Atlas of the Protein-Coding Genes in the Human, Pig, and Mouse Brain’. *Science (New York, N.Y.)* 367 (6482): eaay5947. https://doi.org/10.1126/science.aay5947.

Song, Yiqing, JoAnn E. Manson, Lesley Tinker, Barbara V. Howard, Lewis H. Kuller, Lauren Nathan, Nader Rifai, and Simin Liu. 2007. ‘Insulin Sensitivity and Insulin Secretion Determined by Homeostasis Model Assessment (HOMA) and Risk of Diabetes in a Multiethnic Cohort of Women: The Women’s Health Initiative Observational Study’. *Diabetes Care* 30 (7): 1747–52. https://doi.org/10.2337/dc07-0358.

Southern, E. M. 1975. ‘Detection of Specific Sequences among DNA Fragments Separated by Gel Electrophoresis’. *Journal of Molecular Biology* 98 (3): 503–17. https://doi.org/10.1016/S0022-2836(75)80083-0.

Spieth, Peter Markus, Anne Sophie Kubasch, Ana Isabel Penzlin, Ben Min-Woo Illigens, Kristian Barlinn, and Timo Siepmann. 2016. ‘Randomized Controlled Trials – a Matter of Design’. *Neuropsychiatric Disease and Treatment* 12 (June): 1341–49. https://doi.org/10.2147/NDT.S101938.

Spracklen, Cassandra N, Momoko Horikoshi, Young Jin Kim, Kuang Lin, Fiona Bragg, Sanghoon Moon, Ken Suzuki, et al. 2020. ‘Identification of Type 2 Diabetes Loci in 433,540 East Asian Individuals’. *Nature* 582 (7811): 240–45. https://doi.org/10.1038/s41586-020-2263-3.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. ‘Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles’. *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Suhre, Karsten, Mark I. McCarthy, and Jochen M. Schwenk. 2021. ‘Genetics Meets Proteomics: Perspectives for Large Population-Based Studies’. *Nature Reviews Genetics* 22 (1): 19–37. https://doi.org/10.1038/s41576-020-0268-2.

Sun, Dawei, Lewis Evans, Francesca Perrone, Vanesa Sokleva, Kyungtae Lim, Saba Rezakhani, Matthias Lutolf, Matthias Zilbauer, and Emma L. Rawlins. 2021. ‘A Functional Genetic Toolbox for Human Tissue-Derived Organoids’. *eLife* 10 (October): e67886. https://doi.org/10.7554/eLife.67886.

Suzuki, Ken, Konstantinos Hatzikotoulas, Lorraine Southam, Henry J. Taylor, Xianyong Yin, Kim M. Lorenz, Ravi Mandla, et al. 2023. ‘Multi-Ancestry Genome-Wide Study in >2.5 Million Individuals Reveals Heterogeneity in Mechanistic Pathways of Type 2 Diabetes and Complications’. *medRxiv*, March, 2023.03.31.23287839. https://doi.org/10.1101/2023.03.31.23287839.

Thanabalasingham, Gaya, and Katharine R. Owen. 2011. ‘Diagnosis and Management of Maturity Onset Diabetes of the Young (MODY)’. *BMJ (Clinical Research Ed.)* 343 (October): d6044. https://doi.org/10.1136/bmj.d6044.

The GTEx Consortium. 2020. ‘The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues’. *Science (New York, N.Y.)* 369 (6509): 1318–30. https://doi.org/10.1126/science.aaz1776.

‘The UCSC Genome Browser Database: Update 2006 - PubMed’. n.d. Accessed 8 October 2023. https://pubmed.ncbi.nlm.nih.gov/16381938/.

Tomita, Tatsuo. 2010. ‘Immunocytochemical Localisation of Caspase-3 in Pancreatic Islets from Type 2 Diabetic Subjects’. *Pathology* 42 (5): 432–37. https://doi.org/10.3109/00313025.2010.493863.

Udler, Miriam S., Jaegil Kim, Marcin von Grotthuss, Sílvia Bonàs-Guarch, Joanne B. Cole, Joshua Chiou, Christopher D. Anderson on behalf of METASTROKE and the Isgc, et al. 2018. ‘Type 2 Diabetes Genetic Loci Informed by Multi-Trait Associations Point to Disease Mechanisms and Subtypes: A Soft Clustering Analysis’. *PLOS Medicine* 15 (9): e1002654. https://doi.org/10.1371/journal.pmed.1002654.

Uebi, Tatsuya, Yumi Itoh, Osamu Hatano, Ayako Kumagai, Masato Sanosaka, Tsutomu Sasaki, Satoru Sasagawa, et al. 2012. ‘Involvement of SIK3 in Glucose and Lipid Homeostasis in Mice’. *PLoS ONE* 7 (5): e37803. https://doi.org/10.1371/journal.pone.0037803.

Viñuela, Ana, Andrew A Brown, Alfonso Buil, Pei-Chien Tsai, Matthew N Davies, Jordana T Bell, Emmanouil T Dermitzakis, Timothy D Spector, and Kerrin S Small. 2018. ‘Age-Dependent Changes in Mean and Variance of Gene Expression across Tissues in a Twin Cohort’. *Human Molecular Genetics* 27 (4): 732–41. https://doi.org/10.1093/hmg/ddx424.

Viñuela, Ana, Arushi Varshney, Martijn van de Bunt, Rashmi B. Prasad, Olof Asplund, Amanda Bennett, Michael Boehnke, et al. 2019. ‘Influence of Genetic Variants on Gene Expression in Human Pancreatic Islets – Implications for Type 2 Diabetes’. *bioRxiv*, January, 655670. https://doi.org/10.1101/655670.

———. 2020. ‘Genetic Variant Effects on Gene Expression in Human Pancreatic Islets and Their Implications for T2D’. *Nature Communications* 11 (1): 4912. https://doi.org/10.1038/s41467-020-18581-8.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. ‘Five Years of GWAS Discovery’. *The American Journal of Human Genetics* 90 (1): 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029.

Vujkovic, Marijana, Jacob M Keaton, Julie A Lynch, Donald R Miller, Jin Zhou, Catherine Tcheandjieu, Jennifer E Huffman, et al. 2019. ‘Discovery of 318 Novel Loci for Type-2 Diabetes and Related Micro- and Macrovascular Outcomes among 1.4 Million Participants in a Multi-Ethnic Meta-Analysis.’ Preprint. Genetic and Genomic Medicine. https://doi.org/10.1101/19012690.

Vujkovic, Marijana, Jacob M. Keaton, Julie A. Lynch, Donald R. Miller, Jin Zhou, Catherine Tcheandjieu, Jennifer E. Huffman, et al. 2020. ‘Discovery of 318 New Risk Loci for Type 2 Diabetes and Related Vascular Outcomes among 1.4 Million Participants in a Multi-Ancestry Meta-Analysis’. *Nature Genetics* 52 (7): 680–91. https://doi.org/10.1038/s41588-020-0637-y.

Wainberg, Michael, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N. Barbeira, David A. Knowles, David Golan, Raili Ermel, et al. 2019. ‘Opportunities and Challenges for Transcriptome-Wide Association Studies’. *Nature Genetics* 51 (4): 592–99. https://doi.org/10.1038/s41588-019-0385-z.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. ‘RNA-Seq: A Revolutionary Tool for Transcriptomics’. *Nature Reviews. Genetics* 10 (1): 57–63. https://doi.org/10.1038/nrg2484.

Watson, J. D., and F. H. Crick. 1953. ‘Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid’. *Nature* 171 (4356): 737–38. https://doi.org/10.1038/171737a0.

WHO, World Health. 1999. 'Definition, Diagnosis and Classification of Diabetes Mellitus and Its Complications : Report of a WHO Consultation. Part 1, Diagnosis and Classification of Diabetes Mellitus'. https://apps.who.int/iris/handle/10665/66040.

Wijst, Monique G. P. van der, Harm Brugge, Dylan H. de Vries, Patrick Deelen, Morris A. Swertz, and Lude Franke. 2018. 'Single-Cell RNA Sequencing Identifies Celltype-Specific Cis-eQTLs and Co-Expression QTLs'. *Nature Genetics* 50 (4): 493–97. https://doi.org/10.1038/s41588-018-0089-9.

Wilding, J P H. 2014. 'The Importance of Weight Management in Type 2 Diabetes Mellitus'. *International Journal of Clinical Practice* 68 (6): 682–91. https://doi.org/10.1111/ijcp.12384.

Xie, Xiaoman, Casey Hanson, and Saurabh Sinha. 2019. 'Mechanistic Interpretation of Non-Coding Variants for Discovering Transcriptional Regulators of Drug Response'. *BMC Biology* 17 (1): 62. https://doi.org/10.1186/s12915-019-0679-8.

Zaghlool, Shaza B., Anna Halama, Nisha Stephan, Valborg Gudmundsdottir, Vilmundur Gudnason, Lori L. Jennings, Manonanthini Thangam, et al. 2022. 'Metabolic and Proteomic Signatures of Type 2 Diabetes Subtypes in an Arab Population'. *Nature Communications* 13 (1): 7121. https://doi.org/10.1038/s41467-022-34754-z.

Zhang, Kui, Wenxing Yang, Hao Dai, and Zhenhua Deng. 2020. 'Cardiovascular Risk Following Metformin Treatment in Patients with Type 2 Diabetes Mellitus: Results from Meta-Analysis'. *Diabetes Research and Clinical Practice* 160 (February): 108001. https://doi.org/10.1016/j.diabres.2020.108001.

Zhong, Wen, Fredrik Edfors, Anders Gummesson, Göran Bergström, Linn Fagerberg, and Mathias Uhlén. 2021. 'Next Generation Plasma Proteome Profiling to Monitor Health and Disease'. *Nature Communications* 12 (1): 2493. https://doi.org/10.1038/s41467-021-22767-z.

Zhou, Celine Bellenguez, Chris CA Spencer, Amanda J Bennett, Ruth L Coleman, Roger Tavendale, Simon A. Hawley, et al. 2011. 'Common Variants near ATM Are Associated with Glycemic Response to Metformin in Type 2 Diabetes'. *Nature Genetics* 43 (2): 117–20. https://doi.org/10.1038/ng.735.

Zhou, Yuan Ming Di, Eli Chan, Yao-Min Du, Vivian Deh-Wei Chow, Charlie Changli Xue, Xinsheng Lai, et al. 2008. 'Clinical Pharmacogenetics and Potential Application in Personalized Medicine'. *Current Drug Metabolism* 9 (8): 738–84. https://doi.org/10.2174/138920008786049302.

Zhou, Robert Myers, Ying Li, Yuli Chen, Xiaolan Shen, Judy Fenyk-Melody, Margaret Wu, et al. 2001. 'Role of AMP-Activated Protein Kinase in Mechanism of Metformin Action'. *The Journal of Clinical Investigation* 108 (8): 1167–74. https://doi.org/10.1172/JCI13505.

Zhou, Helle Krogh Pedersen, Adem Y. Dawed, and Ewan R. Pearson. 2016. 'Pharmacogenomics in Diabetes Mellitus: Insights into Drug Action and Drug Discovery'. *Nature Reviews Endocrinology* 12 (6): 337–46. https://doi.org/10.1038/nrendo.2016.51.

Zhou, Sook Wah Yee, Eric L. Seiser, Nienke van Leeuwen, Roger Tavendale, Amanda J Bennett, Christopher J. Groves, et al. 2016. 'Variation in the Glucose Transporter Gene SLC2A2 Is Associated with Glycaemic Response to Metformin'. *Nature Genetics* 48 (9): 1055–59. https://doi.org/10.1038/ng.3632.

Zyla, Joanna, Michal Marczyk, January Weiner, and Joanna Polanska. 2017. 'Ranking Metrics in Gene Set Enrichment Analysis: Do They Matter?' *BMC Bioinformatics* 18 (May): 256. https://doi.org/10.1186/s12859-017-1674-0.