CRANFIELD UNIVERSITY

CORENTIN MOLITOR

A BIOINFORMATICS AND GENOTYPING APPROACH EXPLORING PERSONALISED NUTRITION

School of Water, Energy and Environment
PhD in Environment and Agrifood

PhD
Academic Year: 2018 - 2021

Supervisor: Dr Fady Mohareb
Associate Supervisor: Prof Andrew Thompson
November, 2021

CRANFIELD UNIVERSITY

School of Water, Energy and Environment
PhD in Environment and Agrifood

PhD

Academic Year 2018 – 2021

CORENTIN MOLITOR

A BIOINFORMATICS AND GENOTYPING APPROACH EXPLORING PERSONALISED NUTRITION

Supervisor: Dr Fady Mohareb
Associate Supervisor: Prof Andrew Thompson
November, 2021

© Cranfield University 2021. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright owner.

ABSTRACT

Personalised nutrition is at its early stages but shows the potential of improving the health of the general population, at a time when diabetes and obesity are becoming worldwide epidemics. However, it will need to be based on rigorous scientific research, as well as being accompanied by public policies and ethical considerations.

Research is making great progress towards the understanding of the impact of genetics on complex diseases, which involve hundreds, or thousands, of variants, each having varying effect on the disease. Personalised medicine aims at harnessing this genetic information to tailor prevention and treatment according to each individual.

Unfortunately, the links between the genotype and the phenotype are not yet fully understood. And while the content of publicly available genetic databases is exponentially growing, they are often using different formats and means of access, making it difficult to get complete information. Moreover, evaluating the genetic predisposition of an individual to a disease is not straightforward, and while Polygenic Risk Score models can help in this regard, they are often only based on common variants, which might lead to misevaluation of the risk for rare-variants carriers.

In this thesis will be presented (i) VarGen, an R package to merge information from different genetic databases, which has the potential to infer new variant-disease relationships. (ii) a new method to improve Polygenic Risk Score models, which includes variants obtained from VarGen on top of the common variants from standard polygenic analyses. (iii) the results of a microRNA differential expression analysis, aiming at identifying the impact of microRNAs, on the development of severe Hypoxic-Ischemic Encephalopathy in new-borns.

Keywords:

VarGen; R package; Polygenic Risk Scores; Genome Wide Association Studies; Polygenic Risk Score, diabetes; obesity; body mass index; microRNA; Hypoxic-Ischemic Encephalopathy.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my family, especially to my parents, who always advised and supported me.

Fady, thank you for your help and guidance, from when I started the MSc at Cranfield, you have been an excellent supervisor. Tom, thank you for your help throughout the years, as well as the interesting discussions (and beers). I would also like to thank the other members of the bioinformatics team, Maria, Ewelina, Mariam, Emma, Faisal and recently Alexey, I had a great time working with you.

Thanks to my external supervisor, Prof Andrew Thompson, for your help, and all the rest of the *Molecular Plant Sciences and Bioinformatics Forum* members (Zoltan, Carol, Sofia, Emmanuel, Kyle, and many others) for all the interesting scientific discussions.

I would also like to thank my friends, for the much-needed relaxing breaks during the past three years, Nicolas, Audrey, Lucile, Dimitri, Benoit, Pierre, and Mathieu.

Thank to Café Comet in B83 for their smiles and delicious home-made cakes.

I would like to thank Matthew Brember for his help during the development of VarGen, as stated in Chapter 3, and Dr Alex Gutteridge for sharing his thoughts on the PRS analysis.

This work has received funding by the European Union's Horizon 2020 Research and Innovation Programme through NUTRISHIELD project under Grant Agreement No. 818110. This thesis reflects only the authors views; the European Union is not liable for any use that may be made of the information contained therein.

“When we try to pick out anything by itself, we find it hitched to everything else in the Universe”

John Muir

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

API	Application Programming Interface
BMI	Body Mass Index
BMIQ	Body Mass Index Quantitative trait locus
CADD	Combined Annotation-Dependant Depletion
DM (T1DM/ T2DM)	Diabetes Mellitus (type 1 / type 2)
DNA	DeoxyriboNucleic Acid
EBI	European Bioinformatics Institute
ECM	ExtraCellular Matrix
eQTL	Expression Quantitative Trait Loci
FANTOM5	Functional ANnoTation Of the Mammalian genome 5
FP / TP	False Positive / True Positive
GATK	Genome Analysis Tool Kit
GTEx	Genotype-Tissue Expression
GO	Gene Ontology
GRC	Genome Reference Consortium
GWAS	Genome Wide Association Study
HIE	Hypoxic-Ischemic Encephalopathy
HLA	Human Leukocyte Antigen
InDel	Insertion Deletion
Kbp / Mbp / Gbp	Kilo base pair / Mega base pair / Giga base pair
LD	Linkage Disequilibrium
MR	Mendelian Randomisation
miRNA	microRNA
NCBI	National Center for Biotechnology Information
NHGRI	National Human Genome Research Institute
OMIM	Online Mendelian Inheritance in Man
PRS	Polygenic Risk Score
RNA	RiboNucleic Acid
rsID	Reference SNP cluster ID
SNP	Single Nucleotide Polymorphism
TSS	Transcription Start Site
VCF	Variant Call Format
WHO	World Health Organisation

1 Executive summary

This thesis project was performed as part of the European Union's Horizon 2020-funded project Nutrishield (GA 818110), which focused on diabetes and obesity in young people as use-case models to develop a personalised nutrition platform. There is a general expectation from society that food consumed with the EU is safe, and current nutritional advice is given as a 'one size fits all' strategy. However, not every individual responds similarly to the same food or nutrient. This is determined by genetic factors, such as allergies or the tendency to develop certain diseases, as well as by acquired factors, such as the development of the microbiome, the amount of stress and exercise in daily life. In addition to the above, poor nutritional practice can lead to nutrition-related health conditions, including obesity, diabetes, heart diseases or cancer. One way to prevent these conditions is to identify high-risk individuals and personalise their lifestyle and diet. As a key component of risk identification is the genotype, this thesis will focus on the study of genetic variants linked to obesity and diabetes, followed by the development of polygenic models estimating the genetic risk of developing these diseases.

The literature review is presented in Chapter 2. As studying the genome is an integral part of personalised nutrition, the first part of the review will focus on genome sequencing and variant calling. The second part will describe the traits studied in the clinical studies from the Nutrishield project, namely diabetes mellitus and obesity. Finally, the current state of the art of personalised medicine will be mentioned.

The first objective of the thesis was to generate databases of variants related to diabetes and obesity. These variants will then be compared against the genotypes of the individuals enrolled in the Nutrishield clinical studies. Instead of building these databases manually, I developed VarGen, an R package to automatically retrieve a list of variants related to a trait, based on publicly available data. This package, and the results obtained with it are described in Chapter 3.

Chapter 4 describes a novel method to optimise Polygenic models, which estimate, based on information from many variants, the relative genetic risk to develop a disease. This method was tested on body mass index and validated on diabetes type 2. This new method uses variants from public studies as well as the results obtained from Chapter 3, with VarGen. Identifying individuals most at risk of becoming obese is an integral component of prevention and personalised nutrition.

Chapter 5 will conclude the core part of the thesis. Here, will be presented the conclusions about Chapter 3 and 4, with their limitations and general reflections on the use of genetics for personalised nutrition and medicine.

Finally, Chapter 6 presents the results of an additional project done in collaboration with a Nutrishield partner, *Hospital Universitari i Politècnic La Fe.* The goal was to perform microRNA differential expression analysis to identify potential biomarkers for the development of severe Hypoxic-Ischemic Encephalopathy in neonates.

The scripts used to perform the analyses and produce the figures are available via GitHub: https://github.com/MCorentin/PhD_scripts.

2 Literature review

2.1 Sequencing & variant calling

2.1.1 The human reference genome

The first draft of the human genome was released in 2001 by the International Human Genome Sequencing Consortium [1] [2]. The resulting sequence, generated from a mosaic of different donors, was ~3.2 Gbp long. It was a milestone in science and genetics, nevertheless obtaining this sequence on its own was not enough to understand how our genome works. It is now necessary to understand how DNA and its modifications are affecting our cells.

The Genome Reference Consortium (GRC) is responsible for maintaining the reference. As of 2021, the current version of the human genome is GRCh38 [3] (or hg38) but some databases and tools are still referring to GRCh37 (or hg19) coordinates. Fortunately, it is possible to convert the coordinates from GRCh37 to GRCh38 with the LiftOver tool.

Despite being the 20th version of the human reference genome, GRCh38 is not complete and still contains gaps, regions of unknown bases represented by Ns. The reference is still being worked on, notably around the repetitive regions and the heterochromatin, which are hard to sequence and assemble. Instead of waiting until a new reference version is released, the GRC decided to release assembly corrections as ‘fix patches’ (e.g., GRch38.p13 is the 13th patch for GRch38) so that researchers constantly have access to the most accurate information. It should be noted that the reference does not represent an actual genome, but a melting pot of the sequences found at each locus in different individuals. However, alternate sequences differing from the primary assembly, such as haplotypes and novel loci, can be of interest when studying certain populations or traits, these are represented as ‘novel patches’ in the assembly. Both fix patches and novel patches can be accessed from Ensembl [4]. On the same note, some ethnicities are not well represented by the main reference, which can affect the interpretation of genotyping results [5]. In this project, GRch38 served as the reference for all analyses.

Recently, the Telomere-to-Telomere Consortium, in a preprint, announced the release of the first complete sequence of a human genome [6]. This is promising towards getting a complete reference, but this sequence has limitations compared to GRch38, i.e., it is based on only one individual and is missing the Y chromosome.

2.1.2 Advances in sequencing technologies

Sequencing consists of determining the order of the nucleotides in a DNA sequence. Different sequencing technologies are available, each useful in certain applications. Due to current technical limitations, it is not possible to read entire chromosomes directly, and the DNA must be cut into pieces, called *fragments*, before the sequencing can take place. Each fragment is then sequenced as a *read,* their length depending on the sequencing platform. Second Generation Sequencing produces short reads, a few hundred bases long, which are very accurate (~99.9% of bases are correct). Third Generation Sequencing produces long reads, from a few kbp to hundreds of kbp long, but with a high error rate (~10% of the bases are wrong). Long reads are mostly useful in genome assembly, where they can span repeats and other problematic regions. For variant calling, on one hand their high error rate makes it difficult to separate true variants from sequencing errors, on the other hand, their length is essential to detect structural variations. Hence, the choice of the sequencing technology is dependent on the kind of variations one wants to detect. Recently, a new sequencing technology producing highly accurate long-reads, was developed and showing potential for improving the detection of structural variants [7].

Due to the falling costs and improvements in sequencing technology, it is now possible to get up to a terabase of DNA sequence in two days and for a few thousand dollars [8]. This led to an explosion of the production of sequencing data, and now the bottleneck is around the computational power needed to analyse and understand the function of the different parts of the DNA.

2.1.3 Variants and genotyping

2.1.3.1 What is a variant?

Approximately 0.6% of the nucleotides are different between two persons, corresponding to ~20 million variants [9]. These variations can occur on a single nucleotide, in which case they are called **Single Nucleotide Polymorphisms** (SNPs), or can span several nucleotides, in which case they are called **InDels** (Insertions, Deletions). The latter is composed of two categories, ‘deletions’ (the individual is missing nucleotides compared to the reference) or ‘insertions’ (the individual has extra nucleotides compared to the reference). Usually, the base corresponding to the reference is called the ‘*reference allele*’ and the variation is called the ‘*alternative allele*’.

Previously identified and annotated human variants are assigned a unique Reference ID (rsID) by dbSNP [10], a genetic database managed by the National Center for Biotechnology Information (NCBI). Each rsID refers to a locus containing a certain type of variation (SNP or InDel) and is stable across different human assemblies, thus providing a point of reference for variant analysis. Variants submitted to dbSNP are mapped to the most recent reference genome and merged to create a non-redundant list of rsIDs. As of November 2018 (Build 152), dbSNP contains more than 650 million entries.

Modifications of the genome also include Copy Number Variations (CNVs), which correspond to large insertions or deletions in the genome. CNVs are sometimes linked to diseases [11] , a well-known example being Down Syndrome, where the whole length of chromosome 21 is duplicated [12]. The difference between InDels and CNVs resides in their lengths, while no official consensus exists, traditionally variations longer than 1 kbp are considered CNVs.

Overall, SNPs, InDels and CNVs can be beneficial, neutral, or causative of disorders.

2.1.3.2 Variant calling

Variant calling is the process of comparing an individual genome against a reference to find differences, the so-called variants. Several variant calling tools

exist, the most popular are ‘samtools mpileup’ [13], ‘DeepVariant’ [14] and ‘GATK’ (Genome Annotation Tool Kit) [15]. The rationale is the same between the different tools, first the reads are pre-processed, then aligned against the reference genome and finally variants are called based on the differences detected during the alignment. Sometimes a calibration step is performed to improve the accuracy and avoid false positives, this is the case for GATK’s best practices [16].

The accuracy, or confidence, of the variant calling is dependent on different factors. First, the quality of the reads is important, as low quality will lead to uncertainty when trying to differentiate between actual variants and sequencing artifacts. This is why it is necessary to perform a quality control step, for example with FastQC [17], and remove low quality bases and sequencing adapters, as needed. Second, the number of reads available at a certain base, also called coverage or depth, will determine the confidence of any variant detected at this position. Indeed, the confidence will increase if more reads possess the alternative allele, this is especially true for heterozygous variants since they are only present in half of the reads.

The standard output from variant callers is a Variant Call Format (VCF) file. The current version of the specifications is v4.3, available at https://github.com/samtools/hts-specs. Below is an example of a VCF file:

```
##fileformat=VCFv4.3
##fileData=20090805
##reference=file:///seq/references/reference.fa
##contig=<ID=1,length=88663952>
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples with Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##FILTER=<ID=q10,Description=Quality below 10">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM POS     ID      REF     ALT     QUAL    FILTER INFO                    FORMAT NA0001
1      5093    .       T       A       12      PASS   NS=1;DP=5;AF=0.75       GT:GQ  1/1:48
1      5094    .       C       T       3       q10    NS=1;DP=15;AF=0.25      GT:GQ  0/1:49
1      6018    .       G       C       32      PASS   NS=1;DP=25;AF=0.50      GT:GQ  0/1:21
1      6059    .       A       G       45      PASS   NS=1;DP=16;AF=0.75      GT:GQ  1/1:54
1      7201    .       GA      G       39      PASS   NS=1;DP=13;AF=1.00      GT:GQ  1/1:45
```

Figure 1: Example of a VCF file. The header (lines beginning with ## or #), contains metadata related to the variant calling and describes the content of each column. NA0001 is the genotyped sample name. Finally, the last 5 lines are representing a variant each.

The file begins with metadata lines, starting with '##', containing optional information about the variant calling which led to the creation of the VCF file. This is organised as 'key=value' pairs. Next comes the header line, starting with '#', containing the name of the columns, 8 of which are mandatory:

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
Chromosome	Position	Identifier rsid	Reference allele	Alternative allele	Quality score	Filter status	Additional information

The first five columns contain information about the variant itself, its position, identifier, and alleles. The QUAL column contains a Phred-scaled score representing the confidence that a variant is actually present at this position. The FILTER column indicates if the variant passed all filters, confirming the call, if not, different codes pinpoint to the failed filters (e.g.: 'q10' means a quality below 10). The INFO column contains additional information, separated by semi-colons and following a 'key=value' format. Examples of INFO fields include DP, for the combined depth across samples; AF, for the allele frequency. If there is genotyping information in the VCF, then more columns are added: FORMAT and samples IDs (one per sample). The FORMAT column describes the format that will be followed by each sample column, which contains genotyping information for one sample.

The rest of the file consists of data lines, one variant per line.

Variant calling does not give any indication regarding the deleterious nature of each variant. Therefore, an annotation step is necessary to prioritise and identify variants linked to disorders.

2.1.3.3 Variant annotation

Annotation is an important step as it identifies the potential impact of each variant. Many tools are available to perform this task. Those considered in this project are MyVariant.info [18], Variant Effect Predictor [19] and SnpEff [20]. In addition to the tools, one can choose different transcriptomes for the annotation. The choice of both tool and transcriptome is not trivial and can have an impact on the results [21].

The most common approach to annotate a variant is to predict how it will affect the genes and thus the resulting proteins. The type of consequences depends on the location of the variant on or around the gene. For example, Variant Effect Predictor contains more than 30 categories, such as *intergenic variant*, *start lost* or *splice donor variant*. And an *intergenic variant*, because it is located in a non-coding part of the gene, will often have a much smaller impact than a *stop gained* variant which will shorten the resulting protein, rendering it less effective or even useless. Nothing prevents a variant to be annotated with more than one consequence, or to be associated with more than one gene. SnpEff, in addition, has a predefined high-level categorisation of the SNPs, depending on their severity: *MODIFIER, LOW, MODERATE* and *HIGH*.

That being said, non-coding variants should not be overlooked, as they can regulate or disrupt biological processes by impacting gene expression [22]. However, studying the impact of variants on gene expression is not straightforward and linking a non-coding variant to a gene is not trivial. Still, some projects have been developed to estimate the effect of variants on gene expression. A successful example of this is the ambitious Genome Tissue Expression (GTEx) project [23], which identified expression Quantitative Trait Locus (eQTLs) in 49 tissue types.

The annotation of variants provides an excellent mean of filtering variants of interest. This must be done with caution however, as sometimes a low impact variant might still be the one responsible for the phenotype. Finally, it is important to remember that variant annotation relies on *in-silico* prediction and must be confirmed via experimental methods.

2.1.3.4 Investigating causation

Associations between a risk factor and a trait are usually hard to assess. And it is even more complex to investigate causation, notably because of the interaction between the genotype and the environment, confounding factors and reverse causation.

Genes and the environment can interact in different ways. Ottman described five different models for these interactions [24] (i) the genotype produces or aggravates an environmental risk factor (ii) the genotype increases the effect of the environment, but the genotype is irrelevant without exposition (iii) the environment aggravates the effect of the genotype, but without effect to low-risk genotypes (iv) the genotype and the environment are both required to exacerbate the risk (v) the genotype and the environment both have an effect and this effect is lower or greater when they happen together. In nutrition, some of these interactions can hinder attempts to lose weight, for example SNPs found in PERIOD2 were associated with snacking, stress from dieting and skipping breakfast in carriers [25].

Other important factors that hinder causation investigation are confounding factors and reverse causation. While it is possible to account for confounding factors, current methods are not powerful enough to obtain statistically robust results. Moreover, Westfall and Yarkoni recently demonstrated that even moderate unreliability in measuring these confounders can lead to high Type 1 error rates [26]. Different methods exist to infer causation between a risk factor and a trait. The current gold standard to study the impact of exposures on traits is to perform a Randomised Controlled Trial (RCT), but it would not be ethical to subject individuals to certain risk factors (e.g.: smoking, alcohol consumption). Moreover, some diseases are triggered a long time after the initial exposure (e.g.:

cancer) and for these, RCTs would need to span over decades to obtain significant results. Finally, individuals participating in RCTs are often selected to avoid co-morbidities and belong to specific age groups, hence, are not representative of the whole population [27].

Mendelian Randomisation (MR) is able to circumvent these limitations. MR uses genetic variation as an Instrumental Variable (IR) to study the impact of a risk factor on a disease or trait. Instrumental variables were originally developed in the field of econometrics. Instead of directly studying the impact of an exposure X on an outcome Y, IR uses a variable Z, associated with the exposure X, to study the impact of X on Y. This is useful to study the impact of X on Y while ignoring confounding factors (that might affect X, but not Z). In MR, Z is a genetic variant, X is the risk factor and Y is the disease or trait of interest.

In MR, the random allocation of alleles from parents to offspring ensure that variants are not correlated with confounders (especially, lifestyle and socio-economic factors). Second, diseases do not alter germline variants, so genotype-disease associations are not affected by reverse causation [27].

MR relies on three assumptions. (i) The genetic variant (used as an IV) needs to be robustly associated with the risk factor, preferably in multiple studies (ii) During the trial, the variant needs to be randomised with respect to confounders (iii) There should be no horizontal pleiotropy, i.e.: the variant should not affect the outcome via a pathway that does not involve the exposure of interest [27].

There are still limitations pertaining to MR. First, there is the issue of population stratification, alleles might be distributed differently in different population which violate the assumption of randomisation. Thus, MR analysis should be performed on homogeneous populations, use multiple genetic variants and/or focus on parent-offspring groups. Second, pleiotropy is hard to assess, and the variants used as instrumental variable might be affecting the outcome via another pathway than the exposure under study.

The importance and reliability of MR studies will increase as more variant-phenotype associations are made, notably through Genome Wide Association Studies (GWAS).

2.1.3.5 Genome Wide Association Study

A Genome Wide Association Study (GWAS) aims at identifying genetic markers, usually SNPs, associated with a trait. They are based on the comparison between the genomes of individuals with and without the trait (or with varying phenotypes for continuous traits) in a population. Thus, each SNP is assigned a p-value and an effect size, depending on the allele frequency difference between the cases and controls.

GWAS are becoming popular due to their many advantages. They focus on the whole genome, thus are not limited to coding regions. They do not need any prior knowledge about the trait under study and they can be used for both continuous (e.g.: height) and discrete (e.g.: presence / absence of diabetes) traits. For example, GWAS have successfully identified genetic variants linked to diabetes mellitus type 2 [28], coronary artery disease [29] and even Body Mass Index (BMI) [30]. Since the creation of this method, GWAS led to many discoveries in complex traits, furthering our understanding of genetics and the development of new therapeutics [31].

GWAS rely on statistical significance thresholds to differentiate between True Positives (TP) and False Positives (FP). Two thresholds are usually considered, 'suggestive' and 'significant'. The suggestive threshold, $p\text{-}value < 1*e^{-05}$, was suggested by Lander and Kruglyak and represents the threshold where one false positive is expected per genome scan [32]. The significant threshold, $p\text{-}value < 5*e^{-08}$, comes from the Bonferroni Correction, where the original p-value is divided by the number of independent common variants across the genome [33]. The latter threshold is the most popular and works very well with common variants, however it might be less reliable when dealing with rare variants [34].

Some limitations must be kept in mind when designing or analysing the results of a GWAS. They need a large population to be able to confidently identify common

variants, thus can become expensive to set up. In certain cases, the variants found do not explain all the variance observed in the trait, the so-called 'missing heritability' [35]. One of the main caveats from GWAS comes from the variant calling strategy, using genotyping chips, resulting in identifying not the causal variant, but one in Linkage Disequilibrium with it, which can hinder interpretation and generalisation of the findings to other populations [36]. On a broader view, interpreting the effect of the significant variants is often challenging, especially in the non-coding parts of the genome [22]. Indeed, one of the surprising insights gathered from the GWAS is that non-coding variants are playing an important role in complex human traits and diseases. A study, based on 151 GWAS, found that ~90% of trait/disease-associated SNPs were in non-coding regions (intergenic or intronic) [37].

GWAS results are often given as *summary statistics*, which contain all the genotyped variants with their associated p-values. Additional information can be provided, such as the effect allele and the odds ratio. GWAS summaries statistics for a wide range of phenotypes are accessible from several resources. Notably, the GWAS Catalog, maintained by the NHGRI-EBI [38], GWASdb [39] and GWAS central [40].

Box 1: GWAS glossary

Effect Size: estimation of the SNP *impact* on the genetic variance for the trait. Given as **Odds Ratio** for discrete traits or **Beta** for continuous traits.

Standard Error: standard error of the effect size estimate. Depends on the cohort's size.

Effect allele: allele responsible for the effect size observed. note: it is not always the minor allele at this locus.

Non-effect allele: the other allele, not responsible for the effect size observed.

2.1.3.6 Polygenic Risk Scores

GWAS marked a shift in our understanding of genetics. It is now clear that complex diseases are often due to the accumulation of a large number of small impact variants, in contrast to the rare monogenic variants, which increase the risk of developing a disease by several folds. However, clinical risk identification mostly relies on the latter. While being invaluable for the concerned individuals, these mutations are only affecting a small portion of the population, thus the current identification method is potentially missing out high-risk individuals. Fortunately, recent approaches managed to integrate genome-wide polygenic scores that can harness the wealth of genomic data now available [41]. Better risk identification will lead to better prevention and better understanding of the diseases, which in turn open new therapeutic avenues. These new approaches are relying on 'Polygenic Risk Scores' (PRS). A PRS is a model assigning a relative risk for a given individual based on their genetic profile. They are created from a set of genetic markers (variants) that are linked to a certain disease. These markers are often found from GWAS (see 2.1.3.5) and can be given a weight representing how strongly they are correlated with the disease.

For a PRS analysis, one needs two sets of data, the **base** and the **target**. The base set will be used to define the list of variants to be included in the model (one can use results from a GWAS). Once defined, the PRS model will be applied to the target set, containing genotype information, which will result in a risk score being assigned to each individual in the set. One should keep in mind that the scores obtained are relative within the target group.

The unit of the PRS depends on the trait under study. For a continuous trait, the PRS unit will follow the effect size estimate from the GWAS analysis it is based upon. For example, if a GWAS is reporting the change in heart rate in *beats per minute*, then any PRS based on this data will use the same measure. For a discrete trait, usually when doing a 'case vs control' comparison, the PRS will be reported as *log of odds ratio* (log(ORs)) [42].

Developing a PRS model is not straightforward, as one must differentiate between noise and effect for each variant. Moreover, some variants are inherited

together, making it harder to assess the impact of each particular SNP. Relatedness between individuals can inflate the relation between the genotype and the phenotype. In addition, population structures can also create structures in genetic variations, this can have an impact on a PRS analysis, as the base and target sets might come from different populations. On the same note, this is why PRS models are not generalizable between different ethnicities.

The most common approach to develop a PRS model is the C+T method, which consists of Clumping variants to get a subset of independent SNPs, followed by a Threshold on the GWAS association *p-value*. The C+T method is implemented in Plink [43] and PRSice-2 [44]. Other approaches include the Bayesian model of LDpred2 [45] or the penalised regression of lassosum [46].

The reader should keep in mind that, rather paradoxically, a good PRS is not necessarily a good screening test. Mainly, because PRS are generally comparing the extreme high and low risk individuals, missing out those falling in the middle [47]. There is also the issue of 'penetrance': how many people within the high-risk quantiles will develop the disease? Wald et al. demonstrated that a good PRS for coronary artery disease had a detection rate of only 15%, meaning that it would miss 85% of the cases [47]. For schizophrenia, even if a PRS in the top decile would provide a fivefold increased risk, there would still be a >95% chance of not developing the disease [48]. However, this does not mean that PRS are useless, as they still provide an excellent complementary assessment to the standard clinical risk factors. Torkamani et al. identified three main use of a PRS model: PRS-informed therapeutic intervention, PRS-informed disease screening and PRS-informed life planning [49]. Moreover, PRS remains a powerful tool to understand the genetic causes of diseases and identify high-risk individuals.

2.2 Diabetes mellitus

2.2.1 A brief history of diabetes

Diabetes mellitus (DM) is one of the oldest known diseases, first described in Egypt and India around 1500 B.C., as a condition described by ‘too great emptying of urine’ or a disease causing ‘urine attracting ants’ [50] [51]. This refers to two symptoms of diabetes, Polyuria (abundance of urine) and the presence of sugars in urine (*mellitus* is Latin for ‘sweet like honey’). Sushruta and Chakara in ancient India, around 500 BC, identified the two main type of diabetes, currently named type I and type II [50] [51]. Aretaeus Cappadocian, in 100 A.D., wrote the first comprehensive description and named the disease as diabetes (Greek for ‘siphon’), ‘no essential part of the drink is absorbed by the body while great masses of the flesh are liquefied into urine’ [52].

Our knowledge of diabetes at the molecular level improved considerably during the 19th and 20th centuries. Indeed, in the middle of the 19th century, Claude Bernard discovered the role of the liver in the pathway of gluconeogenesis [53]. The next step was made by Paul Langerhans, in 1869, when he identified the *islets of Langerhans* a key component in the understanding of diabetes [54]. In 1889, von Mering and Minkowski designed an experiment on dogs, and found that removing the pancreas led to diabetes [55]. Almost at the same time, in 1909 and 1910, Mayer and Schaefer discovered the product of the islets of Langerhans: the hormone *insulin* (from Latin, meaning island) [56] [57]. Insulin is the key hormone in diabetes pathogenesis and management. Indeed, ten years later, in 1921, Banting, Best and Collip, also experimenting on dogs, managed to reverse induced diabetes with a treatment based on canine insulin [58]. This was the first proof that insulin deficiency was the central component of diabetes. Just one year later, an event will start a revolution in diabetic therapy, Leonard Thompson, a 14-year-old boy, was saved with an infusion of bovine insulin at the Toronto General Hospital [58]. The next revolution in therapy came with the manufacture of recombinant human insulin in 1978 by Goeddel and his colleagues of Genentech [59] [60]. A timeline of the milestones described above is available as Figure 2.

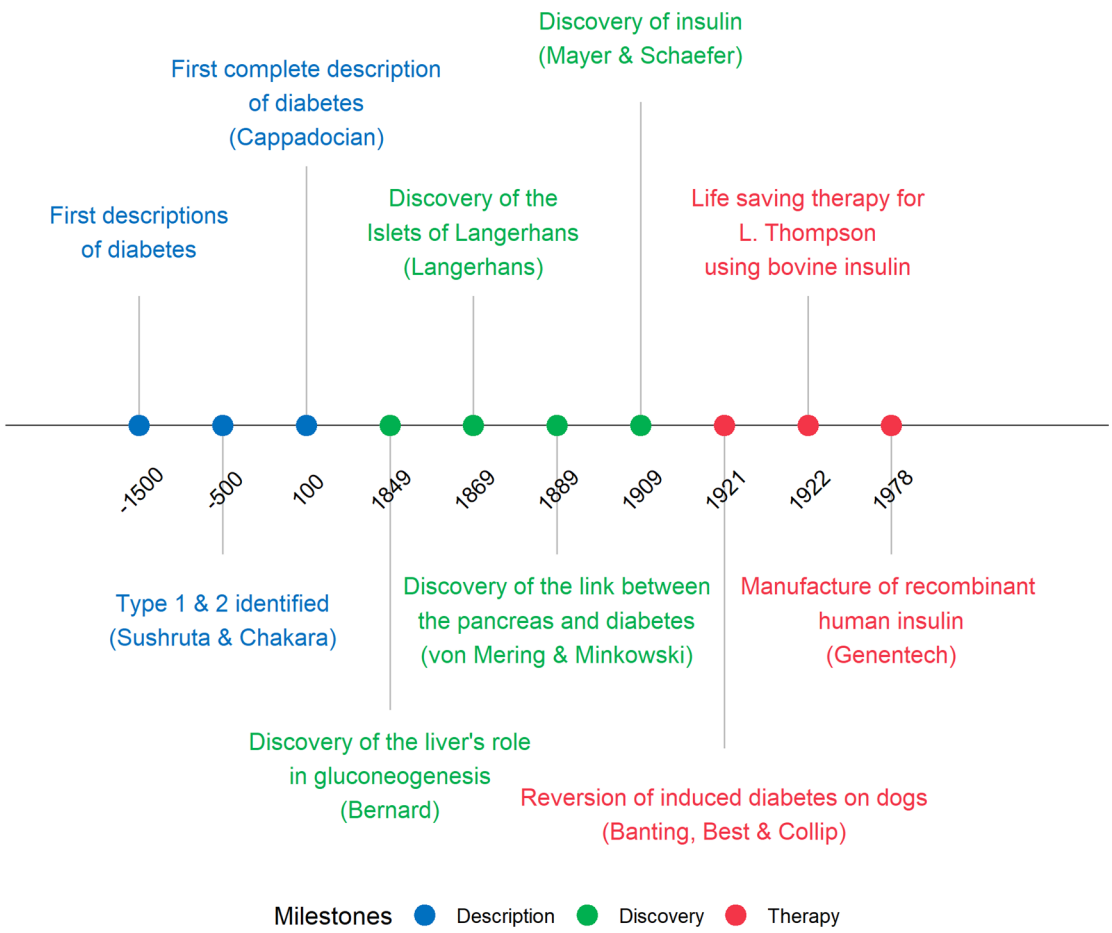


Figure 2: Timeline of diabetes mellitus history. The milestones are arbitrarily classified into three categories (Description, Discovery, and Therapy).

2.2.2 Diabetes: beyond two types.

The current definition from the World Health Organisation (WHO) for diabetes mellitus is '*a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.*' [61].

The two recommended guidelines for the screening of DM are the one issued in 1997 by the *American Diabetic Association*, based on the fasting plasma glucose, and the one from 2006 by the WHO, based on the Oral Glucose Tolerance Test.

Three main types of diabetes mellitus have been identified so far, gestational diabetes mellitus, diabetes mellitus type 1 and type 2. There are other, rarer, type of diabetes: monogenic diabetes and secondary diabetes (see Table 1). In addition, there are two intermediate conditions called ‘Impaired glucose tolerance’ and ‘Impaired fasting glycaemia’, which indicate a high risk of progressing to type 2 diabetes mellitus (T2DM). These correspond to a higher-than-normal blood glucose and fasting plasma glucose levels respectively, but not high enough to diagnose the patient as diabetic.

Table 1: Description of the traditional main types of diabetes. (MODY = Maturity Onset Diabetes of the Young)

	Aetiology	**Definition**	**Symptoms**
Diabetes mellitus type 1	Autoimmune disease destroying the pancreatic beta-cells	Deficient insulin production	Increased thirst / appetite Abundance of urine Weight loss Tiredness
Diabetes mellitus type 2	Genetics, excess body weight and physical inactivity	Deficient insulin production and / or Ineffective use of insulin	
Gestational diabetes	Not enough hormone production by the placenta to meet the extra needs during pregnancy.	Deficient insulin production. Ends after the delivery.	
Monogenic diabetes (e.g., MODY)	Mutation in an autosomal dominant inherited form of diabetes	Deficient insulin production	
Secondary diabetes	Appears as a co-morbidity of other diseases or drugs	Very diverse category	

Recent studies are suggesting more granularity in diabetes than the historical two types. One study identified the existence of two endotypes of type 1 diabetes depending on the age at diagnosis [62]. Another study assessed the role of genetics and the growing impact of the environment, including microbiota and nutrition, in the heterogeneity of diabetes type 1 [63]. For diabetes type 2, some studies divided patients into subgroups with differences in genetics, disease progression and complications [64] [65]. For example, Ahlqvist et al. identified five subtypes of dibetes by performing hierarchical and k-means clustering based

on different biomarkers: age at diagnosis, BMI, HbA_{1c}, glutamate decarboxylase antibodies and homeostatic model assessment of β-cell function and insulin resistance [66]. The five subtypes correspond to different facets and aetiologies of diabetes: (i) severe autoimmune diabetes (equivalent to type 1 diabetes) (ii) severe insulin-deficient diabetes (iii) severe insulin-resistant diabetes (iv) mild obesity-related diabetes (v) mild age-related diabetes. These findings highlight the heterogeneity of diabetes and could open new therapeutic avenues, tailored towards the specificities of each patient.

Diabetes often leads to complications in other organs. People with diabetes have a higher risk of developing: periodontal disease, diabetic retinopathy, cardiovascular disease, renal disease, and lower limb amputation [67].

2.2.3 The genetics of diabetes mellitus

Genetics plays a role in both type 1 and 2 diabetes mellitus.

Type 1 diabetes mellitus (T1DM) is partly inherited, with a 0.4% risk in the general population which rise to 6 or 7% if the individual has an affected sibling [68], plus certain Human Leukocyte Antigen (HLA) haplotypes are associated with the disease. HLA is a genetic system, located on chromosome 6, which role is to generate antigens [69]. It spans ~3.6 Mbp and is composed of 3 regions (i) **class I** which houses the *HLA-A, HLA-B* and *HLA-C* genes, encoding for the heavy chains of the class I molecules (ii) **class II**, which is of interest for diabetes, and is divided in three subregions, DR, DP and DQ which are responsible for producing *HLA-DR* antigen specificities, DP and DQ molecules respectively (iii) **class III** which contains genes leading to the production of additional components for the HLA molecules [69]. As mentioned, certain polymorphisms in the DR and DQ regions are strongly associated with T1DM, for example the haplotype *DR4-DQ8/DR3-DQ2* has an average Odds Ratio of 16 for diabetes [68]. But some non-HLA genes are also playing a role in T1DM and studying them can help to shed light on the aetiology of the disease, as will be discussed below.

T2DM has a strong genetic component, as shown by the higher concordance rate in monozygotic compared to dizygotic twins [70] and the 40% to 70% lifetime risk

of developing diabetes when one parent or both has T2DM respectively [71]. In recent years, GWAS identified many loci related to T2DM, some of them involved in previously unexpected mechanisms, such as the *circadian rhythm*, *zinc transport* or *cell cycle regulation* [72].

As of 2021, the Online Mendelian Inheritance in Man (OMIM) database [73] is linking 4 genes to T1DM (OMIM: 222100) and 27 genes to T2DM (OMIM: 125853) (see Table 2). This difference in the number of associated genes could be due to a difference in the complexity or in the number of available studies between the two types.

Table 2: List of genes associated with type 1 and type 2 diabetes mellitus, according to the Online Mendelian Inheritance in Man database. The only gene in common between the two sets is 'HNF1A'.

Trait	Hugo symbol
Type 1 diabetes mellitus	*ITPR3, PTPN22, IL6,* ***HNF1A***
Type 2 diabetes mellitus	***HNF1A****, HNF1B, HNF4A, PDX1, IRS1, IRS2, PTPN1, HMGA1, TCF7L2, LIPC, PAX4, SLC2A2, PPP1R3A, AKT2, MAPK8IP1, IGF2BP2, RETN, SLC30A8, MTNR1B, GPD2, GCK, ENPP1, WFS1, KCNJ11, ABCC8, PPARG, NEUROD1*

The only gene in common between the two types is *HNF1A*, which is a transcription factor for multiple genes activated in pancreatic islet cells and in the liver. Mutations on *HNF1A* were identified in monogenic forms of diabetes, but its role in the pathogenesis of diabetes is not yet fully understood. A recent protein-protein interaction analysis found associations between *HNF1A* and proteins involved in the uptake of glucose and hormone production in the beta cells [74].

Most of the T1DM genes are associated with the immune system, *PTPN22* encodes a suppressor of T-cell activation [75], while *IL6* is an interleukin with roles in immunity, tissue regeneration, and metabolism [76]. This and the previously discussed role of the HLA support the concept of T1DM being an autoimmune disease. The last gene, *ITPR3*, is a mediator of intracellular calcium release, which is important for Ca^{2+} dependent insulin secretion, and SNPs within this gene were associated with T1DM in Swedish individuals [77].

An enrichment analysis was performed on the T2DM genes against all *the Homo sapiens* genes in the PANTHER database [78], with the annotation derived from the Gene Ontology (GO) database (Released 2021-07-02) [79]. Overrepresented GO terms in the T2DM gene list were, as expected, mostly related to insulin, pancreatic cells, and glucose, with terms such as *insulin secretion, insulin receptor signaling pathway, detection of glucose, glucose metabolic process, hepatocyte differentiation*. Other terms include, *reverse cholesterol transport, NADH metabolic process, response to drug,* some terms are also enriched in obesity, as will be discussed in Section 2.3.4.2. The complete list of enriched terms is available in Table A.1-1.

2.2.4 A 21st century epidemic

Diabetes is on the rise around the world, as shown by the 'Diabetes Atlas 8th edition' (2017) from the International Diabetes Federation [80]. This report highlights that 10 million more adults were diagnosed with diabetes in 2017 compared to 2015. In addition, 34 million more adults were at risk of developing diabetes in 2017 compared to 2015. Moreover, the report estimated that between 30 and 80% of adults with diabetes were undiagnosed [80]. Once thought to be a disease affecting only adults, T2DM is now increasingly detected among children and adolescents. There is an growing trend in both prediabetes and T2DM for this population, estimated to continue in 2030 [81].

It is now clear that patient involvement is a key component of diabetes management. As no cure exists yet, treatment and prevention are aimed towards improvements in the patients' quality of life.

The Horizon 2020-funded project Nutrishield is aiming at helping this side of the disease management by personalising the diet. Indeed, the best prevention approach for T2DM is known since the antiquity and is still advocated today, i.e. exercise and a healthy diet [50], especially if tailored toward each individual [82].

2.3 Obesity

2.3.1 A short definition

Obesity is a chronic disease, characterised by excess of fat content and a modification of the adipose tissue. The adipose tissue contains the fat cells (adipocytes) and is found in the hypodermis (the tissue found between the muscle and the dermis). The WHO defines obesity via the Body Mass Index (BMI), which is obtained by dividing a person's weight in kilograms per square of his height in meters (kg/m^2). An adult is considered overweight with a BMI greater or equal to 25 and obese with a BMI greater or equal to 30 [83]. BMI is easy to measure and informative, however, it depends on ethnicity and does not reflect the repartition of adipose tissue in the body, which is important to assess the potential health impacts of obesity [84].

2.3.2 Adipose tissue or adipose organ?

The adipose tissue has, for a long time, thought to be hormonally inactive, as its role was reduced to energy storage and thermal insulation. However, recent studies have recognised it as an important endocrine organ, interacting with various organs, including the central nervous system [85]. A key component of appetite regulation and obesity is leptin (as discussed in 2.3.4.1), and the adipose tissue is the main producer of this hormone. Thus, the endocrine role of the adipose tissue is more complex and important than initially thought. The impact of this organ on health can be different depending on the type of adipose tissue (white or brown) and the location of this tissue (subcutaneous or visceral).

2.3.2.1 White versus brown adipose tissues

Two main types of adipose tissues have been identified, white and brown. They differ in their role and composition.

The white adipocytes' main function is to store energy. This is reflected by their spherical shape, due to a large, single lipid droplet filling 90% of the cell volume. The mitochondria in these cells are sparse, thin and elongated [86]. This is the most prevalent type of adipose cells in adults, and it plays an important role in

many biological processes. For example, the leptin hormone, a regulator of appetite, is produced by the white adipose tissue.

The brown adipocytes' main function is thermogenesis. Thus, they are filled with several, small liquid droplets, and many mitochondria, giving them their colour. The role of mitochondria is to produce *adenosine triphosphate*, which, in brown adipose tissue, is circumvented by the *uncoupling protein 1* (UCP1), allowing the energy to be released as heat. This explains the prevalence of brown tissues in new-borns and small mammals. As a result of the low ratio of body volume to body surface, they need powerful, non-shivering, thermogenesis to combat cold temperatures, compared to adults of larger mammals [86]. The amount of brown adipose tissue in the adult population is undetermined but was estimated to be ~10%.

It has been theorised that white adipose tissue cells, in response to cold, sometimes transform into *beige* cells, a process called 'transdifferentiation'. Beige cells are halfway between a brown and a white adipocyte, having positive UCP1 expression, medium mitochondrial density, and multiple lipid droplets. Enhancing this transdifferantiation process is considered as a potential therapeutic approach to rebalance the adipose metabolism and treat obesity [87].

In terms of clinical impact, an excess of white adipose tissue is linked to obesity, while brown adipose tissue is associated with a lower BMI. An imbalance in the amount of white adipocytes can lead to metabolic dysfunctions, such as hyperglycemia, diabetes, or cancer [88]. Moreover, white adipose tissue is affiliated with inflammation, notably with an increased secretion of molecules such as TNF-alpha and interleukin-6, which can also impact the insulin signalling pathway [89]. This might explain the inflammatory aspect of obesity pathophysiology and its link with diabetes.

2.3.2.2 Subcutaneous versus Visceral adipose tissues

Subcutaneous and visceral are two different types of adipose tissues, defined by their anatomical locations, which also differ in their morphologies, mechanisms,

and impacts on health. As their names suggest, subcutaneous fat is located beneath the skin, while visceral fat is lining the internal organs of the body.

An excess of visceral fat is associated with cardiovascular diseases, and visceral adipose tissue amounts to ~10% and ~20% of the total fat mass in lean and obese subjects respectively. The remaining 80-90% corresponding to subcutaneous adipose tissue. The expansion of waist circumference can be used as a proxy to monitor the increase in visceral adipose tissue [90].

In terms of composition, visceral fat depots are mainly composed of white adipocytes, which serve as energy storage. Subcutaneous depots, on the other hand, comprise both white and brown adipocytes, as well as interstitial tissue. Adipocytes have a shorter lifespan in subcutaneous compared to visceral tissue, and 'younger' adipocytes are not associated with metabolic disorders [88].

In terms of endocrine function, leptin, angiotensinogen, and glycogen synthase are favourably expressed by the subcutaneous tissue, while the insulin receptor, 11β hydroxysteroid dehydrogenase, and interleukin 6 are more expressed in visceral fat depots.

The combination of insulin-resistance, white adipocytes composition, older cells, and inflammatory endocrine function, explain the higher health implications of visceral over subcutaneous fat.

2.3.3 Causes and health impacts of obesity

The aetiologies of obesity are multiple and complex. The main risk factors are genetics, socio-economic components (with lower revenue being associated with higher obesity rates), and finally the lifestyle, especially the quality of the diet and the amount of physical activity. Other noteworthy factors include the microbiome, and the environment, notably stress, pollution, and some drugs.

Obesity itself is a very heterogeneous disease, whose mechanisms and impact on health depends on several components, such as fat distribution, metabolic disturbance and presence or absence of comorbidities. Indeed, obesity can lead to diabetes, arthrosis, and cardiovascular diseases. Moreover, the potential

psychological repercussions of this disease, such as depression, anxiety, and addiction should not be ignored, especially in children. All of these factors may contribute to the reduction of the lifespan of individuals, but, fortunately, even small weight loss leads to noteworthy health benefits [91].

Obesity seems to also affect the lungs, and obese individuals are more at risk of being hospitalised for respiratory infections compared to healthy weight individuals. This is caused both by fat deposits around the thorax, which are affecting the mechanical functions of the lungs and the increased production of inflammatory molecules in the adipose tissue, which increases airway inflammation [92]. This makes obesity a recognized risk factor for asthma, and explains in part the higher prevalence of adverse outcomes after an infection with SARS-CoV-2 noticed for individuals with a higher BMI [93].

2.3.4 The genetics of obesity

The genetics behind obesity in humans are hard to study for several reasons. First, obesity is a highly polygenic disease, with genes affecting different processes, such as energy balance and appetite. Second, obesity itself is a heterogeneous disease, composed of different subtypes. Third, obesity is cofounded by the environment (lifestyle, diet), so genetically susceptible individuals do not necessarily show the phenotype. Thus, obesity can be seen as a web of complex interactions between many genes and the environment [94].

Yet, the impact of genetics in obesity is obvious. Maes et al. reviewed the familial resemblance of BMI from twin studies [95]. They highlighted a higher BMI correlation within monozygotic (0.74) than dizygotic twin pairs (0.32), suggesting a heritability of BMI between 50 and 90%. Moreover, the BMI correlation was higher between biological parent-offspring pairs (0.19) than adoptive pairs (0.06) confirming the lesser impact of cultural transmission. However, the recent surge in cases of obesity worldwide points towards an important role of the environment, with the increase of sedentarity and availability of ultra-processed foods which are highly caloric and nutrient poor [96]. The overall picture of obesity is then clear: genetics determine the susceptibility to obesity while the environment reveals the phenotype.

2.3.4.1 The initial discoveries: leptin and the monogenic variants

The central hormone in obesity is leptin. The level of this hormone follows the proportion of body fat and stimulates or reduces food intake as needed. This strict mechanism allows for a precise regulation of body weight. Leptin is produced by the white adipose tissue to act on nerve cells, more precisely on the hypothalamus. Obese individuals are often deficient or resistant to leptin [97]. In humans, leptin is produced by the *LEP* gene, located on chromosome 7, while the leptin receptor, *LEPR*, is located on chromosome 1.

Several monogenic variants found in the leptin pathway were associated with extreme obesity in mice and humans. (i) Mutations in the pro-opiomelanocortin (*POMC*) gene are associated with obesity. *POMC* is the precursor of α-melanocyte-stimulating hormone (*α-MSH*), which is a leptin target and acts to decrease food intake. (ii) Mutations in an MSH receptor, the melanocortin 4 receptor (*MC4R*), leads to leptin resistance and is responsible for ~5% of the cases of extreme obesity. *MC4R* is expressed in the brain and is critical in maintaining body weight balance. The melanocortin system can be activated to reduce food intake and promote energy expenditure, through the activation of the Pomc neurons by leptin, insulin, or serotonin [98]. (iii) Finally, mutations located on the leptin receptor, *LEPR,* have been associated with obesity.

The mutations directly affecting leptin are rare, however 10-15% of morbid obesity are due to a gene defect in the neural circuit on which leptin acts. Leptin and the monogenic forms of obesity revealed the importance of the central nervous system in this disease, which was later confirmed by the results obtained from GWAS [99].

2.3.4.2 Obesity: a polygenic trait

As mentioned above, obesity was initially thought to be caused by leptin dysfunction. The identification of monogenic rare variants leading to extreme obesity reinforced this idea. This changed when GWAS were performed on obesity. It has nowadays become apparent that many variants, impacting

hundreds of supposedly unrelated genes, are associated with BMI and the picture became much more complex.

As of 2021, the Online Mendelian Inheritance in Man (OMIM) database [62] is linking 11 genes to obesity (OMIM: 601665) and 9 genes to 'body mass index quantitative trait locus' (BMIQs) (OMIMs: 602025, 607447, 607514, 612362, 612460, 614411, 615457, 617885, and 618406) (see Table 3). As for diabetes (see 2.2.3), an enrichment analysis was performed on the obesity genes against all the *Homo sapiens* genes in the PANTHER database [78], with the annotation derived from the Gene Ontology (GO) database (Released 2021-07-02) [79]. Non exhaustively, the overrepresented GO terms in obesity are related to feeding: *adult feeding behavior, regulation of appetite*, the nervous system: *neuropeptide signaling pathway, regulation of transmission of nerve impulse*, the immune system: *regulation of glucocorticoid secretion, negative regulation of interleukin-1 beta production*, temperature regulation: *temperature homeostasis*, *response to cold*, and fat cell differentiation (see 2.3.2.1): *white fat cell differentiation, regulation of brown fat cell differentiation.* Moreover, certain terms are like those obtained with diabetes type 2, possibly hinting at parallel mechanisms between the two diseases: *response to insulin, positive regulation of MAPK cascade* and *circadian rhythm.* The complete list of enriched terms is available in Table A.1-2.

Table 3: List of genes associated with obesity, according to the Online Mendelian Inheritance in Man database.

Trait	Hugo symbol
Obesity	*ADRB2, ADRB3, AGRP, CARTPT, ENPP1, GHRL, NR0B2, POMC, PPARG, SDC3, UCP3*
BMIQs	*ADCY3, AQP7, FFAR4, FTO, MC3R, MC4R, MRAP2, PCSK1, UCP2*

The interpretation of GWAS variants is sometimes complex. For example, Classnitzer et al. demonstrated that a variant identified on the *FTO* gene, *rs1421085 T-to-C*, was triggering an over-expression of two distal genes: *IRX3* and *IRX5* [100]. This over-expression favours the transformation of pre-

adipocytes to white, lipid-storing adipocytes instead of beige, energy-dissipating adipocytes.

This polygenic aspect of obesity is even more intricate when considering gene-environment interactions. For example, a meta-analysis performed by Kilpeläinen et al. found that physical activity reduced the impact of FTO variants on obesity risk [101]. Garaulet et al. found evidence that two SNPs from the PER2 gene (*rs2304672C>G* and *rs4663302C>T*), related to the circadian clock, were associated with snacking, diet-induced stress and bored-eating, hindering attempts to lose weight for the carriers [25]. While certain interactions, e.g., the diet and physical activity, are straightforward, there are more intricate factors at play, such as smoking or sleep disorders [96].

2.3.5 A 21st century epidemic

As with diabetes mellitus, the number of people with obesity has dramatically increased over the last century, as highlighted in Figure 3. In 2017, the WHO, in a study done in collaboration with the Imperial College London, estimated that obesity has tripled since 1975 [102]. In 2003, the WHO declared that obesity reached epidemic proportion, with more than 1 billion overweight adults, in 2016 more than 1.9 billion adults were overweight, including 650 million who were obese.

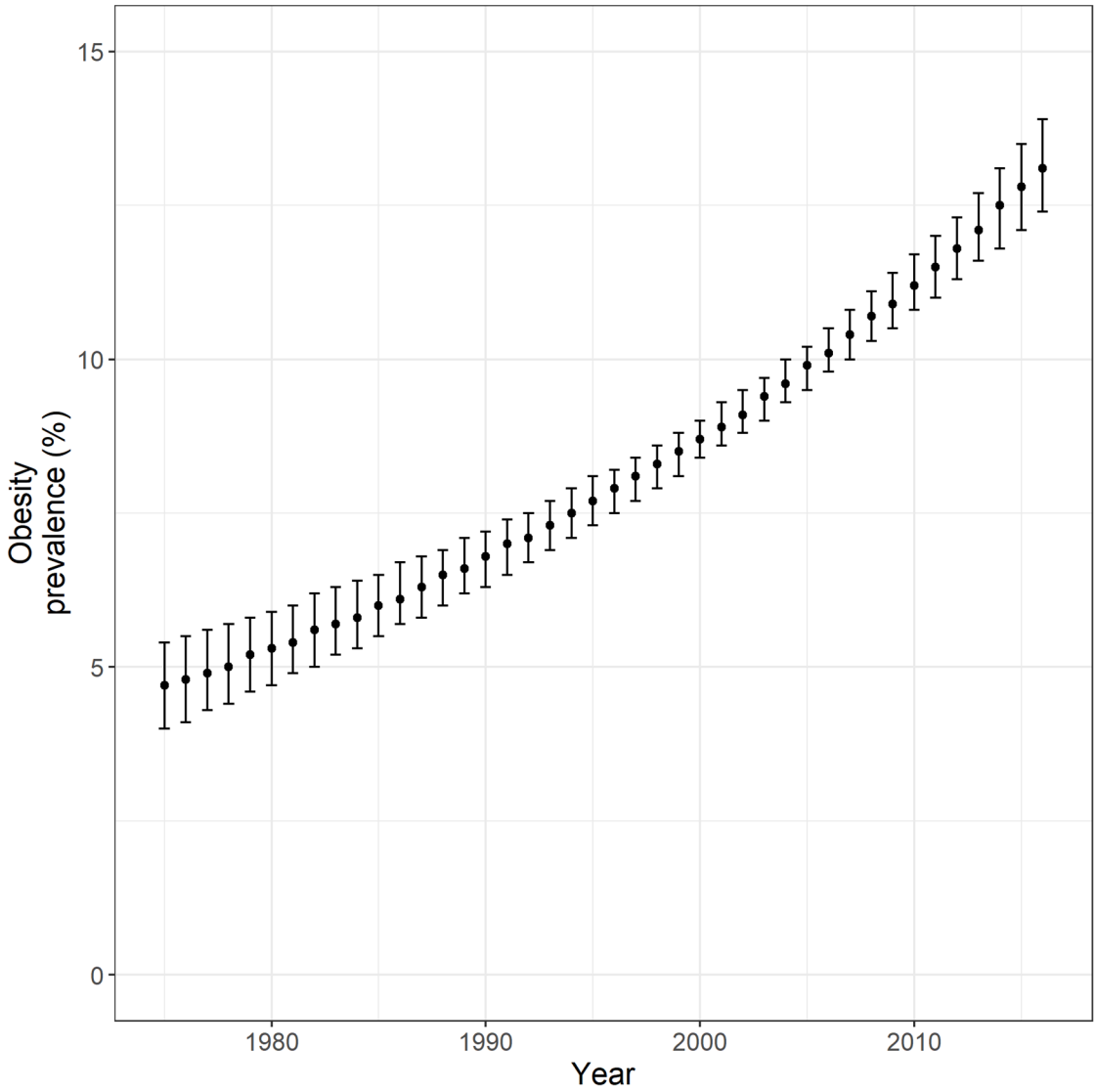


***Figure 3: Prevalence of obesity (BMI >= 30) in the world between 1975 and 2016. The bars represent the 95% credible interval. Data obtained from the World Health Organisation website, based on the study by Abarca-Gómez et al.* [102].**

2.4 Personalised Nutrition

2.4.1 Precision medicine

The increasing interest in personalised, or precision, medicine became apparent in the last decade. The ‘one size fits all’ approach for treatment is reaching its limits, with the highest-grossing drugs in the United States being efficient for only 1 in 4 people at best, but this number can be as low as 1 in 25 or even 1 in 50 [103]. Personalised medicine consists of taking individual variability into account when designing prevention strategies or considering treatments. A common approach is to divide a disease into relevant subgroups, called endotypes, representing distinct mechanisms, which might benefit from different therapeutic solutions. This technique was successfully applied to a range of disorders, an example of this was already mentioned previously here, with the identification of five endotypes of diabetes by Ahlqvist et al. [66] (see 2.2.2). Historically, cancer treatment has benefited from personalisation, and recent advances in the molecular understanding of tumour heterogeneity allowed to shift from an organ-centric view, to a vision based on genetic variants and molecular alterations [104].

The main hurdle that personalised medicine faces is the ‘curse of dimensionality’, when the amount of data collected is much higher than the number of analysed samples. This makes the separation of noise from true signal a difficult task, especially when looking at specific or weak signals. This can be resolved by removing redundant variables, adding more samples to the analysis, or by using dimensionality reduction methods (e.g., Principal Component Analysis or variable importance selection). Recent biobank projects, such as the UK Biobank [105], involve hundreds of thousands of participants, which will allow discoveries to be made on a scale that was never seen before.

An optimal precision medicine strategy will adapt to the uniqueness of each individual, and some scientists are arguing for the implementation of ‘one person trials’ [103]. However, we will need to be conscious of statistical pitfalls, more specifically, clinical trials will need to consider intra-individual variability, for example by assessing if the same patient responds favourably more than once

to the same treatment [106]. This will be especially important when considering complex phenotypes with no clear-cut pathophysiology.

2.4.2 The current state of personalised nutrition

The ideal of personalised nutrition is to consider the environmental and personal characteristics of an individual, to provide dynamic nutritional advice throughout life. If implemented correctly, it has the potential to reduce the risk of developing metabolic disorders or improving their management. This contrasts with the ‘one size fits all’ strategy, which is found in most current public health recommendations.

There are two main challenges facing personalised nutrition today. First, most nutritional studies lack the proper duration or number of participants to infer statistically relevant effects of the diet on the metabolism. Second, energy intake is very difficult to measure precisely [107]. Most studies rely on food frequency questionnaires, which are often time consuming and become a burden for the patient. Moreover, self-reported questionnaires are flawed and under-estimate the actual energy intake [108].

To solve these limitations, much shorter food frequency questionnaires, based on adherence to the Mediterranean diet, were developed. The Mediterranean diet was chosen for its positive association with cardiovascular health [109]. Such a questionnaire, the Mediterranean Diet Adherence Screener (MEDAS), which consists of 14 questions (see Table A.2-1), was used successfully in several studies. A higher score on this test was associated with lower coronary artery disease risk, BMI and waist circumference [110]. Another questionnaire, MEDLIFE, includes 28 items, distributed among three blocks, fifteen items are about food consumption, seven about traditional Mediterranean dietary habits and six about physical activity and social interaction habits [111]. This index is the first to include physical activity, which is important when considering the association between the diet and lifestyle.

Some interesting findings, showing the potential of personalised nutrition to improve the health of the general population, have been gathered from recent

studies. In this literature review, we will focus on the PREDIMED [112], Food4Me [113] and Zeevi et al. [114] studies, as they each use different approaches to the problem:

The PREDIMED (Prevención with Dieta Mediterránea) study measured the impact of the adherence to the Mediterranean diet on body mass index, waist circumference and waist-to-height ratio, in 7,447 participants with high cardiovascular risk [112]. The study assessed the Mediterranean diet score via the MEDAS questionnaire described above. The most important factors associated with lower abdominal obesity, were a high intake of nuts and low intake of sweet beverages. This study validated the approach of using MEDAS instead of long, burdensome, food frequency questionnaires, to measure adherence to the Mediterranean diet.

The Food4Me study was the first to implement an internet-based approach to enrol participants and assess the effect of personalised advice on their diet [113]. The 1,269 participants, who fully completed the randomized controlled trial, were split into four levels of personalisation (i) level 0, giving standard, non-personalised dietary advice (ii) level 1, giving personalised advice based on the participant's baseline diet (iii) level 2, giving personalised advice based on the baseline diet and phenotypic data (blood biomarkers and anthropometrics) (iv) level 3, giving advice based on the baseline diet, phenotypic data and genotypic data (consisting of five diet-responsive genes). The personalised diet was generated through a series of decision trees developed specifically for this study. After 6-months, the quality of the diet of the control arm was compared to the personalised arms (levels 1-3) via food frequency questionnaires. The individuals who got personalised nutrition advice consumed less red meat, salt, and saturated fat, while consuming more folate, resulting in a healthier diet. The Food4Me study proved that providing internet-based personalised advice is an effective approach to improve the diet in the general population. This study did not find that adding the phenotype and genotype to the personalisation provided added value, but one may argue that considering only 5 diet-related loci is not enough to measure a significant impact. Moreover, the study was using remotely

collected biological samples, which might lead to measurement errors, and self-reported questionnaires, which, as mentioned previously, tend to be unreliable [108].

Zeevi et al. [114] built a model to predict glycemic responses in a cohort of 800 participants and to adapt their diet accordingly. This biomarker was chosen because elevated blood glucose levels are a risk factor for type 2 diabetes. The study lasted for one week and blood glucose levels were measured continuously via a glucose monitor using subcutaneous sensors. The model was based on the gradient boosting regression algorithm [115], merging predictions from thousands of decision trees. The input included lifestyle reports, e.g., food intake, exercise, and sleep, obtained via a smartphone-adjusted website, anthropometrics, medical background, and microbiota profiling via 16S rRNA sequencing. The model found 21 beneficial, 28 non-beneficial and 23 non-decisive microbiome features (e.g., growth of *Eubacterium rectale* was found to be beneficial, as it is associated with lower post-meal glucose levels). This study showed the interest of using a specific biomarker as a target, as well as the importance of the microbiome and new technologies, such as machine learning, in personalised nutrition. Indeed, the model provided advice in the intervention arm of the study, which successfully lowered post-meal glucose levels. It is to be noted that some aspects of the study have been criticized in the literature, notably the statement by Zeevi et al. about the great inter-individual variability in glycemic response to the same meal, Wolever argues that the results obtained are better explained by intra-individual variability [116].

These studies highlight how versatile personalised nutrition is in its implementation, be it via short questionnaires, internet-based studies or aiming at regulating specific biomarkers. They also demonstrate the potential of personalised nutrition to provide healthier alternative adapted to each individual and curb the epidemics of non-communicable diseases, notably type 2 diabetes, and obesity.

2.4.3 Factors needed to personalise the diet

Personalised nutrition can be as simple as stratifying the advice given to relevant subsets of individuals, or as complex as integrating biological, social and lifestyle data to give tailored advice to each individual.

Anthropometrics and socio-demographics data are important, as they will help both in assessing the risk an individual has of developing obesity or diabetes, as well as framing to what extent the individual will be able to follow the dietary advice given to them. The same reasoning applies to lifestyle data, current diet and physical activity need to be taken into account when devising a new diet.

Recent developments in biology are promising for the future of personalised medicine and nutrition. Genetics has always been a staple of personalisation, as our genotype is the blueprint of our phenotype. And recent discoveries, notably gained through GWAS, on the impact of variants on diet, obesity and diabetes is and will be invaluable for the future of personalised nutrition. But other players are revealing their potential. First, the multi-omics revolution is proving how merging data from different -omics technologies (genomics, proteomics, and metabolomics) can help us to make sense of complex systems and mechanisms as well as being better at monitoring dietary intake than questionnaires. Second, it is now clear that including the microbiome in the personalised nutrition equation is indispensable. For example, the microbiome of obese individuals tends to be less diverse, less complex and containing different groups than healthy individuals. A study in mice found that the composition and diversity of the microbiota, in interaction with the diet, was affecting metabolism and was a risk factor for the development of obesity [117].

As mentioned in the previous Sections, obesity and diabetes are complex diseases, which imply complex phenotypes. The precise exploration of their pathophysiology, including measurement of biomarkers, characterisation of subtypes and of individual variation between affected individuals, will be a necessary step for improving personalised nutrition. Understanding the differences between subtypes and their mechanisms, will be key to design diets

aiming at mitigating their impact. In other words, disease stratification will be needed for diet stratification.

Recent studies also highlighted the role of timing in nutrition. The circadian clock is an internal, biological clock, which responds to external time cues, such as light, to maintain endocrine and metabolic pathways. The interplay between circadian biology, the impact of nutrition and the microbiota has gained interest in the past decade [118]. Among other processes, the circadian clock is regulating lipid homeostasis, and mouse mutants with disrupted circadian rhythm were more at risk of becoming obese [119]. Moreover, the enrichment analyses performed in 2.2.3 and 2.3.4.2 for both diabetes and obesity found an over-representation of the *circadian rhythm* GO term. Thus, proper timing of food intake could be another interesting facet of personalised nutrition.

The ethical implications of personalised nutrition need to be considered before being translated to clinical practice and advice is given to the general population. Indeed, personalised nutrition deals with genomic, social, and personal information, plus, it is important to assure that the correct diet is given to the correct person. The position of the *International Society of Nutrigenetics / Nutrigenomics on Personalized Nutrition* on this matter was released in 2016 [120]. Emphasis was given on the importance of obtaining informed consent, protecting the privacy of genetic information, and using validated genetic knowledge. The society also mentioned the need to frame the legal regulations surrounding personalised nutrition.

2.5 Aims and objectives

The aim of this thesis was to explore the impact of the genotype on diabetes and obesity, with the overall purpose of allowing diet personalisation. Indeed, the genotype is one of the most important determinants of our phenotype, therefore every attempt at making personalised nutrition must account for it.

2.5.1 Generating lists of variants

First, comprehensive databases of Single nucleotide Polymorphisms (SNPs) related to diabetes and obesity were gathered. Instead of producing them manually when needed, this step was automated by developing *VarGen*, an R package that can integrate data from different sources to find variant-trait relationships. The main workflow from VarGen both retrieves variants related to a trait and has the potential to find new variant-trait associations. An alternative pipeline, more specific, called *VarPhen*, retrieves variants linked to a list of phenotypes given by the user. Moreover, to help the user to estimate the importance of each variant, an annotation function was added to the package.

VarGen was benchmarked against two similar tools, DisGeNET [121] and VarFromPDB [122]. Obesity and Alzheimer's disease were chosen as use-cases for the benchmark, and the relevance of the variants retrieved with VarGen was assessed by comparing the overlap of the results between each tool.

To understand the impact of the variants gathered previously on the biology of obesity and diabetes, a pathway analysis was performed using the outputs from VarGen and VarPhen. The tool *Pascal* generated gene and pathway scores for each list of variants (i.e., for obesity, diabetes type 1 and diabetes type 2). Then, we focused on the top 15 pathways for each list to explore in detail the biological processes that were affected by the variants.

2.5.2 A new method to refine Polygenic Risk Score models

Obesity and diabetes are due to the cumulative effect of a multitude of variants, with varying impact, therefore estimating an individual's overall genetic risk is not straightforward. One possible approach is to use Polygenic Risk Score (PRS)

models, which summarize, as a single value, the genetic risk cumulated by a set of variants.

PRS models are often based on variants obtained from GWAS analyses and applied to a target population. While being effective, PRS models suffer from the fact that variants obtained through GWAS are often common resulting in a potentially biased risk estimation. Indeed, carriers of rarer variants, with a higher impact on the phenotype, actually have a higher or lower risk than assessed by the PRS model. Here, we present a new method to refine the estimations obtained from a PRS model, which uses variants obtained from VarGen. Based on the variants from VarGen, we can produce a second PRS and detect subsets of individuals with a high or low risk for this other set of variants, thus revaluating their overall genetic risk.

This method was tested and validated on Body Mass Index (BMI) and diabetes. The UK Biobank, a biomedical database containing anthropomorphic and genetic information for ~500,000 individuals from the United Kingdom, was used as the target set of the PRS models. Knowing the information about BMI and diabetes for the participants of UK-Biobank allowed us to assess the accuracy of the genetic risk prediction of the refined PRS models. In future works, genetic risk scores will be used, in combination with other clinical factors, within Nutrishield's platform for personalised nutrition.

3 VarGen: an R package to discover and annotate variants associated to a disease

The work presented in this Chapter has been published in *Oxford Bioinformatics* [123] and is available on GitHub as an open-source project under the MIT license https://github.com/MCorentin/VarGen.

3.1 Background and motivation

As described in Chapter 2, identifying the genetic component of complex diseases is crucial to understand, prevent, and treat them. This is becoming increasingly important for diabetes and obesity as they are becoming epidemics. Moreover, personalised medicine is showing promises towards improved prevention and care. The identification of disease risk is not trivial, as complex diseases often involve multiple factors, including the environment, the microbiome, lifestyle, and genetics. Concerning genetics, as mentioned in Section 0, recent findings are suggesting that, taking into account the accumulation of many small impact variants rather than focusing on monogenic, high impact variants may provide a more precise and generalisable risk assessment in the general population.

With the recent advances in DNA sequencing (see Section 2.1.2), it is now possible to produce genetic data relatively quickly and affordably. This led to a surge of publicly available, high-quality, genotyping information in the past ten years. For example, The National Center for Biotechnology Information (NCBI) genetic variants archive, dbSNP [10], went from 18 to 660 million human variants between build 130 (in 2009) and build 151 (in 2017). Unfortunately, the information is often spread between different resources, hence there is a need for tools to aggregate the content of these databases and facilitate the exploration of variant-disease relationships.

This is the incentive behind the development of VarGen, an easy-to-use R package designed to fetch and score variants linked to a disease using several public databases.

3.2 Vargen Workflows

VarGen has two main workflows, one based on sensitivity (VarGen) and the other on specificity (VarPhen). Alternatively, it is possible to run the analysis based on a specific list of genes.

3.2.1 VarGen

3.2.1.1 A pipeline for variant discovery

The VarGen pipeline can be launched from the function: `vargen_pipeline` and returns a list of variants linked to a given disease.

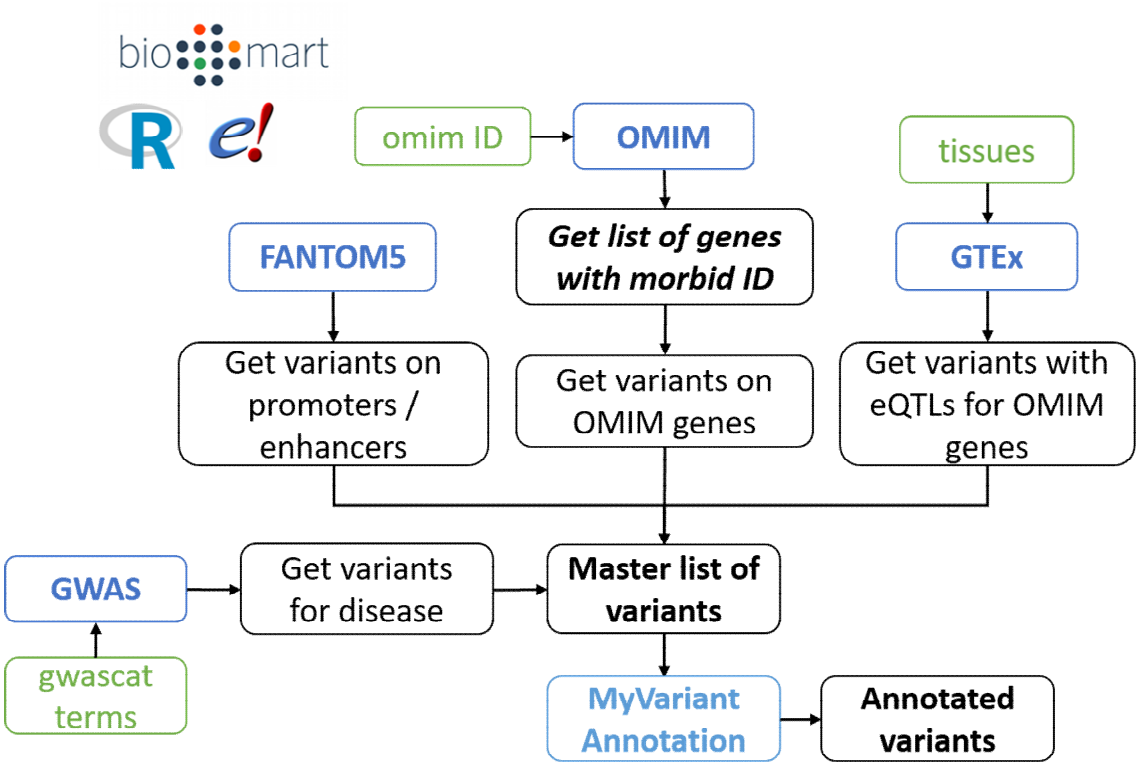


Figure 4: VarGen workflow, user input is represented in green and databases in blue. The pipeline is centred on the list of genes obtained from OMIM. VarGen gets the variants located directly on those genes, as well as on their enhancers and promoters.

The pipeline, described in Figure 4, typically starts from one or more disease identifiers from the Online Mendelian Inheritance in Man (OMIM), entered by the user. First, VarGen gets the list of genes associated with these identifiers, subsequently called the 'OMIM genes', and returns all the variants located directly on them. The next step is to get the variants located on the promoter regions of

the OMIM genes. VarGen retrieves this information from the Functional Annotation Of Mammalian Genomes 5 (FANTOM5) database. It is expected that variants on the promoters can affect gene expression if the mutation affects the binding site of activators or repressors. There is increasing evidence of the importance of non-coding variants on complex traits and diseases [22]. If the user provided one or more tissues of interest, VarGen will use the Genotype Tissue Expression (GTEx) database to get the variants affecting the expression of the OMIM genes in these tissues. Finally, the user can provide one or more GWAS traits and VarGen will query the GWAS Catalog to get the list of variants associated with each trait.

This pipeline was designed as a discovery analysis, with the potential to identify new variants related to the disease. Consequently, all variants returned by VarGen do not necessarily have an impact on the disease of interest. To estimate the importance of each variant, an annotation function was added to VarGen: `annotate_variants`. This function takes the rsID of the variants as input and sends requests to the 'MyVariant.info' API to retrieve the annotation (see Section 3.2.1.2). This function was developed with the help of Matthew Brember as part of his MSc thesis project at Cranfield University (2019).

All the positions are referring to the GRCh38 version of the human assembly. As FANTOM5 and GTEx v7 are based on GRCh37, VarGen automatically lift-over the positions obtained from these databases to GRCh38.

3.2.1.2 VarGen output

The main output from VarGen is a list of variants related to the disease. For each variant the following information is reported:

- **chr:** the chromosome on which the variant is located.
- **pos:** the variant position. For InDels, the starting position is reported.
- **rsid:** the variant identifier.
- **ensembl_gene_id:** the ensembl gene identifier related to the variant.
- **hgnc_symbol:** the Hugo symbol for the gene related to the variant.

- **source:** the variant source in VarGen, can be *omim*, *fantom5*, *gtex* or *gwas*. In the case of *gtex* the tissue is also specified in parenthesis.
- **trait:** the trait(s) related to the variant. For the variants found with *omim*, *fantom5* or *gtex* this corresponds to the omim identifier. For the *gwas* variants, this corresponds to the GWAS trait.

The annotation with MyVariant.info will add the following columns to the output:

- **ref:** the allele in the human reference (GRCh38)
- **alt:** the alternative allele. In the case of multi-allelic variants, VarGen creates one line per alternative allele and annotates them separately.
- **CADD phred score:** ranging from 1 to 99, based on the rank of each variant relative to all possible 8.6 billion substitutions in the human reference genome. A higher value means a more deleterious variant [124].
- **fathmm-xf score:** between 0 and 1, a higher value means a more deleterious variant, with more confidence the closer to 0 or 1 [125].
- **fathmm-xf prediction**: can be 'D' for Damaging if the fathmm-xf score is higher than 0.5, or else 'N' for Neutral.
- **Annotation type:** provides information about the variant context (e.g., coding, non-coding, regulatory region).
- **Consequence:** provides information about the variant impact (e.g., regulatory, downstream, stop_gained).
- **ClinVar clinical significance:** reports the clinical significance of the variant from ClinVar (e.g., benign, pathogenic) [126].
- **SnpEff impact:** high-level assessment of the variant putative impact (high, moderate, modifier or low) [20].

Example of outputs are available in Table B.1-3 and Table B.1-4.

VarGen provides different ways to visualise the results. The GWAS variants can be represented in a Manhattan plot, with the function `plot_manhattan_gwas`. The input is a list of GWAS traits and chromosomes to plot. An example of Manhattan plot, for traits related to Alzheimer's disease, is presented in Figure 5, note: the y-axis does not represent a stronger SNP effect on the trait but rather a

stronger association between the SNP and the trait. Manhattan plots for obesity, diabetes mellitus type 1 and 2 are presented in Figure B.1-1, Figure B.1-2 and Figure B.1-3 respectively.

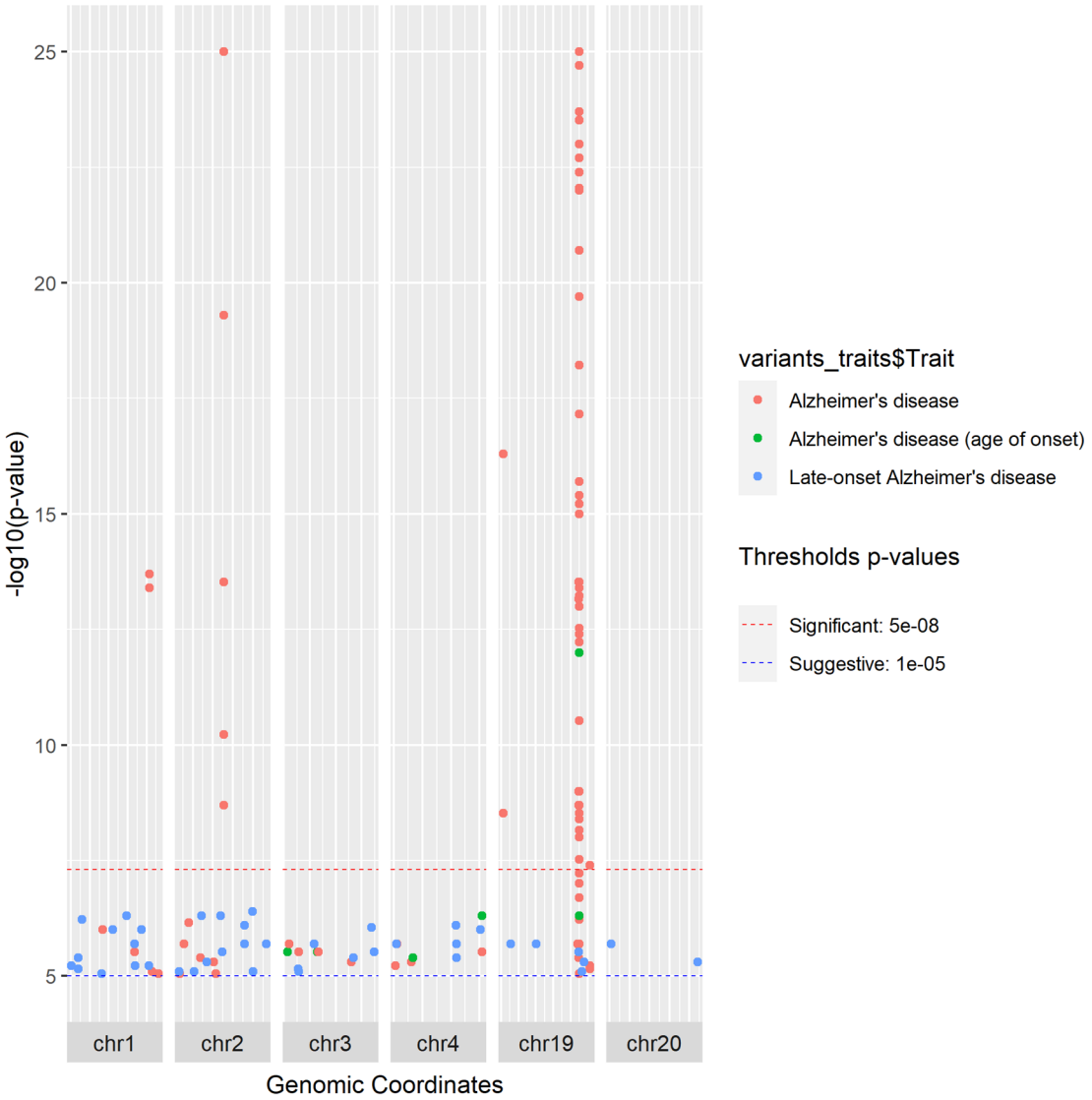


Figure 5: Manhattan plot produced with the 'plot_manhattan_gwas' function of VarGen. Each dot is a variant, coloured by its corresponding GWAS trait. The x-axis represents the genomic coordinates, split by chromosome, here only 6 chromosomes are represented for the sake of clarity. The y-axis represents the –log10(p-value), a higher value means a more significant relation between the variant and the trait. The two thresholds 'Significant' and 'Suggestive' are described in Section 2.1.3.5. There is an interesting locus on chromosome 19, containing many SNPs associated with Alzheimer's disease.

A customised visualisation was developed, with the help of Matthew Brember as part of his MSc thesis project at Cranfield University (2019), to represent

VarGen's output, such a figure can be created by running the `vargen_visualisation` function. It takes a list of annotated variants from VarGen as input and creates one plot per gene present in the list. An example, for the *SIM1* gene (*ENSG00000112246*), which is related to obesity, is shown in Figure 6.

Figure 6: Example of custom visualisation created with the vargen_visualisation function from VarGen. This plot gives information about the variants found by VarGen on the SIM1 gene (ENSG00000112246). At the top, the chromosome is represented (here chromosome 6) with a red bar pinpointing the gene location. Just below the chromosome there is an axis indicating the genomic position (here from 100.38 to 100.46 Mb). The three tracks below are relative to this axis. The first track from the top contains the five different transcripts of this gene (three are on the last line), with the coding parts drawn in purple. The second track has the variants found in this gene, as green bars, grouped by consequence. The last track contains the same variants, as blue and red dots, with the CADD score as the y-axis. The red dots correspond to a list of rsIDs given by the user.

3.2.2 Alternative pipelines

3.2.2.1 VarPhen: a more specific pipeline

As mentioned before, VarGen was designed to be a discovery pipeline. This can lead to very large outputs, which the user needs to filter manually. In order to give more flexibility to the package, a more specific pipeline was created, called

VarPhen. This alternative pipeline only outputs variants that are confirmed to be related to the phenotypes of interest. The input is a list of phenotypes, which can be obtained by giving a list of keywords to the `get_phenotype_terms` function. This function queries BioMart (see 3.3.5) and attempts to match any of the keywords given by the user to the list of phenotypes available under the "phenotype_description" filter in the "*Ensembl Variation*" mart, using grep. It means that any phenotype containing one of the keywords as part of their name will be returned by the function. The user can then input these phenotypes, or a subset of them, into the `get_variants_from_phenotypes` function, which will query BiomaRt using the same filter "phenotype_description", but this time, returning a list of variants related to these phenotypes. The phenotype-variant relationships are retrieved by BioMart using the following sources:

- COSMIC (Catalogue Of Somatic Mutations In Cancer)
- ClinVar (Variants of clinical significance from ClinVar)
- dbGaP (The database of Genotypes and Phenotypes)
- EGA (European Genome-phenome Archive)
- GIANT (Genetic Investigation of ANthropometric Traits)
- HGMD-Public (The Human Gene Mutation Database)
- MAGIC (Meta-Analyses of Glucose and Insulin-related traits Consortium)
- NHGRI-EBI GWAS Catalog

A flowchart presenting an overview of VarPhen is available as Figure 7. As with VarGen, the list of variants obtained can be annotated with the `annotate_variants` function.

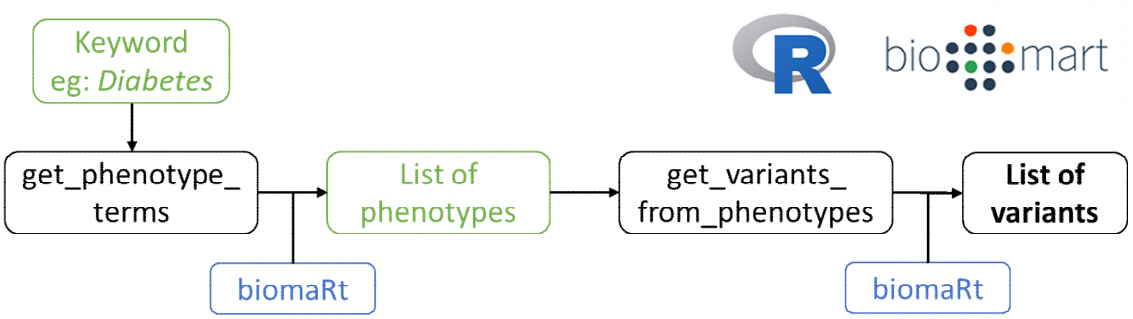


Figure 7: Flowchart of the VarPhen pipeline. The user can enter one or more keywords (e.g.: diabetes) to find phenotype terms and their associated variants.

3.2.2.2 A custom list of genes

As an alternative to the main pipeline, the user can run `vargen_custom`. This function follows the same steps as `vargen_pipeline` (see Section 3.2.1.1) but directly accepts a list of Ensembl gene identifiers (e.g., *ENSG00000197594*) instead of OMIM disease identifiers. This function is very similar to `vargen_pipeline`, it gets the variants directly located on the genes, then the variants on the promoters from FANTOM5. There is also the possibility to enter tissues for GTEx and GWAS terms. This pipeline is very useful if the user wants to focus on specific genes.

3.2.2.3 GWAS variants only

If the user is only interested in GWAS variants, it is possible to run the GWAS step independently. First, the user can search for a list of terms of interest with the `list_gwas_traits` function. Then the `get_gwas_variants` will retrieve the variants linked to the terms in the GWAS Catalog.

3.3 List of ressources accessed by VarGen

VarGen retrieves variants linked to phenotypes by integrating information from public databases (see Section 3.2.1.1). This Section will describe the content and purpose of each database.

3.3.1 The Online Mendelian Inheritance in Man database

The Online Mendelian Inheritance in Man (OMIM) is a database of human genetic disorders [73], with a focus on the link between genes and phenotypes. It has been continuously updated since 1966 and is one of the most consulted resources for genetic disorders, both by clinicians and researchers. As of October 2021, it contains 16,588 genes and 6,205 phenotypes with known molecular basis (data from https://omim.org/statistics/entry). The gene-to-disease relationship is obtained by manual curation of the peer-reviewed literature. This resource started as a series of twelve catalogues of mendelian traits and disorders, published between 1966 and 1998. The online version has been available since 1987. It is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine.

VarGen accesses OMIM to get the list of genes associated with a certain phenotype. This is an important step as VarGen's workflow is centred around the genes found with OMIM.

3.3.2 The Genotype Tissue Expression database

The Genotype Tissue Expression (GTEx) project started in 2013 [127]. This was an effort to build a data resource and tissue bank to assess the impact of variants on gene expression in different tissues. RNA from 948 donors were isolated post-mortem from 54 non-diseased tissues, the list of tissues is available as Table B.1-2.

The criteria for inclusion are listed below:

- 21 <= Age (years) <= 70
- 18.5 < BMI < 35
- Less than 24 hours between death and tissue collection
- No whole blood transfusion within 48 hours prior to death
- No history of metastatic cancer
- No chemotherapy or radiation therapy within 2 years prior to death
- Generally unselected for presence or absence of diseases or disorders

GTEx collected a wide range of data: whole genome genotyping, exome genotyping, RNA sequencing, whole exome sequencing, and whole genome sequencing [128]. The data are available through different portals: the National Center for Biotechnology Information (NCBI), the GTEx data portal and the database of Genotype and Phenotype [129].

One of GTEx aims was to study expression Quantitative Trait Loci (eQTLs) in different tissues. One can define eQTLs as the analysis of the impact of variants on gene expression [130]. Because of this, GTEx is complementary of GWAS, since most of the variants discovered thus far are not found in protein coding regions, meaning they probably have an impact on gene regulation [128]. Combining the results from GWAS and GTEx will help the research community to make sense of the mechanisms altered by the disease-related variants. It has been estimated that eQTLs play an important role in complex traits and diseases [131].

The GTEx project was completed in 2020. This was accompanied by a study providing insights about the impact of variants on gene expression (eQTLs) and splicing (sQTLs) in 838 individuals over 49 tissues [23]. This study found that 94.7% and 66.5% of the protein-coding genes were regulated by eQTLs and sQTLs respectively.

VarGen harnesses the data from GTEx to get the eQTLs that are affecting the expression of the genes found with OMIM. The user can choose which tissues should be taken into account.

3.3.3 The Functional Annotation Of Mammalian Genomes 5

The Functional Annotation Of Mammalian Genomes 5 (FANTOM5) was a project investigating the transcription regulation activities in mammalian cells [132]. The consortium studied cellular functions by quantification of RNA molecules in a range of different cells, mostly primary cells but also cell lines and tissues (consisting of multiple cell types). More than a thousand human and mouse samples were analysed. FANTOM5 used a variation of the *Cap Analysis of Gene Expression* protocol, using single molecule sequencer, to study regulation. This allowed the quantification of Transcription Start Sites (TSS) at a single base resolution. The consortium developed a central repository of tools and databases to access the results generated from the project, including SSTAR for data exploration, ZENBU for visualisation of the results on a genome browser and the FANTOM Five ontology.

VarGen uses the results from FANTOM5 to get the locations of the enhancers of the genes obtained from OMIM. Indeed, variants located in promoters might affect gene regulation and play a role in diseases [22].

3.3.4 The Genome Wide Association Study Catalog

The *NHGRI-EBI GWAS Catalog of human genome-wide association studies* is a central repository storing the results of GWAS [38]. It was founded in 2008 by the NHGRI due to the increasing number of published associations. The idea was to provide researchers with a catalogued and standardised access to this wealth of data. The GWAS Catalog is filled by manual curation of the literature, out of which is extracted information about each study. Namely, details about the publication, cohort (size, ancestry, and country of recruitment) and SNP-disease associations (rsID, p-value, gene, and risk allele). Finally, one or more traits are allocated to each study, depending on the studied phenotype. Detailed information about each SNP is obtained from the rsIDs, using the Ensembl Application Programming Interface (API).

A study can be added to the GWAS Catalog if it includes an array-based genotyping and analysis of more than 100,000 SNPs selected to tag variation

across the genome. Sequencing data imputed to genotyping arrays are eligible as well if they follow the same criteria. A study is excluded if it was not published in English, is limited to certain candidate genes, measures somatic variations, or does not include new GWAS data. Once a study has been declared eligible, the statistically significant SNP-trait associations (p-value < $1x10^{-5}$) are added to the GWAS Catalog.

As of June 2020, the GWAS Catalog contains 4,580 publications and 187,403 associations for 4,287 traits (source: https://www.ebi.ac.uk/gwas/home).

VarGen accesses this resource to get the variants associated to a list of traits entered by the user. This can be run independently of the rest of the pipeline if the user is only interested in GWAS variants.

3.3.5 BioMart: at the crossroad of biological data

BioMart [133] is an open-source interface to a wide range of biological databases. This allows researchers to perform searches based on different criteria and regroup the results using a single interface. This service can be accessed via different means: the web interface, the Perl API or from its integration within a larger project (Ensembl, UniProt, HapMap etc…). VarGen queries BioMart via the R package *biomaRt*, available through Bioconductor.

BioMart is hierarchically organised in Marts, datasets, and attributes/filters. A Mart is a relational database with a schema compliant with BioMart definitions. Each Mart consists of different datasets, often one per species. Each dataset contains different type of data. For example, the *btaurus_snp* dataset from the *Ensembl Variation* Mart contains information about variants for the *Bos taurus* species. A query can retrieve the whole dataset or just a subset based on **attributes** to select fields of interest (e.g., chromosome, position, etc) and **filters** to select values of interest based on one or more attribute (e.g., to only get data related to chromosome 1). Both attributes and filters are dependant on the selected Mart and dataset. If several Marts have shared attributes or filters, it is possible to merge them in a query. The structure of a BioMart query is described in Figure 8.

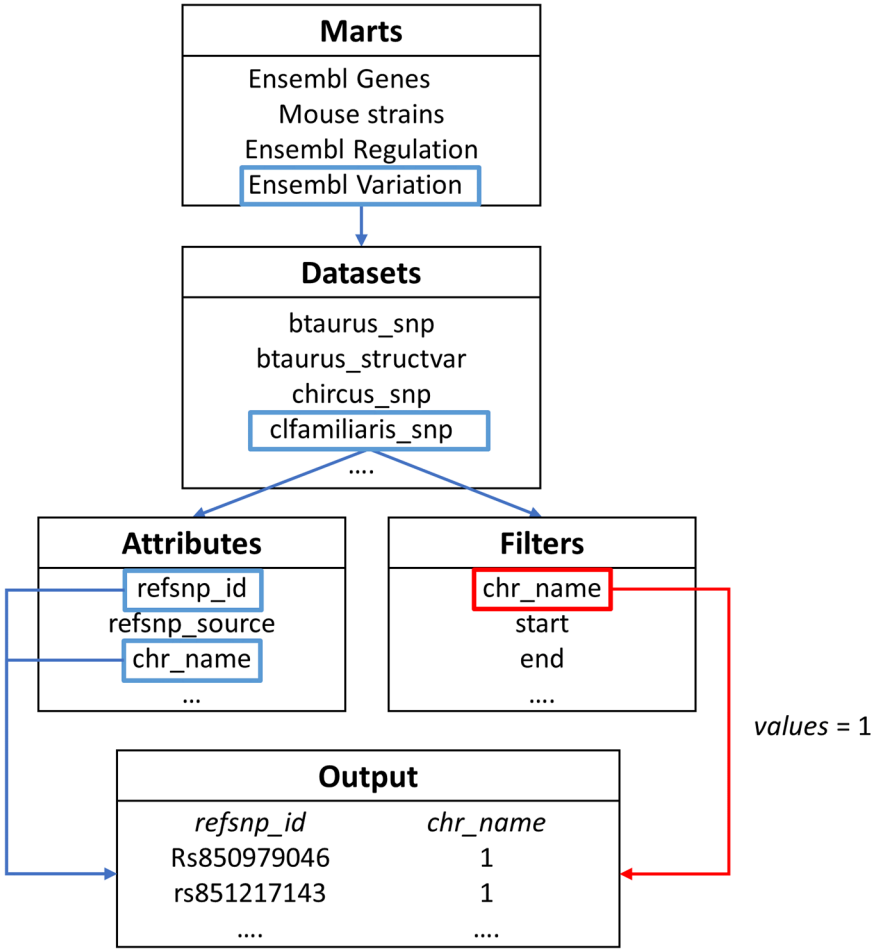


Figure 8: Structure of a BioMart query. The first step is to select a Mart, here Ensembl Variation. Each Mart has different datasets, usually each ensembl dataset correspond to a species, here the clfamiliaris_snp dataset was selected. Finally, the user chooses the data to display (Attributes) and the restrictions on the results (Filters), here the refsnp_id and chr_name will be displayed, for chromosome 1, due to the filter 'chr_name = 1'.

In VarGen, the two Marts of interest are *Ensembl Genes* which contains information about genes and *Ensembl Variation* which contains information about variants. For both Marts, VarGen accesses the *hsapiens_gene_ensembl* dataset, which corresponds to human data. However, the user does not need to be knowledgeable about Biomart as VarGen provides high-level functions to connect to and query these Marts, viz `connect_to_gene_ensembl` and `connect_to_snp_ensembl`.

3.3.6 MyVariant.info: an API for variant annotation

MyVariant.info is a REST API providing variant annotation as a service [18]. It aggregates ~1,500 annotation fields from 19 sources (see Table B.1-1). VarGen retrieves annotations from some of these sources, including Combined Annotation-Dependant Depletion (CADD) [124], FATHMM-XF [125], SnpEff [20] and ClinVar [126]. The annotation is time efficient as the information from *MyVariant.info* is retrieved from pre-annotated variants. More details about the fields retrieved by VarGen for the annotation is available in Section 3.2.1.2. To retrieve all these annotations, VarGen requires the R package *myvariant*, a wrapper developed to easily query the *MyVariant.info* services.

CADD implements a machine learning algorithm to classify the variants as neutral or deleterious, based on more than 60 features [124] (e.g., VEP consequence, SIFT score, variant type). Instead of relying on a small number of known pathogenic variants, CADD takes the opposite approach and trains its model on fixed 'neutral' variants in the human population since the split with the chimpanzee. Since these variants have an allele frequency of 95-100%, they are assumed to be neutral. Deleterious variants are based on simulated random variants not subjected to evolutionary pressure. This allows annotation for non-coding variants, which have important roles in understanding the genetic basis of certain disorders [22]. The raw CADD scores are then transformed into Phred scores, ranking all the possible variants in the human genome (see Equation 1). VarGen retrieves the Phred score from CADD, to assess the deleteriousness of each variant.

$$\boldsymbol{CADD\ phred\ =\ -10\ log_{10}(\frac{SNP\ rank}{\#\ possible\ SNPs})}$$

Equation 1: Calculation of the CADD Phred score for one SNP, dividing the rank of the SNP against all the other possible SNPs in the human genome. The ranks are based on the raw CADD score.

FATHMM-XF is the extended version of FATHMM-MKL, an algorithm designed to predict the functional consequence of coding and non-coding variants. Each prediction is given as a confidence score, with a low value indicating benign impact and a high value indicating deleteriousness. The scores are obtained from

a supervised machine learning prediction based on both pathogenic variants, from the Human Gene Mutation Database, and neutral variants, from the 1000 genome project. The machine learning algorithm, a Hidden Markov Model, is based on the following features: sequence conservation across species, proximity to genomic features, chromatin accessibility and nucleotide sequence.

SnpEff is an open-source tool capable of predicting the putative impact of variants. The prediction is based on their genomic locations, (e.g.: *intronic*, *splice site*, *untranslated region*) and information from ENSEMBL, UCSC and organism-specific databases. The output can be given in VCF or text format and contains information about the variant itself, the genetic information (e.g.: *gene name*, *transcript ID*) and the coding effect of the variant (e.g.: *synonymous, non-synonymous, frameshifts*). SnpEff also provide a high-level assessment of the variant impact, namely *LOW*, *MODIFIER*, *MODERATE* and *HIGH*.

ClinVar is an archive provided by the NCBI, which stores clinically important variants. Since the data are based on submissions by the community, there can be conflicting information for some of the variants, in which case, ClinVar reports every submitted values. To add some weight to each submission, ClinVar also records the underlying evidence that led to the assessment of the variant's impact. VarGen will extract the 'clinical significance' annotation, which follows the recommendation from the American College of Medical Genetics and Genomics [134]. The list of terms that can be associated to a variant are listed below:

- Affects
- Likely pathogenic
- Confers sensitivity
- Other
- Uncertain significance
- Association not found
- Protective
- Conflicting data from submitters
- Pathogenic
- Drug response
- Risk factor
- Association
- Likely benign
- Benign
- Not provided

3.3.7 VarGen access to resources

VarGen accesses the resources described previously via BioMart, Application Programming Interfaces (APIs) or local files. This subsection will describe how each resource is queried by VarGen.

The information from OMIM is retrieved with BioMart (see Section 3.3.5). The `get_omim_genes` function creates a query to get the genes associated with a certain OMIM identifier (the BioMart filter name is *mim_morbid_accession*). The function returns the ensembl gene id, locus, HGNC symbol and the OMIM description. The GTEx eQTLs are downloaded as local files, one per tissue, containing the significant variant-gene pairs. VarGen reads the files corresponding to the tissues given as input and retains the variants associated with the OMIM genes obtained from the previous step. The FANTOM5 information is obtained from a local file, the *enhancer_tss_associations.bed,* which contains the associations between the transcription start sites and the genes. The GWAS Catalog can either be downloaded or accessed online at the user discretion. Finally, the annotation is performed with the *MyVariant.info* API. A summary of the resources accessed by VarGen is available in Table 4. All the mandatory local files can be downloaded automatically via the `vargen_install` function.

Table 4: Description of the databases accessed by VarGen. For each database, the user input is described, as well as the data retrieved.

Database	Access	User input	Data retrieved
OMIM	Online via BiomaRt	OMIM identifiers	List of OMIM genes related to a disease
FANTOM5	Local file	Correlation threshold	Enhancers / promoters of the 'OMIM genes'
GTEx	Local files One per tissue	List of tissues	Variants leading to a change in the 'OMIM genes' expression
GWAS Catalog	Local or Online	List of GWAS traits	Variants associated with the GWAS traits given as input
MyVariant	Online via *myvariant*	List of rsIDs	Variant annotation

3.4 VarGen benchmarking

VarGen was benchmarked against two other similar tools, DisGeNET [121] and VarFromPDB [122]. Since the true set of all the variants linked to most diseases is not known, we must rely on the overlap between different tools to assess the relevance of the variants obtained with VarGen. Two traits were chosen as use cases, obesity (OMIM: 601665) and Alzheimer's disease (OMIM: 104300). The benchmarks were run with R version 3.6.3, disgenet2r v0.0.9 and VarfromPDB v2.2.10.2.

3.4.1 First use case: obesity

3.4.1.1 Methods

For VarGen, the following input was given to `vargen_pipeline`:

- The OMIM identifier *601665* (corresponding to 'OBESITY').
- The following GTEx tissues: *Adipose subcutaneous* and *Adipose visceral*.
- The following GWAS traits: *Obesity (extreme), Obesity-related traits, Obesity, Obesity (early onset extreme), Obesity and osteoporosis, Obesity in adult survivors of childhood cancer exposed to cranial radiation, Obesity in adult survivors of childhood cancer not exposed to cranial radiation, Type 2 diabetes (young onset) and obesity, Obesity without metabolic disease.* The data were extracted from the GWAS Catalog version e96.
- The *fantom_corr* parameter was set to 0.20.

For VarPhen, the keyword *obesity* was used to retrieve 39 phenotype terms having a link with obesity, they are listed in Table B.1-5. These phenotypes were then given as input in the `get_variants_from_phenotypes` function.

For DiGeNET, the disease identifier *C0028754*, corresponding to obesity, was given as input for the function `disease2variant`. It was run twice, one time with all the databases and the other time with the curated ones.

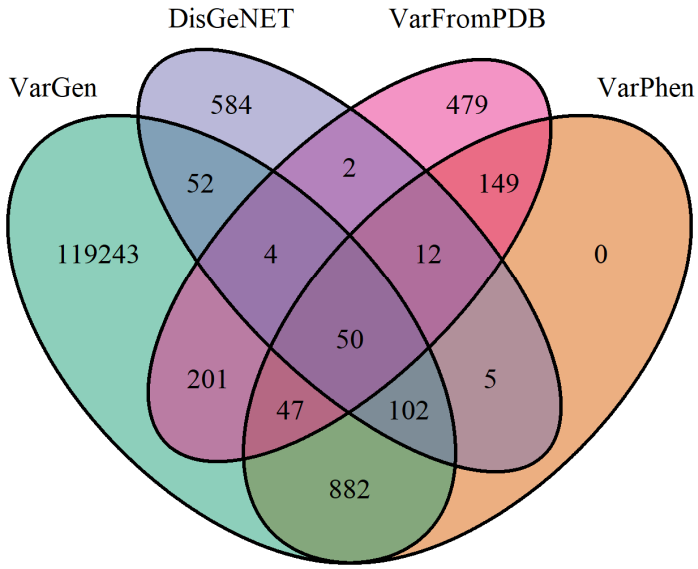
For VarFromPDB, the authors' guidelines were followed to obtain the results. *Obesity* was entered as the keyword to the pipeline. When trying to run the

orphanet step, the following error appeared "*length of 'dimnames' [2] not equal to array extent*". I tried to solve the error without success, so this step was skipped.

3.4.1.2 Results

The overlap between the variant identifiers obtained with the different packages was represented as a Venn diagram (see Figure 9).

A.



B.

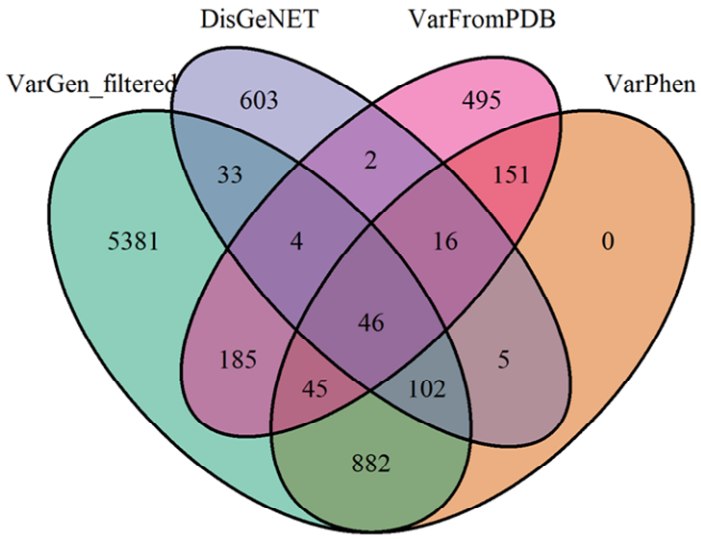


Figure 9: Venn diagrams representing the variants retrieved by the different pipelines: VarGen, DisGeNET, VarFromPDB and VarPhen. Obesity (OMIM: 601665) was chosen as the use case. A. Venn diagram using the raw output for all the pipelines. B. Venn diagram using the filtered VarGen dataset, with the following strategy: all the variants from the GWAS Catalog and with clinical significance were kept, and the remaining variants were filtered if their CADD Phred score was below 10.

Concerning the unfiltered results, the highest number of shared variants, 882, is between VarGen and VarPhen, despite the two pipelines implementing a different approach and data sources. DisGeNET is sharing 208 and 169 variants with VarGen and VarPhen respectively, while VarFromPDB is sharing 302 and 258 variants with VarGen and VarPhen respectively. In contrast, DisGeNET and VarFromPDB only share 68 variants, of which, only 2 are not found by VarGen nor VarPhen.

Some variants are discovered by only one package. Most of the 584 variants only discovered by DisGeNET are from literature mining and GwasDB, two resources not implemented yet in the other packages. From the 479 variants found uniquely with VarFromPDB, 408 are not directly linked with obesity, but other phenotypes (Intellectual Disability, Bardet-Biedl syndrome, etc) explaining the low overlap with the other tools. Many variants are found only by VarGen, since it reports variants affecting the genes related to a disease and not variants directly linked to the disease. Hence, some of the 119,243 variants uniquely found by VarGen are potentially false positives. This can be diminished by filtering variants based on their Phred score, source, and clinical significance, while keeping most of the variants found in common with the other databases. See below for more details. VarPhen has the best sensitivity / specificity ratio. It is ideal if the user does not want to filter the results manually.

To alleviate the potential amount of false positive obtained with the VarGen pipeline, the annotation step was implemented. From the annotated results, it is possible to filter the hits or to rank them based on their deleteriousness. More precisely, keeping the variants with a CADD score > 10, while keeping all the variants from GWAS or with information about clinical significance managed to reduce drastically the number of variants found only with VarGen while keeping a similar overlap with the other tools (see Figure 9B).

In summary, both VarGen and VarPhen are more sensitive than current existing alternatives. Specificity is also achieved by VarPhen and by filtering the results from VarGen. The variants uniquely detected by the other pipelines can be explained by different input databases for DisGeNET and the multiple

phenotypes used by VarFromPDB. Finally, the fact that only two variants are found by both DisGeNET and VarFromPDB suggest that no important variant is missed by VarGen nor VarPhen.

3.4.2 Second use case: Alzheimer's disease

The methods and results for this use case are available in Appendix 6.2.5B.2.

3.5 The lists of variants

Lists of variants for obesity, diabetes type 1 and type 2 were generated with VarGen, the goal was to compare the genotypes obtained during the Nutrishield clinical trials to these lists of variants.

3.5.1 Methods

For each disease, three outputs were produced, to provide different levels of sensitivity / specificity for each disease, (i) the raw output from VarGen (ii) the filtered output from VarGen and (iii) the output from VarPhen.

(i) For the VarGen pipeline, the `vargen_pipeline` function was run, with the input values detailed in Table 5 for each disease. The Fantom correlation threshold was set to 0.20. The list of variants was then annotated with `annotate_variants`.

(ii) The filtered list of variants was based on the output obtained from (i). Were kept, only the variants from the GWAS Catalog, with information about clinical significance and/or a CADD score higher than 10. This filtering should remove most of the less interesting variants while keeping the most impactful variants.

(iii) For the VarPhen pipeline, the phenotypes were searched with the `get_phenotype_terms` function from VarGen; with the following keywords *obesity*, *INSULIN-DEPENDENT* and *NONINSULIN-DEPENDENT* respectively for obesity, type 1 diabetes, and type 2 diabetes. These phenotypes were then given to `get_variants_from_phenotypes` to get the list of variants. The variants were then annotated with `annotate_variants` but no further filtering was performed, as VarPhen is specific enough.

The following versions were used: R v3.6.3, VarGen v0.2.1, the GWAS Catalog ‘e100_r2021-01-14’ and GTEx v8.

Table 5: List of input given to vargen_pipeline to generate the lists of variants for obesity, diabetes type 1 and diabetes type 2.

	OMIM	*GTEx tissues*	*GWAS terms*
Obesity	601665	Adipose Visceral Adipose subcutaneous	• Childhood obesity • Obesity-related traits • Obesity (extreme) • Obesity • Obesity without metabolic disease • Obesity (early onset extreme) • Obesity and osteoporosis • Type 2 diabetes (young onset) and obesity • Body mass index
Type 1 Diabetes	222100	Pancreas	• Diabetes mellitus • Type 1 diabetes • Type 1 diabetes and autoimmune thyroid diseases • Fulminant type 1 diabetes • Type 1 diabetes in high-risk HLA genotype individuals (time to event)
Type 2 Diabetes	125853	Pancreas	• Diabetes mellitus • Type 2 diabetes • Type 2 diabetes (young onset) and obesity • Prevalent type 2 diabetes • Type 2 diabetes (age of onset) • Type 2 diabetes and end-stage kidney disease • Type 2 diabetes (adjusted for BMI) • Schizophrenia and type 2 diabetes

Each output was given as input to Pascal [135] to obtain a list of pathways associated with the variants. When running Pascal, the following options were set*: --maxsnp=-1* to consider all genes, regardless of the number of SNPs located on them; *--genescoring=sum* to compute gene scores based on the sum-of-chi-squares, averaging the SNPs association signal across the gene region; --

runpathway=on to calculate the pathway scores. It is to be noted that the standard input for Pascal is a list of variants from a GWAS analysis with their associated p-values. As the variants from VarGen are not associated with a p-value, the CADD score was used as proxy, using Equation 2.

$$pascal\ score = \frac{0.1}{CADD\ score}$$

Equation 2: Transformation of the CADD score into a score for Pascal. The arbitrary value of 0.1 was chosen as it resulted in a range similar to p-values obtained from GWAS.

3.5.2 Results & Discussion

3.5.2.1 Obesity

VarGen retrieved 11 genes linked to obesity (OMIM 601665). The pipeline retrieved 80,294 variants, and after filtering, the list was reduced to 12,762 high-impact variants. Concurrently, VarPhen found 1,677 variants related to obesity phenotypes. The 4,848 variants retrieved from the GWAS Catalog as part of the VarGen pipeline are represented as a Manhattan plot on Figure B.1-1.

The results from the annotation for each pipeline is represented in Figure 10. As expected, the largest amount of variants was obtained with VarGen, which also have the largest amount of 'NA' annotations from snpEff (Figure 10A). Concerning clinical significance, it is interesting to note that, despite the difference in the number of variants, there is a similar distribution between VarGen and VarPhen. This indicates the lack of knowledge about the pathogenicity of most variants and of the interest of integrating several data sources to explore variant-disease relationships (see Figure 10B). Finally, the CADD score distribution is quite low for VarGen and VarPhen, as most of the variants falls below a score of ten. This can be explained by the Phred nature of the CADD scores ($\log_{10}$). Thus, a variant with a Phred score >10 is in the top 10% of all variants (in terms of raw CADD score), while a Phred score >20 indicates a variant in the top 1% (see Figure 10C). CADD correlate with pathogenicity [124], making variants with high CADD scores valuable subjects to study diseases.

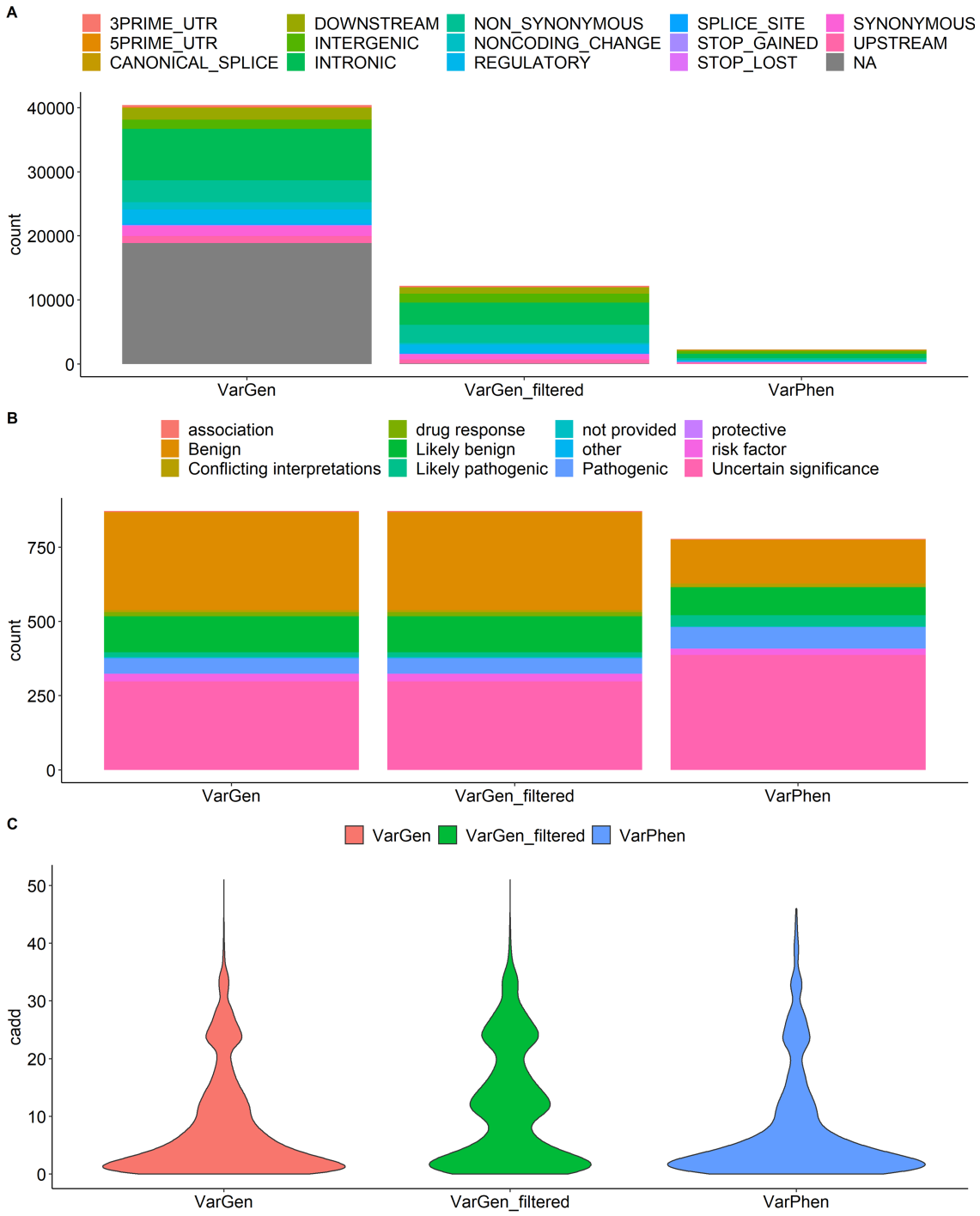


Figure 10: Details about the annotations of the three list of variants obtained with VarGen (raw and filtered) and VarPhen for obesity. Empty annotations ("") were ignored, for the sake of clarity A) Stacked barchart of the consequence terms from snpEff. B) Stacked barchart of the clinical significance terms from clinvar. The distribution is the same between VarGen and VarGen_filtered, since all the variants with clinical significance were kept during the filtering step C) Violin plot representing the distribution of the CADD scores for each pipeline.

The pathway analysis performed with Pascal assigned a chi2Pvalue to each one of the 1,077 pathways in their database. This discussion will focus on the top 15 pathways, represented in Table 6.

Table 6: Top 15 pathways obtained with Pascal from VarGen's filtered list of variants for obesity.

Pathway Name	Chi2Pvalue
Reactome Mitotic Prometaphase	0.0002198
Reactome DNA Replication	0.0004419
Reactome Nuclear Receptor transcription pathway	0.0009311
Reactome Transport to the Golgi and subsequent modification	0.0010427
Reactome Antigen Presentation: Folding assembly and peptide loading of class I MHC	0.0010427
Reactome mitotic MM G1 phases	0.0011293
Reactome Asparagine N-linked glycosylation	0.0025005
Biocarta Nuclearrs pathway	0.0049984
Reactome Synthesis secretion and deacylation of Ghrelin	0.0052137
Reactome Generic Transcription Pathway	0.0055095
Reactome MHC class II antigen presentation	0.0062515
Kegg ECM receptor interaction	0.0068640
Reactome activation of chaperone genes by XBP1S	0.0083420
Reactome Unfolded Protein Response	0.0083420
Reactome class I MHC mediated antigen processing & presentation	0.0088980

The ghrelin hormone regulates appetite stimulation and growth when binding to the *GHS-R1a receptor.* This explains the presence of the `synthesis, secretion and diacylation of ghrelin` pathway in this list. Several mutations located in the ghrelin gene and its receptor are associated with human obesity and short stature [136] [137].

Concerning the `ECM receptor interaction` pathway, Lin et al. proposed a mechanism for induced insulin resistance in obesity caused by the ExtraCellular Matrix (ECM) receptors [138]. More precisely, the activation of ECM receptors pathways in adipose tissues induces adipocyte death, inhibition of angiogenesis and promotion of macrophage infiltration which lead to inflammation and insulin

resistance. Interestingly, in the list, three pathways are linked to the adaptative immune system, namely `class I MHC mediated antigen processing presentation`, `antigen presentation folding assembly and peptide loading of class I MHC` and `MHC class II antigen presentation`. A review from Bastard et al. highlighted the impact of the overexpression of inflammatory molecules, such as *TNF-alpha* and *IL-6*, in obesity and their impact on key steps of the insulin signalling pathway in different model species [89]. Thus, in obesity, ECM receptors enhance macrophage infiltration in adipose tissues, which in turn overexpresses inflammatory molecules, such as *TNF-alpha* and *IL-6*, leading ultimately to insulin resistance.

Two pathways are linked to protein folding, namely `activation of chaperone genes by XBp1s` and `unfolded protein response`. There is yet no clear link between obesity and protein folding, however a study identified that *XBp1s* overexpression reduced obesity in mouse models by the activation of lipolysis [139].

Two pathways are linked to the Golgi complex, `asparagine N linked glycosylation` which contains the `transport to the Golgi and subsequent modification` pathway. The Golgi complex is the central sorting and processing station of the secretory pathway, including lipid regulation. It was identified as a possible therapeutic target for obesity [140].

Another interesting pathway is `nuclears pathways`, some nuclear receptors are responsible for lipid metabolism, storage, or elimination. So, variants affecting this pathway might disrupt lipid metabolism. This also explains the presence of the `nuclear receptor transcription pathway` which is part of the high-level pathway `generic transcription`. Moreover, nuclear receptors are related to lipid sensing, liporegulation and insulin resistance [141].

The remaining three pathways are linked to replication, namely `mitotic prometaphase`, `mitotic M-M/G1 phases` and `DNA replication`. These pathways are very high-level, but mutations in a gene responsible for DNA replication and integrity, *WRN*, were found to exaggerate obesity [142]. Indeed,

when compared with wild types, WRN-deficient mice developed signs of obesity, i.e., weight gain, hyperinsulinemia, and insulin resistance. The same observation was made for mice with a SNP in the helicase domain of the *WRN* protein [143].

3.5.2.2 Diabetes type 1

VarGen queried OMIM and retrieved 4 genes linked to diabetes mellitus type 1 (OMIM 222100). The pipeline retrieved 40,304 variants, after filtering, the list was reduced to 4,105 high-impact variants. Concurrently, VarPhen found 524 variants related to diabetes type 1 phenotypes. The 295 variants retrieved from the GWAS Catalog are represented as a Manhattan plot in Figure B.1-2.

The results from the annotation are presented in Figure 11. In terms of snpEff consequences, the same conclusion as the one reached for obesity can be made here (see Figure 11A). Interestingly, VarGen has twice as many variants with information about clinical significance than VarPhen (see Figure 11B). This could be because the genetics behind diabetes mellitus type 1 are not fully understood yet, thus VarPhen, which retrieves variants that have been directly linked to the disease, is picking up less variants. This highlights the potential of VarGen to discover new variants of interest for less studied diseases and it would be of interest to explore the clinically significant variants found by VarGen to further our understanding of diabetes mellitus type 1. The distribution of the CADD Phred scores is slightly different than the one obtained with obesity (see Figure 11C). Many VarPhen variants have a low CADD score, on one hand it could indicate that our filtering for VarGen (CADD > 10) might be too stringent here; on the other hand, it might just be due to the low number of variants retrieved by VarPhen. One can always merge the output obtained from VarPhen and VarGen, and it would be especially relevant here.

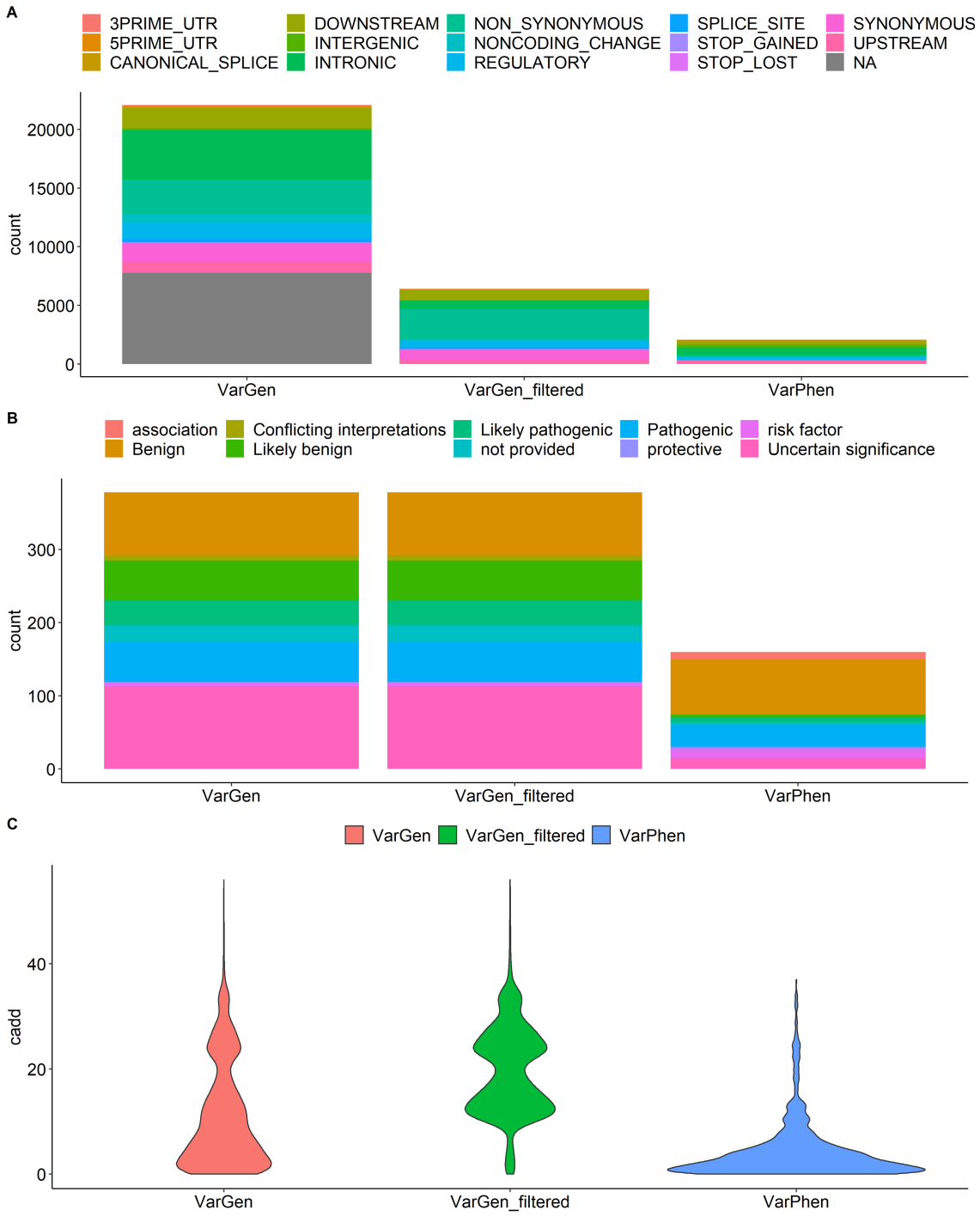


Figure 11: Details about the annotations of the three list of variants obtained with VarGen (raw and filtered) and VarPhen for diabetes mellitus type 1. Empty annotations ("") were ignored, for the sake of clarity A) Stacked barchart of the consequence terms from snpEff. B) Stacked barchart of the clinical significance terms from clinvar. The distribution is the same between VarGen and VarGen_filtered, since all the variants with clinical significance were kept during the filtering step C) Violin plot representing the distribution of the CADD scores for each pipeline.

The pathway analysis performed with Pascal assigned a chi2Pvalue to each one of the 1,077 pathways in their database. This discussion will focus on the top 15 pathways, represented in Table 7.

Table 7: Top 15 pathways obtained with Pascal from VarGen's filtered list of variants for diabetes mellitus type 1.

Pathway Name	Chi2Pvalue
Biocarta Mitochondria Pathway	0.0093023
Kegg Vascular smooth muscle contraction	0.0139534
Kegg Gap junction	0.0139534
Kegg Long term potentiation	0.0139534
Kegg Long term depression	0.0139534
Kegg Taste transduction	0.0139534
Kegg GNRH signaling pathway	0.0139534
Kegg Alzheimers disease	0.0139534
Reactome DAG and IP3 signaling	0.0139534
Reactome antigen activates B cell receptor leading to generation of second messengers	0.0139534
Reactome Opioid Signalling	0.0139534
Reactome PLC beta mediated events	0.0139534
Reactome Elevation of cytosolic Ca2+ levels	0.0139534
Reactome Regulation of insulin secretion by glucagon-like Peptide-1	0.0139534
Reactome Platelet homeostasis	0.0139534

The most obvious pathway related to diabetes in this list is the `regulation of insulin secretion by glucagon like peptide 1`. *Glucagon-like Peptide-1* is secreted in response to glucose and fatty acids, then binds to the beta cells of the pancreas and unfolds a series of cascading events. This leads to enhanced insulin secretion involving the *Protein Kinase A* and *Rap1A*. Thus, deleterious variants impacting this pathway could explain the lack of insulin secretion in diabetes type 1. Interestingly, the `taste transduction` pathway, present in this list, might regulate the secretion of *glucagon-like peptide-1* [144].

Several pathways are linked to Ca2+ flux, namely `elevation of cystosolic Ca2 levels`, `PLC beta mediated events` and `DAG and IP3 signaling`. Bot *DAG* and *IP3* have been linked to insulin [145], indeed *DAG* is involved in

insulin secretion via *PKC*, while *IP3* signal the release of Ca2+ from the endoplasmic reticulum, which causes the secretion of insulin through the membrane. The `elevation of cystosolic Ca2 levels` happens upstream and start the `PLC beta mediated events` which hydrolyses *PIP2* into both *IP3* and *DAG*. Interestingly, another pathway found by Pascal is the `platelet homeostasis` and the elevation of Ca2+ levels is essential for platelet activation. The `GnRH signaling pathway` is also upstream of *IP3* / *DAG* and was shown to be affected by type 1 diabetes [146].

Another pathway from the list, `opioid signalling,` increases intracellular calcium and indirectly impacts `PLC beta mediated events` [147]. A review from Singh et al. highlighted that opioids receptors from the pancreas have a complex influence on insulin homeostasis: on one hand acute opioids exposure increases insulin secretion, on the other hand chronic opioids exposure leads to decreased secretion [148]. This was shown by the impact of opioid receptor agonists on insulin secretion in mice. `Antigen activates B cell receptor leading to generation of second messengers` is a high-level pathway which contains the `PLC beta mediated events` pathway. Similarly, `gap junction` is a more generic pathway which describes the channels of communication between adjacent cells. Gap junction is affected by changes in intracellular Ca2+ levels and might be linked to the aforementioned pathways.

Many biological processes are affected by Ca2+ and insulin, which explains the presence of the `vascular smooth muscle contraction` and `long-term depression` (LTD) pathways in this list. LTD is involved in synaptic plasticity, learning, and forgetting. Low Wang et al. found that insulin affects Vascular Smooth Muscle Cell quiescence and migration, respectively via the *PI3K* and *MAPK* pathways [149]. Dysregulation of intracellular Ca2+ levels seems to affect Alzheimer's disease, Popugaeva et al. highlighted that many Alzheimer's models include endoplasmic reticulum Ca2+ excess [150]. Moreover, there is growing evidence that Alzheimer's disease shares features with diabetes and some researchers even go as far as calling it 'type 3 diabetes' [151] [152]. This justifies the presence of `Alzheimers disease, long-term depression,` and

`long-term potentiation` (both linked to learning and memory). The role of insulin in the central nervous system, via *PI3K*, is another parallel between diabetes, and Alzheimer's [153].

Concerning the `mitochondrial pathway`, Belosludtsev et al. reviewed the role of mitochondrial dysfunctions in diabetes, especially in the mitochondrial calcium transport systems [154]. More precisely, the consequences of mitochondrial dysfunction, i.e., impaired calcium homeostasis, excessive ROS production and mitochondrial permeability transition pore opening, are also present in diabetes, suggesting a role of mitochondrial dysfunction in the aetiology of the disease. Finally, mitochondrial dysfunction might accelerate the complications of diabetes mellitus by the death of β-cells in the pancreas.

In conclusion, the combination of VarGen and Pascal highlighted pathways of interest for diabetes mellitus type 1. Some of these pathways are straightforward and relate to insulin, while others highlight an interesting overlap with Alzheimer's disease. Moreover, the results underlined the important role of Ca2+ flux, and more precisely *IP3* and *DAG*, in diabetes pathogenesis.

3.5.2.3 Diabetes type 2

VarGen retrieved 29 genes linked to diabetes mellitus type 2 (OMIM 125853). The pipeline retrieved 493,860 variants, and after filtering, the list was reduced to 20,064 high-impact variants. Concurrently, VarPhen found 524 variants related to diabetes type 2 phenotypes (of which 231 are in common with diabetes mellitus type 1). The 3,326 variants retrieved from the GWAS Catalog are represented as a Manhattan plot on Figure B.1-3.

The results from the annotation are available on Figure 12. Since diabetes mellitus type 2 is linked with substantially more genes in OMIM, more variants are retrieved with VarGen. Thus, the effect of filtering is even more important here (see Figure 12A). The number of variants with information about clinical significance obtained here is an order of magnitude higher compared to obesity and diabetes mellitus type 1 (see Figure 12B). This could allow for a strong PRS or gene network analysis.

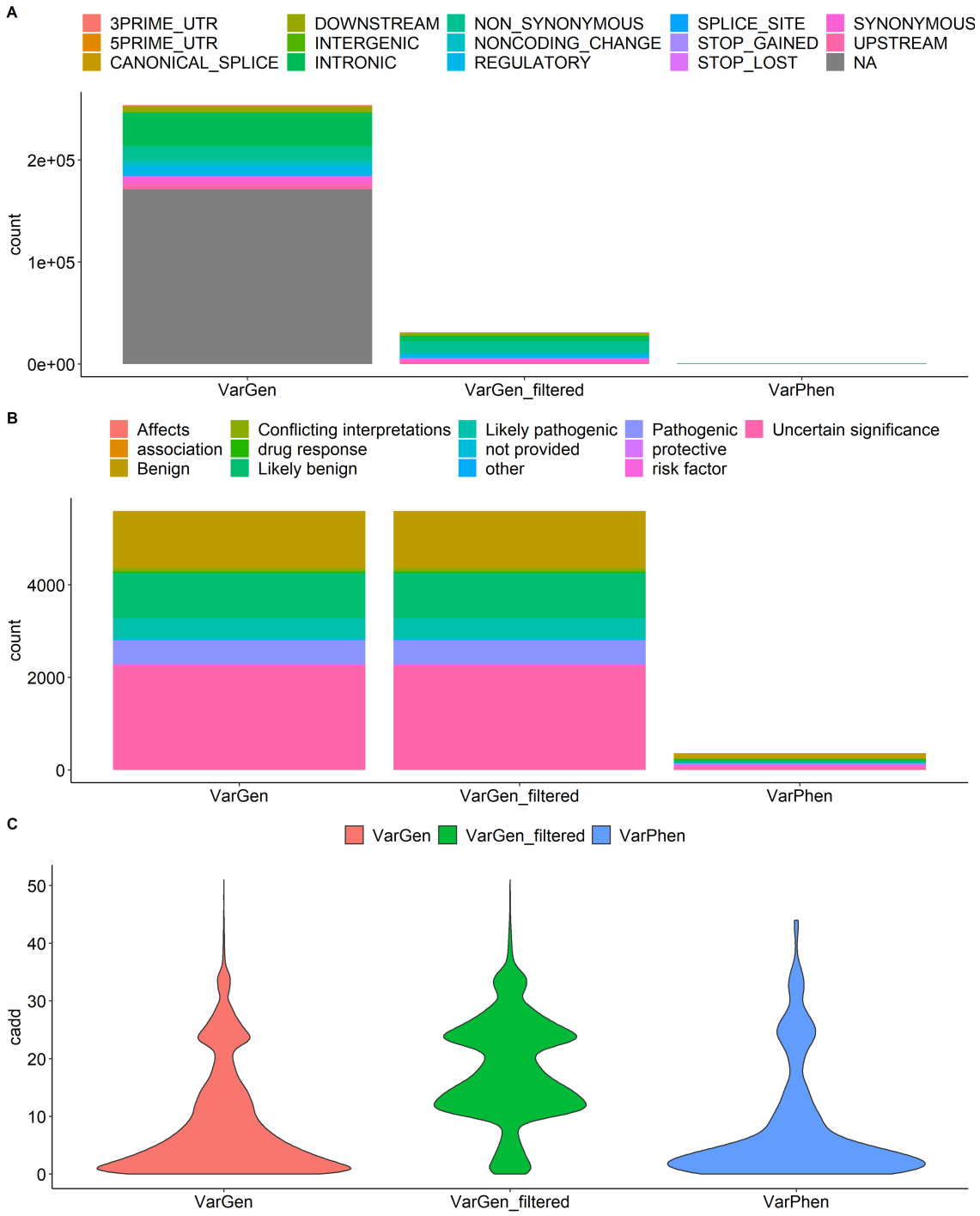


Figure 12: ***Details about the annotations of the three list of variants obtained with VarGen (raw and filtered) and VarPhen for diabetes mellitus type 2. Empty annotations (“”) were ignored, for the sake of clarity A) Stacked barchart of the consequence terms from snpEff. B) Stacked barchart of the clinical significance terms from clinvar. The distribution is the same between VarGen and VarGen_filtered, since all the variants with clinical significance were kept during the filtering step C) Violin plot representing the distribution of the CADD scores for each pipeline.***

The pathway analysis performed with Pascal assigned a chi2Pvalue to each one of the 1,077 pathways in their database. This discussion will focus on the top 15 pathways, represented in Table 8.

Table 8: Top 15 pathways obtained with Pascal from VarGen's filtered list of variants for diabetes mellitus type 2.

Pathway Name	Chi2Pvalue
Reactome Regulation of beta-cell development	0.0001576
Kegg Insulin signaling pathway	0.0001640
Kegg Arrhythmogenic right ventricular cardiomyopathy	0.0001675
Reactome Regulation of gene expression in beta cells	0.0002084
Kegg type II diabetes mellitus	0.0008199
Kegg Melanogenesis	0.0009231
Kegg Thyroid cancer	0.0010050
Kegg Focal adhesion	0.0010122
Kegg Colorectal cancer	0.0014548
Kegg Adherens junction	0.0016475
Reactome Developmental Biology	0.0017427
Reactome Neuronal System	0.0024329
Reactome Nuclear Receptor transcription pathway	0.0029706
Kegg Starch and sucrose metabolism	0.0030355
Kegg Renal cell carcinoma	0.0032300

One of the elements from this list is the general `type II diabetes mellitus` pathway from KEGG, which describes the molecular mechanisms involved in this disease. This general pathway contains the `insulin signalling pathway` which in turn contains the `starch and sucrose metabolism` pathway, both present in the list. This can help to pinpoint to impact of the variants found by VarGen on the general metabolism of type 2 diabetes.

Three pathways are linked to β-cells, the producers of insulin, which are a key component of type 2 diabetes, namely `regulation of gene expression in beta cells`, which in contained in `regulation of beta cell development`, itself contained within the `developmental biology` pathway.

Four pathways are related to cancer, namely, `melanogenesis`, `thyroid cancer`, `colorectal cancer,` and `renal cell carcinoma`. Interestingly, the *insulin growth factors 1* and *2* are anti-apoptotic hormones and are probably necessary for the survival of cancer cells [155]. Epidemiological studies have found links between diabetes type 2 and cancer, notably hyperinsulinemia is a risk factor for cancer as well as a potential target for therapy [156]. Vella et al. observed over-expression of an insulin receptor in thyroid cancer cells [157]. Trevisan et al. found that hyperinsulinemia, insulin resistance, and metabolic abnormalities are risk factors in the aetiology of colorectal cancer [158]. Additionally, type 1 insulin-like growth factor inhibitors have therapeutic value for renal cell carcinoma [159].

`Arrhythmogenic right ventricular cardiomyopathy` is a disease that may result in heart failure. In cardiomyocytes, Ca2+ are released via insulin action on *IP3* receptors, which might influence cardiac metabolism and physiology [160]. Part of this pathway involves the `adherens junction`.

Interestingly, `nuclear receptor transcription pathway` was also found in the pathways related to obesity, this can hint to parallels between the two diseases. On a same note, the *PPARG* gene was found both in obesity and type 2 diabetes with VarGen. Nuclear receptors have a role in the modulation of insulin secretion and lipid induced insulin-resistance [141].

`Focal adhesion` represents the mechanical links between intra and extra cellular substrates and is an important step for insulin secretion.

Another interesting pathway is `neuronal system`, insulin signalling of the central nervous system underlies pathologies such as Alzheimer's disease [153], as mentioned in the analysis of the pathways found for diabetes type 1 (see 3.5.2.2).

3.5.2.4 Overlap between the diseases

The three sets of filtered variants obtained for obesity, diabetes mellitus type 1 and 2 were intersected by rsIDs. The resulting Venn diagram is available as *Figure 13*.

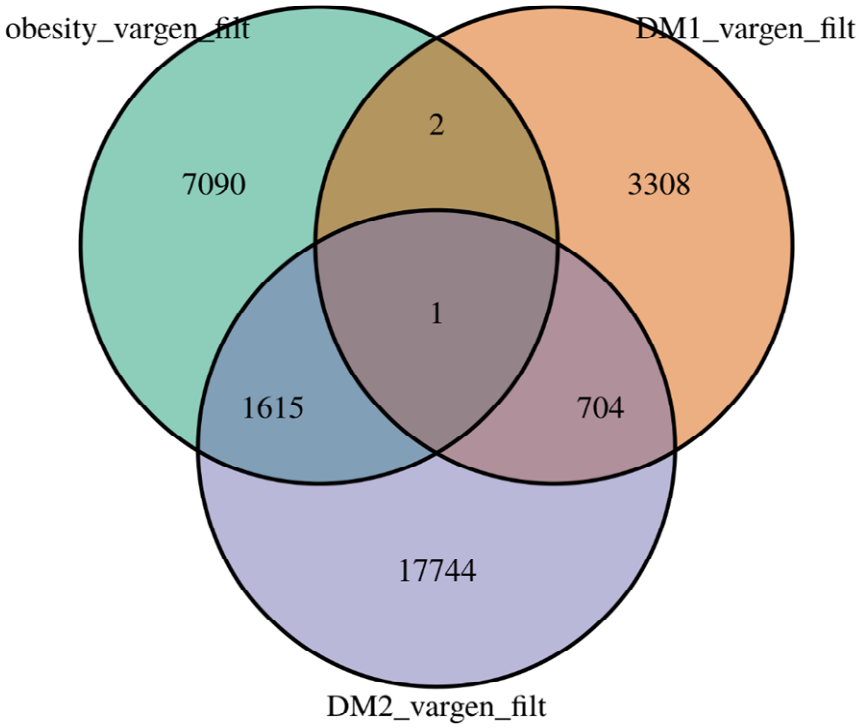


Figure 13: Venn diagram representing the overlap of variants found with VarGen (filtered as described in 3.5.1) for obesity, diabetes mellitus type 1 (DM1) and type 2 (DM2).

Obesity and diabetes type 1 shares only three variant and must have very distinct genetic causes. Interestingly, diabetes type 2 shares variants with both obesity and diabetes type 1. Studying the overlaps between these different diseases might help to understand the common causes of obesity and diabetes, thus shedding some light on the genetics of the metabolic syndrome.

3.6 Conclusion

VarGen is an easy-to-use, versatile, R package that can quickly retrieve variants linked to a disease of interest. It can retrieve known information about well-studied diseases, while finding new knowledge, as seen with obesity and diabetes type 2. It can also help in the discovery of new variants and retrieve interesting pathways for less-studied diseases, as seen with diabetes type 1. While the variants found still need to be validated by experimental approaches, VarGen can speed up the discovery and shortlisting of variants. VarGen is open-source, and we hope that it will be useful for the research community to infer new relationships between variants and diseases.

4 A two-step Polygenic Risk Score for Body Mass Index

4.1 Introduction

Prevention is better than cure. Going one step further, one can even aim at tailoring prevention based on each individual's genetic makeup. This can be achieved by studying the genome and how it affects traits. More specifically, if we know the impact of DNA variants on a disease, it becomes possible to calculate the predisposition of a person to develop this disease. For monogenic traits, such as Huntington disease and cystic fibrosis, it is as straightforward as a 'yes or no' answer since they are caused by mutations on a single gene. However, polygenic traits, such as obesity, are affected by the cumulative effect of a multitude of variants, across the genome, which makes the estimation of genetic predisposition a complex task to tackle. Fortunately, Polygenic Risk Scores (PRS) provide a way to summarise, as a single value, the genetic risk an individual has of developing a certain disease. This single value is obtained by summing all the disease-causing alleles carried by the individual. If available, this sum can be weighted by the effect size of each variant, to reflect the varying impact of each variant, for a more accurate estimation.

The first step to perform a PRS analysis is to identify the disease-related set of variants that will serve to develop the model. This is not a straightforward task, as the variants related to complex traits are often non-coding and affect gene expression rather than protein integrity [22]. Fortunately, one can harness the knowledge gained from the many Genome Wide Association Studies (GWAS) performed this past decade [31] (see 2.1.3.5).

PRS are also applicable to continuous traits, here it was applied to the Body Mass Index (BMI). The goal was to identify individuals more at risk of becoming overweight or obese. Prevention can then be targeted to high-risk individuals; diet and lifestyle modifications remain the best way to prevent obesity and its many adverse health impacts. Obesity is associated with a shorter life span and is a risk factor for many other co-morbidities: diabetes, cardiovascular disease, and cancer, to name a few.

But, is BMI an accurate measure of obesity? Despite the widespread use of this index across the world, some criticism has been raised against it. Müller et al. [84] made the argument that BMI is an oversimplification of obesity and should not be used as a target for GWAS analyses. They argue that BMI is an anthropomorphic and not a biological measurement, and additionally, is not representative of body shape or fat distribution. Moreover, BMI represents the weight at a set time in a person's life, it would be more accurate, albeit more difficult, to study the genetics behind weight control or susceptibility to obesity. In response to this argument, Speakman et al. [161] (i) listed the many genes linked to obesity which were successfully identified from GWAS targeting BMI (ii) mentioned the ease of measuring BMI, which leads to large cohorts, which is key to obtain refined GWAS results. Finally, while the argument of BMI being oversimplistic is true, it remains representative of body fat percentage in the global population, thus stands as an informative avatar of obesity when dealing with many individuals. For these reasons, I decided to maintain BMI as the target for this PRS analysis.

Regarding clinical utility, the reader should keep in mind that PRS, on its own, is not the panacea of individual risk assessment. First, the risk remains relative within the population studied. Second, genetic risk only represents one side of the story, complex diseases also involve other factors, such as the environment and the microbiome. Third, PRS models usually do not take into account, rare, monogenic variants associated with the disease, which are very impactful, albeit for a small portion of the population [49]. Despite these limitations, PRS remains a powerful tool for understanding and preventing diseases, especially when used in combination with other clinical data [49].

Here, a new method to refine PRS models will be presented. It is based on the use of a second PRS model, based on variants obtained with VarPhen, to account for rarer variants. Individuals with an extreme low or high score in this second PRS will be readjusted in the original PRS to reflect this second risk assessment. This approach was validated on BMI and diabetes.

4.2 Methods

This PRS model was developed following the guidelines from Choi et al. [42]. As mentioned in 0, the development of a PRS model requires two sets of data: **the base dataset**, which contains the variants linked to the disease and **the target dataset**, on which the PRS will be applied. Both the target and the base sets are using the latest human reference genome, *GRCh38*. The data were processed with plink v1.90 [43] and R v3.6.3.

4.2.1 Base data: GWAS on obesity

The base data were obtained from the meta-analysis of 82 GWAS and 43 Metabochip studies of BMI, performed by Locke et al. [99], amounting to 339,224 individuals, including 322,154 of European origin. Metabochip is a genotyping array customised for the study of metabolic diseases. The summary statistics of the meta-analysis were downloaded from the GWAS Catalog (study identifier: GCST002783).

The raw data contained 2,555,086 variants, including the non-significant ones. Quality control was performed, by removing duplicated SNPs as well as ambiguous SNPs with complementary alleles (A/T and C/G), to avoid any possible source of mismatch. After this initial filtering step, 2,160,856 SNPs were retained.

4.2.2 Target data: UK biobank

The target set is the UK Biobank, a biomedical database containing anthropomorphic and genetic information for ~500,000 individuals from the United Kingdom (UK), aged between 40-69 years. The participants were recruited at one of 22 centres across the UK, between 2006 and 2010 [105]. UK Biobank has approval from the North-West Multi-centre Research Ethics Committee (REC reference 11/NW/0382). This analysis was conducted under UK Biobank application 55079. All the analyses described in this chapter are using the baseline data, corresponding to the first visit at the centre. The individuals in UK Biobank were genotyped with one of two very similar custom arrays: *UK BiLEVE Axiom Array* or *UK Biobank Axiom Array*, both consisting of ~800,000 genetic

markers [162]. The genotyping data are available under the plink format, namely *ped*, *bim* and *fam* files (see 6.2.5C.1). The coordinates of the genotyping were based on GRch37, since the base and target data need to be both referring to the same version of the human reference genome, the UK Biobank plink files were lift-overed to GRch38 with *liftOverPlink* [163].

First, since the base data were obtained from individuals of European descent, the UK Biobank data were filtered to only keep individuals reporting their ethnic group as 'British' or 'Irish'. This corresponded to more than 90% of UK Biobank. This was performed to reduce bias in the PRS, as most variants reported from GWAS are not causal but in Linkage Disequilibrium (LD) with the causal variants, and this LD pattern depends on population genetic structure [116] (see 2.1.3.5).

Then, standard GWAS quality control was applied to the target data with plink v1.90. SNPs were removed if: their minor allele frequency was lower than 0.01 (`--maf 0.01`); their p-value from the Hardy-Weinberg Equilibrium Fisher's exact test was lower than 1e-6 (`--hwe 1e-6`), since it indicates a higher probability of genotyping error; there were missing from more than 1% of the samples (`--geno 0.01`). Finally, individuals with more than 5% of missing genotype were removed from the analysis (`--mind 0.05`).

Including related samples might include bias in the model. Thus, samples with a 1st or 2nd degree relative also present in the biobank were removed. First, highly correlated SNPs were pruned with the parameter `–-indep-pairwise 200 50 0.25`, meaning that across a sliding window of 200 variants, sliding by 50 variants at a time, SNPs were removed if their Linkage Disequilibrium r^2 was higher than 0.25. Then samples with a F coefficient, estimating the level of inbreeding, outside 3 standard deviations of the UK Biobank mean were removed. Then, relatedness was pre-calculated with the `--make-grm-gz` parameter, and if two samples had a genomic relatedness higher than 0.125, one of the pair was removed (`--rel-cutoff 0.125`). Demographics for the remaining samples are detailed in Table 9.

Table 9: Demographics for the individuals included in the PRS analysis.

Number of individuals	373,397	
Age at recruitment (years)	Median IQR (Q1-Q3)	58 12 (51 – 63)
Gender	Male	46% (n = 172,264)
BMI	Median IQR (Q1-Q3)	26.7 5.7 (24.1 - 29.8)

4.2.3 Polygenic risk score calculation

An important step before calculating the PRS is to remove correlated SNPs while keeping the independent causal variants from the GWAS results, a step called 'clumping'. The variants from the base set were clumped using the aforedescribed processed UK Biobank population for the LD calculations. The following plink parameters were chosen: `--clump-p1 1` so that all SNPs are included for clumping, regardless of their p-value; `--clump-r2 0.1` to remove SNPs with a r^2 correlation higher than 0.1; `--clump-kb 250` to consider a 250 kbp window around the current SNP. After clumping, 80,789 independent SNPs remained.

Population stratification is usually a confounder in GWAS analyses and a way to alleviate this is to add Principal Components as covariates to the model. The first 6 Principal Components of the processed UK Biobank dataset were computed with plink via the `--pca 6` option.

For the PRS calculation itself, a range of different p-value thresholds were considered, with the `--q-score-range` option, namely 0.001; 0.05; 0.1; 0.2; 0.3; 0.4 and 0.5. For example, for threshold 0.2, only the SNPs with a p-value between 0 and 0.2 were included in the model. Plink is using the formula described in Equation 3 to compute the PRS for each sample.

$$PRS_j = \frac{\sum_i^N (ES_i * EA_{ij})}{P * S_j}$$

Equation 3: Plink formula to compute the Polygenic Risk Score for sample j. With N being the total number of variants, ES_i the effect size for SNP i, EA_{ij} the number of effect alleles observed in sample j, P the ploidy (here 2) and S_j the number of non-missing SNPs in sample j.

The best-fit PRS was assessed with the amount of explained BMI variance at each threshold. The best-fit of 4.9% explained variance was achieved with the 0.3 threshold, containing 31,517 variants (see Figure 14). Then, covariates were added to the model, i.e.: sex and the 6 Principal Components calculated previously.

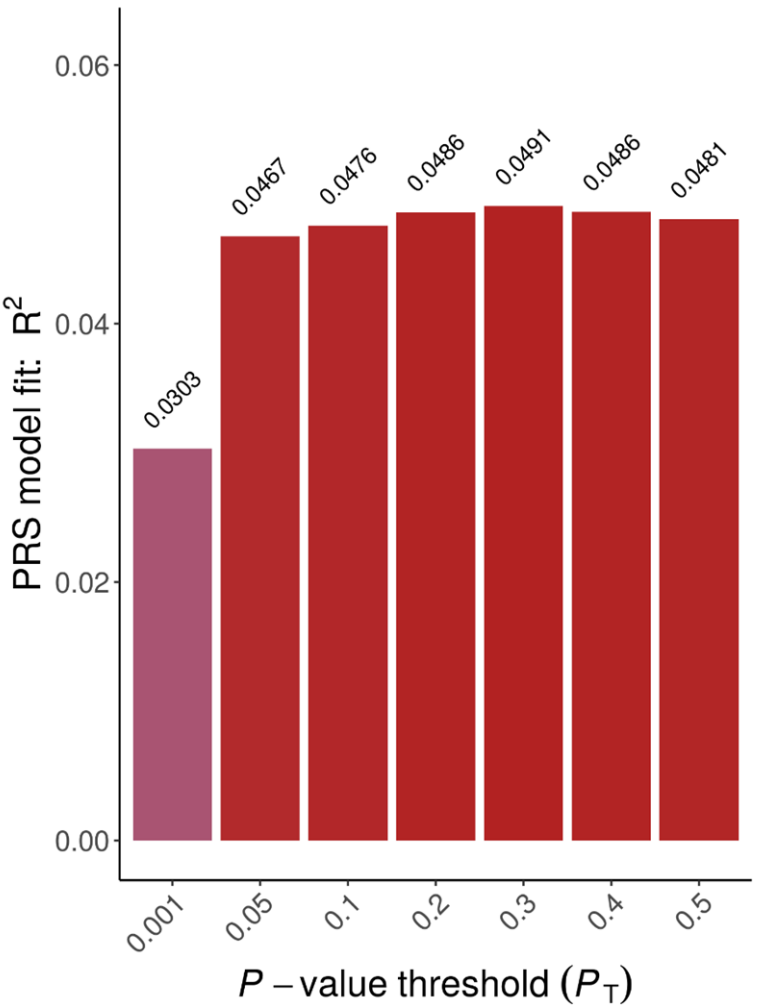


Figure 14: ***R^2 value obtained for each p-value threshold. For each threshold, a linear model of the BMI as a function of the PRS score, sex and the first 6 Principal Components was created. The R^2 obtained from the model was subtracted by the R^2 from a null model containing the covariates without the PRS score.***

The number of participants and SNPs filtered at each step is detailed in Figure 15. The final PRS model consisted of 373,397 individuals and 31,517 SNPs.

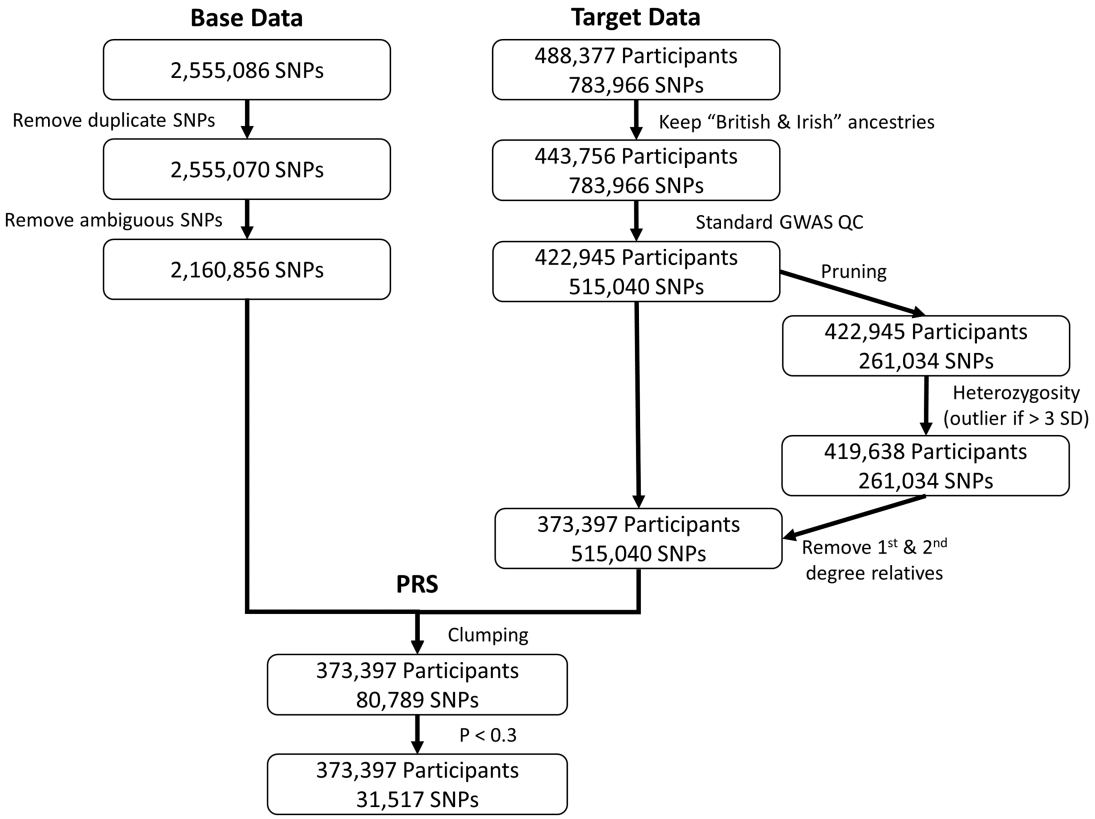


Figure 15: Data preparation workflow for the PRS analysis. For both the base and target datasets. The final PRS was based on 373,397 individuals and 31,517 predictors.

4.2.4 Refining the model with VarPhen

The PRS defined previously suffers from the same flaw as most of the other PRS, its base set, results from a GWAS, is mostly composed of common SNPs. But, as mentioned by Torkamani et al. [49] rarer variants can drastically influence genetic risk estimations, this could lead to over or under estimation of the actual risk for individuals carrying these more impactful variants.

To alleviate this, a second, unweighted, PRS was implemented using the same target set, UK Biobank, but another base set, composed of variants retrieved by VarPhen (see 3.2.2.1). The input to VarPhen, listed in Table 10, consisted of phenotypes related to obesity and body mass index, and the pipeline retrieved 12,348 variants. After filtering out InDels, variants with missing risk alleles or missing from the target set, the base set was composed of 287 valid predictors.

Since VarPhen does not return the effect size, a value of 1 was assigned for each SNP, resulting in an unweighted PRS model.

Table 10: List of phenotypes given as input to VarPhen, in order to get the SNPs related to obesity and BMI.

List of phenotypes	
Obesity	Obesity mild early-onset
Obesity extreme	Morbid obesity
Obesity (early onset extreme)	Obesity-related traits
Obesity early-onset susceptibility to	Monogenic Non-Syndromic Obesity
Obesity due to SIM1 deficiency	Obesity autosomal dominant
Obesity (BMIQ14) susceptibility to	Abdominal obesity-metabolic syndrome 3
Body Mass Index	Body Mass Index Quantitative Trait Locus 4
Body Mass Index Quantitative Trait Locus 9	Body Mass Index Quantitative Trait Locus 10
Body Mass Index Quantitative Trait Locus 12	Body Mass Index Quantitative Trait Locus 18
Body Mass Index Quantitative Trait Locus 19	Body Mass Index Quantitative Trait Locus 20

4.3 Results and discussion

For the remainder of this chapter, the PRS based on the GWAS meta-analysis by Locke et al. [99] will be referred as the **backbone PRS***,* while the PRS made from the set of SNPs obtained with VarPhen will be referred as the **VarPhen PRS***.*

4.3.1 The *backbone PRS* for BMI

The individuals from the target set (UK Biobank) were split into 10 quantiles, each containing more than 37,000 individuals, based on their *backbone PRS* score. The BMI mean of each quantile was computed and plotted (see Figure 16). There is a clear positive correlation between the BMI and the PRS score, especially at the two extreme quantiles. The highest PRS quantile has a BMI mean of almost 30 which corresponds to an obese state. This indicates that the model can accurately estimate the genetic predisposition of an individual to gain weight and even identify those most at risk to develop obesity. This was confirmed with the development of a linear regression model of BMI as a function of PRS_{score}, resulting in a $R^2 = 0.049$ and a p-value < 2.2e-16. These results are consistent with a previous PRS for BMI [164].

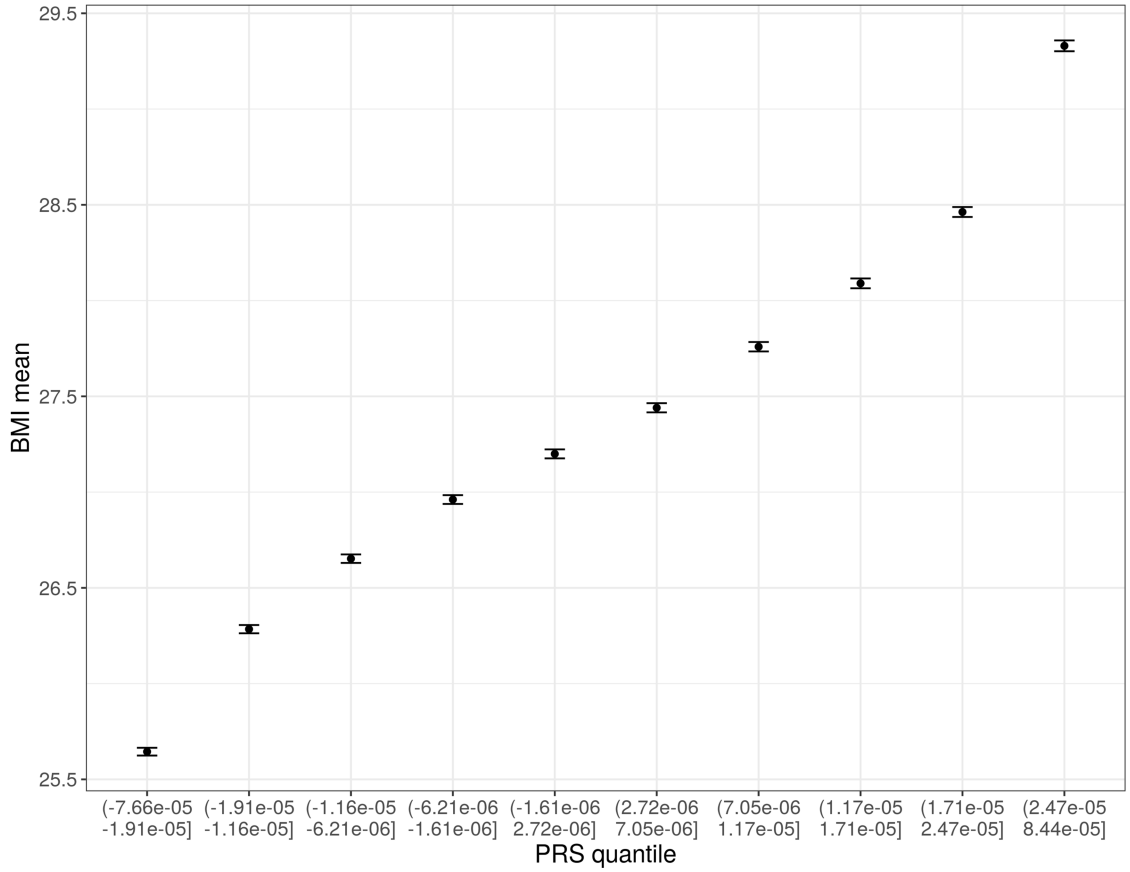


Figure 16: BMI mean for each backbone PRS quantile. Each quantile contains ~37,000 individuals. The bars correspond to the standard error of the mean.

4.3.2 Readjustment with VarPhen

The same approach as the *backbone PRS* was used to visualise the *VarPhen PRS*. The target dataset was split in 10 quantiles and the BMI mean of each quantile was computed (see Figure 17). Interestingly, despite being an unweighted PRS based on only 287 variants, a good correlation between the BMI and the score was observed. A linear regression model of BMI as a function of PRS_{score} resulted in a R^2 of 0.0082 and a p-value < 2.2e-16.

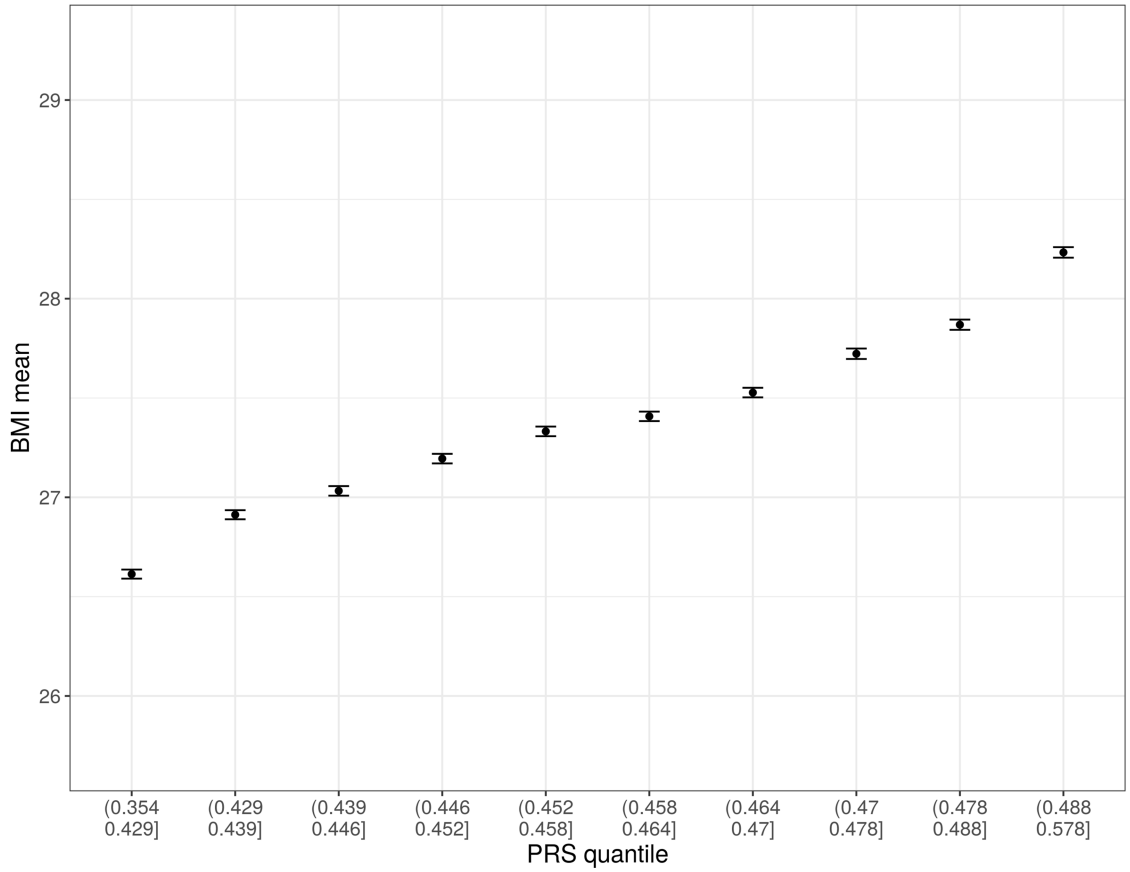


Figure 17: BMI mean for each VarPhen PRS quantile. Each quantile contains ~37,000 individuals. The bars correspond to the standard error of the mean.

The individuals from the lowest (score ≤ 0.429) and highest (score ≥ 0.488) quantiles from the *VarPhen PRS* were assigned to a '**low-risk**' and '**high-risk**' group respectively. This corresponded to ~74,000 individuals in total (37,076 for the low-risk group, 37,256 for the high-risk group). The individuals in the high-risk group have more SNPs in common with the VarPhen set, thus are more at risk of developing obesity. Next, the individuals from the low- and high-risk groups were compared to the rest of the individuals in the *backbone* PRS, for each quantile. Individuals from the high-risk group had higher BMI means at each quantile of the *backbone PRS* (see Figure 18). The opposite was observed for the low-risk group. This can be interpreted as the BMI being mostly controlled by the common SNPs of the *backbone PRS*, while some individuals carry a subset of SNPs, more impactful, which significantly change their genetic risk. Ignoring the results from the 2nd PRS would mean a significant under or over-estimation of the genetic risk for these individuals.

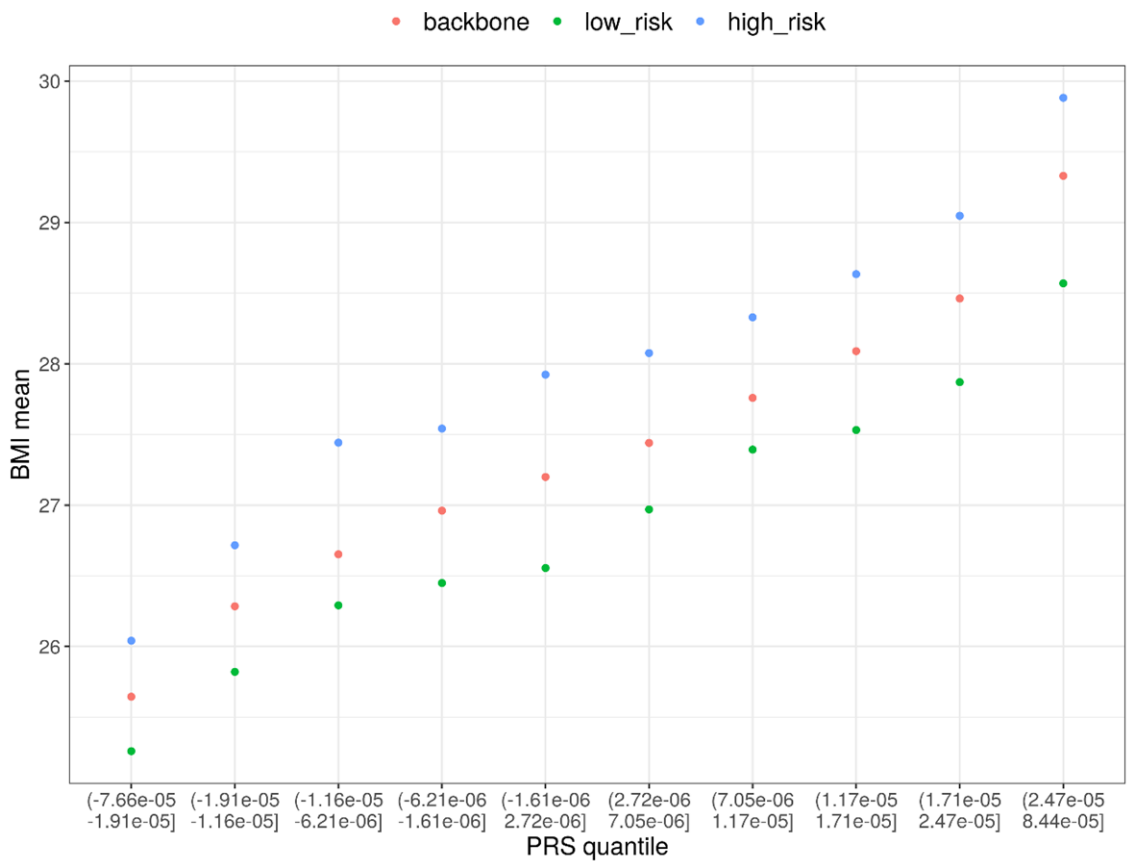


Figure 18: Mean BMI for each backbone PRS quantile (in orange), with the means of the readjusted individuals corresponding to the lowest (in green) and highest (in blue) PRS quantiles of the VarPhen PRS. Each quantile contains ~37,000 individuals, and ~7,500 are readjusted per quantile.

Interestingly, the low- and high-risk individuals are split across the whole 10 quantiles of the *backbone PRS* (see Table 11), showing the complementarity of the two models. However, most of the individuals mapping to a low quantile in the *backbone PRS* are also mapping to a low quantile in the *VarPhen PRS* (the same reasoning applies to the high quantiles). This suggests that the two PRS are correlated despite measuring complementary genetic risks, this is expected, since the two PRS are sharing some predictive SNPs (see 4.4.1).

Table 11: Overlap of the individuals between the backbone quantiles and the VarPhen (VP) quantiles

	VP Q1	**VP Q2**	**VP Q3**	**VP Q4**	**VP Q5**	**VP Q6**	**VP Q7**	**VP Q8**	**VP Q9**	**VP Q10**
Backbone Q1	5475	4716	4318	4184	3643	3536	3520	2906	2767	2274
Backbone Q2	4714	4347	4167	4183	3685	3545	3750	3162	3107	2679
Backbone Q3	4271	4162	4028	4172	3688	3651	3887	3228	3212	3041
Backbone Q4	4032	3898	3941	4039	3610	3630	3928	3407	3470	3385
Backbone Q5	3861	3745	3684	4018	3637	3629	4074	3489	3614	3589
Backbone Q6	3502	3708	3723	3939	3605	3594	4077	3618	3791	3782
Backbone Q7	3286	3583	3614	3972	3670	3701	4096	3536	3873	4009
Backbone Q8	3015	3311	3401	3952	3541	3763	4146	3744	4141	4327
Backbone Q9	2771	3153	3391	3717	3536	3732	4138	3932	4259	4708
Backbone Q10	2412	2730	3058	3523	3309	3616	4375	4011	4767	5539

4.4 Discussion and limitations

4.4.1 Addressing the independence of the two sets

Following a discussion with Dr Alex Gutteridge (Enedra Therapeutics), a point of concern arose concerning the independence of the two models. Indeed, some SNPs are shared between the two base sets (see Figure 19) which might suggest that the readjustment was not genuine but stemmed from running a similar PRS on the same individuals.

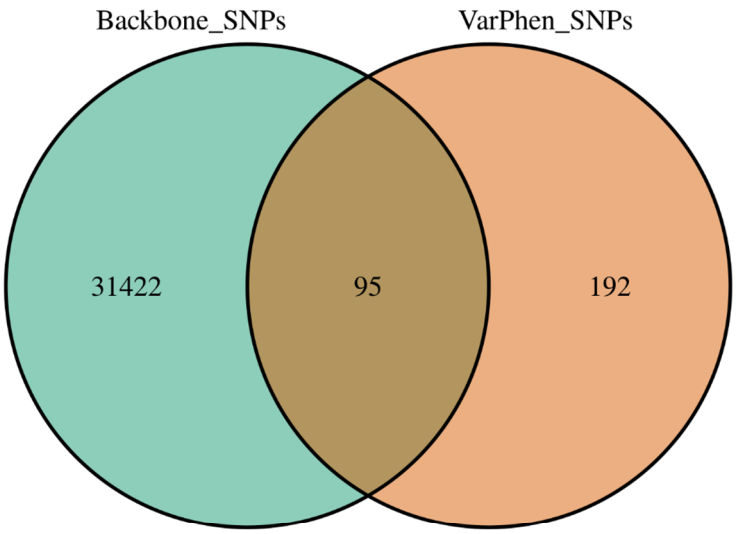


Figure 19: Venn Diagram of the shared predictive SNPs used in the two PRS models

To assess if the readjustment signal was due to an actual effect from the VarPhen SNPs, the same analysis was done but without any SNP in common between the two sets, including SNPs with a LD $r^2 > 0.1$. First, the SNPs from both sets were merged and their LD correlation was calculated with plink using the `--ld-snp-list` and `--r2` options. This resulted in 77 independent SNPs, which were then used as the base set for the *VarPhen PRS*. The same steps as 4.3.2. were applied, to readjust the extreme individuals from the *VarPhen PRS* across the quantiles from the *backbone* PRS.

As seen from *Figure 20*, the readjustment effect did not disappear when the two PRS are built using two independent sets of SNPs, which confirms the genuineness of the readjustment from the *VarPhen PRS*. The observed

readjustment was less impactful, which is expected since the *VarPhen PRS* was built using less SNPs.

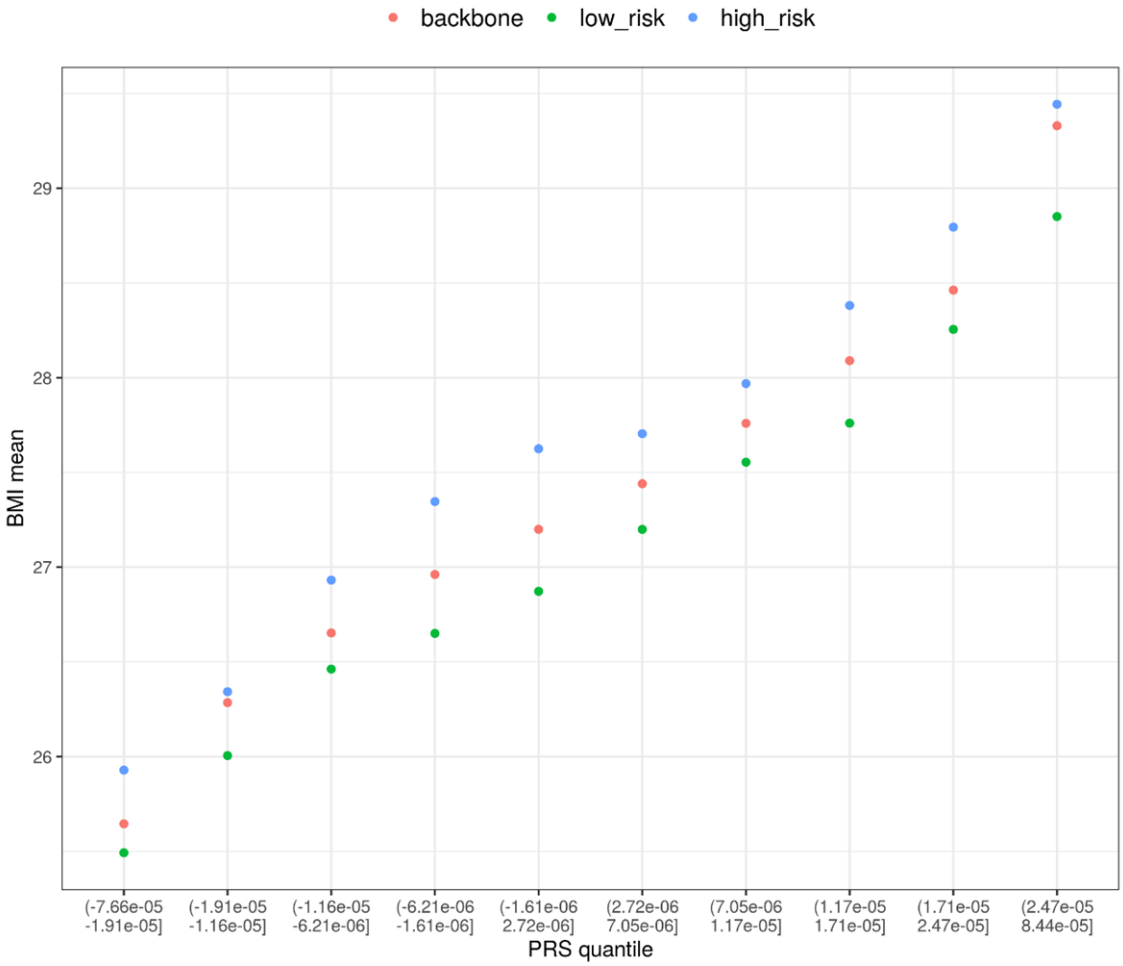


Figure 20:The backbone PRS quantiles with the readjusted individuals corresponding to the lowest (in green) and highest (in blue) PRS quantiles of the independent VarPhen PRS analysis. Each quantile contains ~37,000 individuals, and ~7,500 are readjusted per quantile. Here, the VarPhen PRS base set only contained SNPs that were not in LD with the backbone base set.

The reader should keep in mind that the 2nd PRS does not have to be independent from the 1st PRS. The base set for the 2nd PRS should contain SNPs that are linked to the trait in different databases or the literature, to complement the sets of SNPs obtained from GWAS analyses, regardless of their overlap.

4.4.2 PRS models and pleiotropy

Pleiotropy happens when a locus is affecting two or more unrelated traits. Personalised medicine, including personalised nutrition, will benefit from increased knowledge about pleiotropy, which is currently scarce. Indeed, this will

both help to (i) better understand the genetics of disease and (ii) design better therapeutic strategies, e.g. the repurposing of drugs [165]. Large scale resources and projects, such as *The International Mouse Phenotyping Consortium* will help in this regard [166].

To investigate pleiotropy from the *backbone* PRS model for BMI, the variants that were used to build it were queried against BiomaRt (see 3.3.5), to retrieve other phenotypes they might be involved with. Concretely, the *phenotype_name* and *phenotype_description* attributes from the *Ensembl Variation* Mart were retrieved and subsequently filtered, based on *snp_filter*, using rsIDs from the PRS model as filtering values.

Of the 31,517 queried SNPs, ~10% (3,047) were associated with a phenotype in BiomaRt. The query returned 1,973 unique phenotypes, of which, 65% (1,224) had a single SNP in common with the PRS model. Some of the phenotypes were directly related to body mass index or obesity and were removed, to focus on pleiotropy.

After filtering, 1,880 unique traits were left, and the phenotypes with the most shared SNPs with the PRS model for Body Mass Index were *Height, Cancer* (notably *leukemia*), and some were directly or indirectly related to the metabolic syndrome: *Type 2 diabetes*, *Blood pressure*, *Coronary Artery Disease, Apolipoprotein levels,* and *Cholesterol levels* (see Figure 21). These results could imply the existence of shared genetic risks between the different traits characterising the metabolic syndrome. Future work could focus on assessing if individuals at high genetic risk for obesity are also at high-risk for other metabolic syndromes, if this is the case it could indicate that obesity is not only a risk factor for these diseases but also shares genetic roots with them.

Figure 21: Treemap *of the phenotypes having more than 40 SNPs in common with the Polygenic Risk Score model for Body Mass Index. The phenotypes were retrieved with BiomaRt, using the rsIDs as filters. For each trait, the number of SNPs shared with the PRS is shown in parenthesis.*

Some SNPs are involved with many traits, e.g., *rs1260326* is involved with 145 phenotypes while *rs13107325* is involved with 95. These SNPs might be useful to understand key pathways and mechanisms shared between traits. Accounting for pleiotropy could also help in identifying “subtypes” of risk in PRS. For example, different therapeutic strategies could be designed for someone which is at high-risk for both obesity and coronary artery disease compared to someone at high-risk for both obesity and Diabetes type 2.

4.4.3 Validation with another trait

The two-step PRS approach was validated with another trait: diabetes mellitus type 2.

The base data were generated from the summary statistics of the following GWAS Catalog study: GCST006867. Shortly, Xue et al. [28] performed a GWAS analysis for diabetes mellitus type 2 based on individuals from three datasets: Genetic Epidemiology Research on Aging (GERA), DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) and UK Biobank. Their analysis was based on 659,316 individuals, including 655,666 of European origin, and the raw summary statistics contained 5,052,918 SNPs. The same filtering steps as for BMI were used, namely duplicated SNPs and ambiguous SNPs with complementary alleles (A/T and C/G) were removed, to avoid any possible source of mismatch. After this initial quality control step, 4,280,711 SNPs remained.

The target data are the processed UK Biobank, as described in 4.2.2. For the phenotype, an individual was considered as having diabetes if it was diagnosed by a doctor (data field: 2443). The processed target data contained 17,763 cases and 354,790 controls.

The clumping of the variants (see 4.2.3) resulted in 68,543 variants. As before, a range of different p-value thresholds were applied to find the best model, as assessed by the amount of explained variance. The threshold of 0.4 resulted in the most explained variance and was used to build the PRS model, which was composed of 42,061 variants. Figure 22 presents the results of the PRS, with higher quantiles having higher percentages of individuals with diabetes, showing that the model accurately estimates the genetic risk for diabetes.

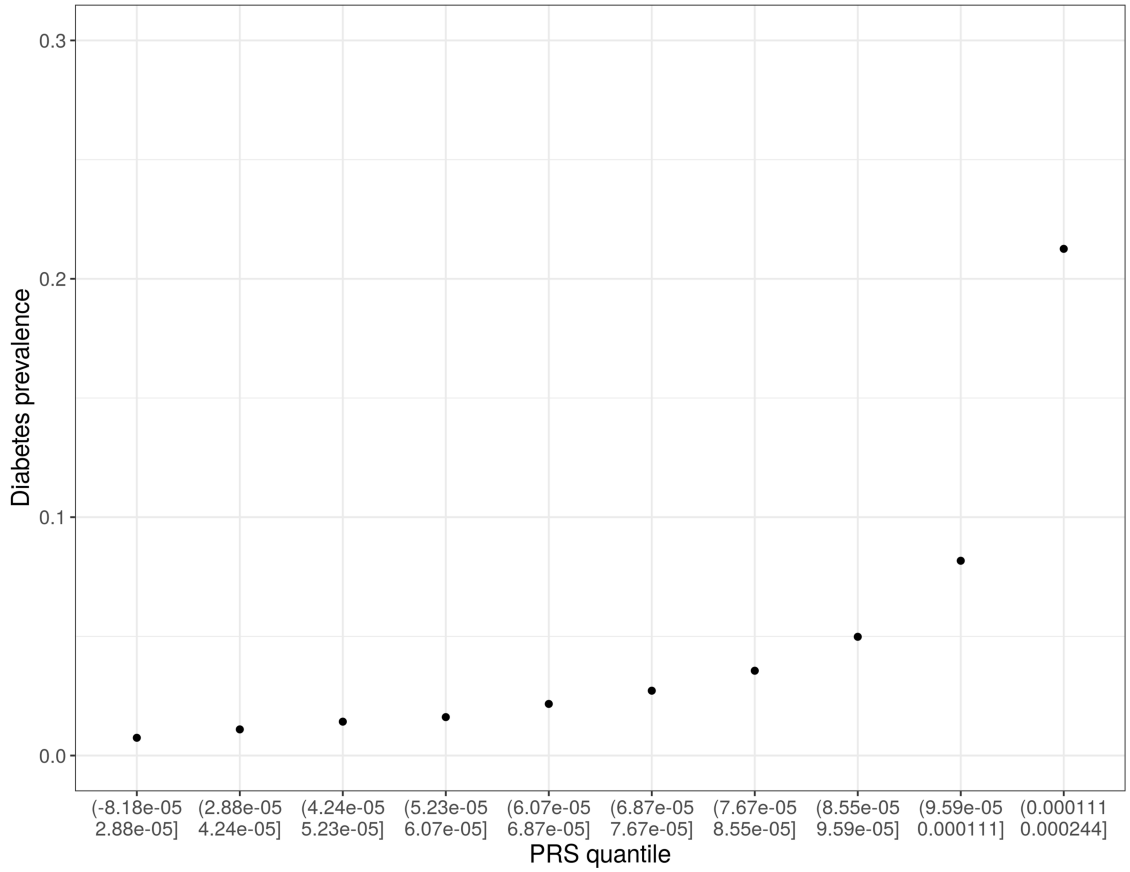


Figure 22: Prevalence of diabetes for each quantile of the backbone PRS. Each quantile contains ~37,000 individuals.

As before, a second, unweighted PRS based on VarPhen was performed. The list of phenotypes entered as input to the pipeline is available in Table 12. This resulted in a PRS model with 290 variants. The prevalence of diabetes for each quantile of this model is presented in Figure 23. As expected, the model based on VarPhen was not as performant as the *backbone PRS*, since it is unweighted and relies on fewer variants. However, this model was still effective for the risk readjustment of the individuals from the extreme quantiles, as will be detailed below.

Table 12: List of phenotypes given as input to VarPhen, in order to get the SNPs related to diabetes mellitus type 2.

List of phenotypes	
Diabetes Mellitus	Diabetes mellitus type 2
Type 2 diabetes	Insulin-resistant diabetes mellitus
Diabetes mellitus type 2 susceptibility to	Type 2 diabetes mellitus
Diabetes Mellitus Noninsulin-Dependent with Acanthosis Nigricans and Hypertension	Diabetes mellitus noninsulin-dependent association with
Diabetes mellitus noninsulin-dependent modifier of	Diabetes type II susceptibility to
Type 2 diabetes mellitus 5 susceptibility to	Diabetes mellitus noninsulin-dependent maternally transmitted

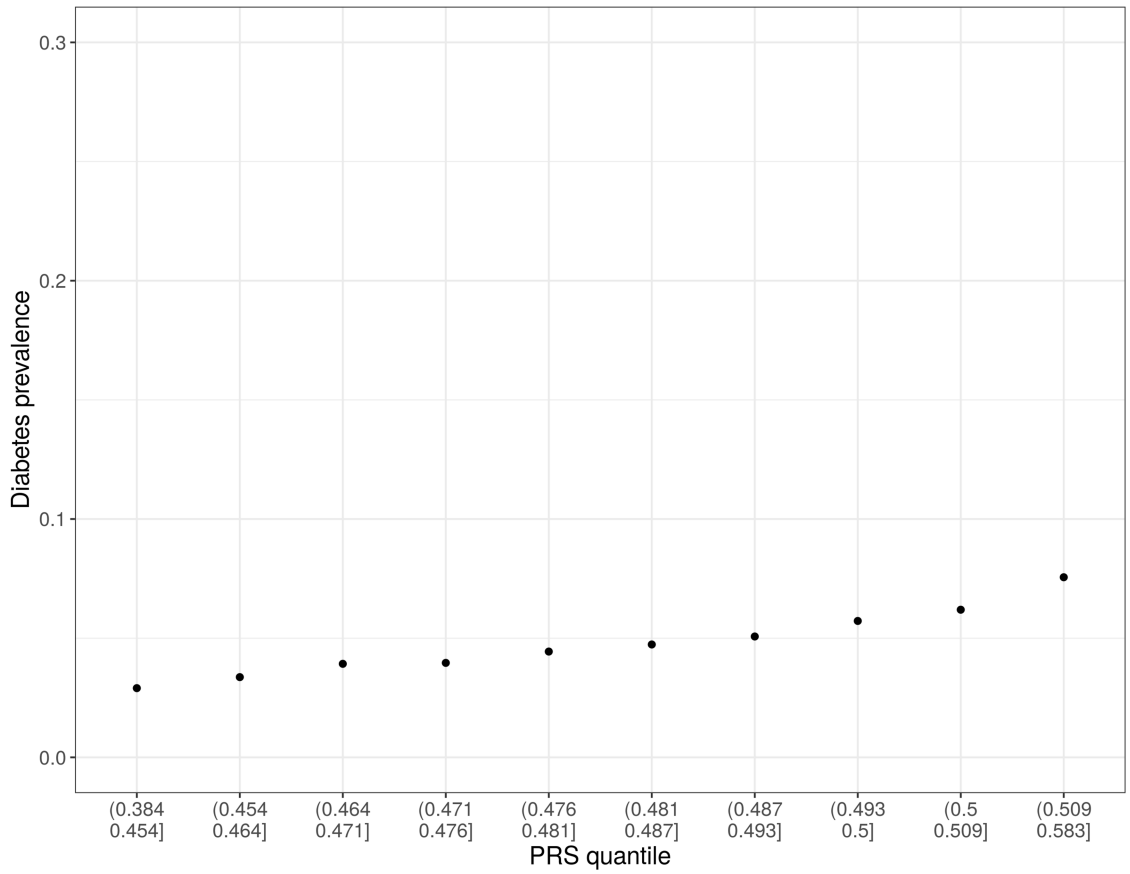


Figure 23: Prevalence of diabetes for each quantile of the VarPhen PRS. Each quantile contains ~37,000 individuals.

As before, the individuals from the lowest (score ≤ 0.454) and highest (score ≥ 0.509) quantiles of VarPhen PRS were assigned to a ‘low-risk’ and ‘high-risk’ group respectively. The prevalence of diabetes was calculated for the low- and high-risk groups for each quantile of the backbone PRS. The results are

presented in Figure 24, as expected the low-risk group has a lower prevalence of diabetes than the whole corresponding quantile, which itself has a lower prevalence than the high-risk group. This is even more significant for the highest quantile.

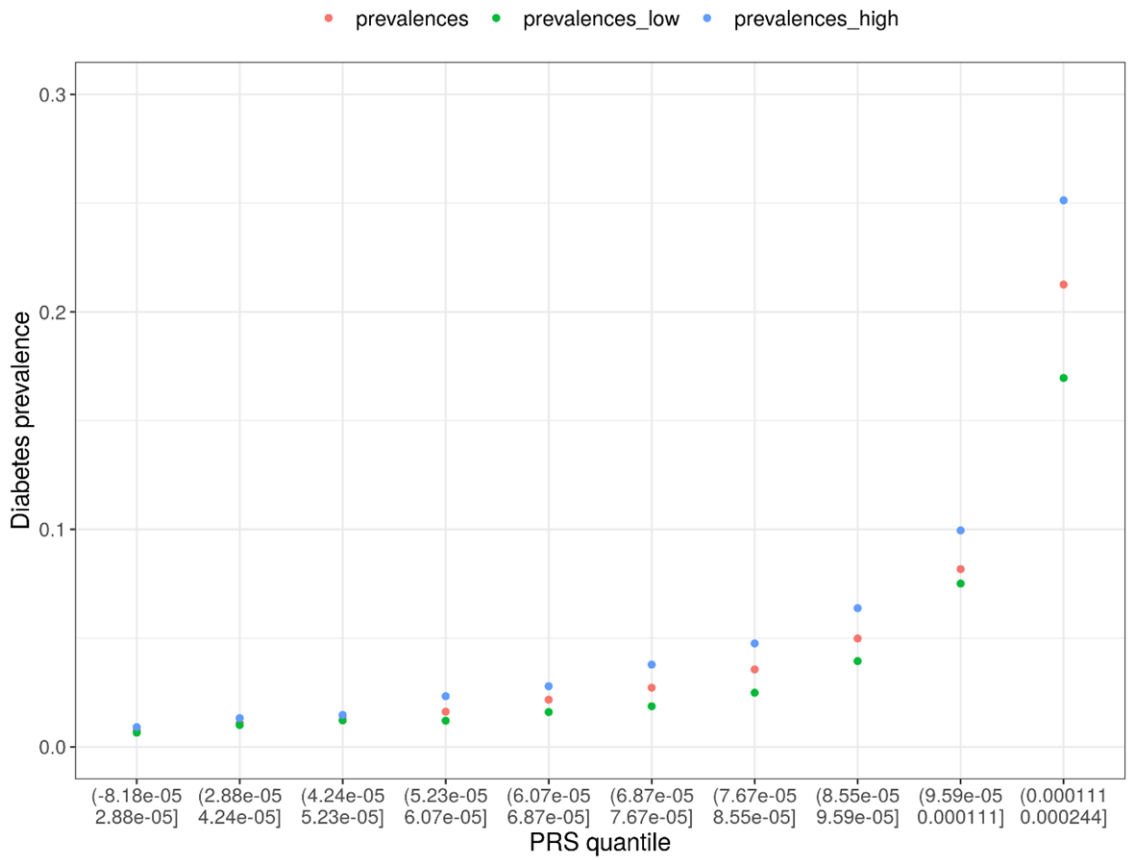


Figure 24: Prevalence of diabetes for each backbone PRS quantile (in orange), with the prevalence of the readjusted individuals corresponding to the lowest (green) and highest (blue) PRS quantiles of the Varphen PRS. Each quantile contains ~37,000 individuals, and ~7,500 are readjusted per quantile.

The results from the *backbone PRS* might be inflated since the base and target sets share some individuals in their analyses (from UK Biobank) [42]. This is a limitation of the *backbone PRS* but does not affect the main argument from this validation analysis. Indeed, these results still demonstrate the validity of the two-step PRS approach to readjust the risk for certain individuals based on a set of confirmed SNPs.

4.4.4 Discussion

The *backbone PRS* based on the GWAS results from Locke et al. [99] provided an accurate estimation of the genetic risk linked to obesity within the UK Biobank. Thus, it can be used as a powerful tool to personalise prevention, for example diet and exercise advice can be given in anticipation to high-risk individuals. However, this PRS alone was not enough to accurately estimate the risk for every individual. To alleviate this, the second PRS, based on the VarPhen SNPs, was developed to readjust individuals across the whole range of risks from the *backbone PRS*. This highlighted the need to integrate variants obtained from different sources to a PRS analysis to make a refined prediction. Nevertheless, despite all our efforts, any PRS on obesity will remain imperfect until we have a complete understanding of the genetics behind obesity and body weight control.

However, the reader may wonder, why not simply merge the two PRS instead of creating a two-way scoring system? As we have seen, the *VarPhen PRS* is a weak predictor of the genetic risk (both for BMI and diabetes), so merging the two PRS would only render the *backbone PRS* less accurate. But the *VarPhen PRS* is still useful to identify extreme individuals which might benefit from a readjustment. For diabetes, this translated to almost a doubling of the risk between the low- and high-risk groups for the highest quantile (and this quantile is the most interesting for prevention purposes). This shows the great benefit of refining the prediction instead of just assigning one score. For BMI the readjustment was more linear, but individuals from the low- and high-risk groups have BMI means similar to those of the neighbouring quantiles, sometimes even more.

This PRS model can only be applied to data generated through the Nutrishield project's clinical trials, if the genetic risk associated with BMI is stable across the lifespan of an individual. Indeed, Nutrishield is focusing on children, while the UK Biobank is composed of individuals aged between 40 and 70. Fortunately, results from several studies suggest a stable genetic component of BMI [164] [167], meaning that genetic risk estimation is transposable between different age groups.

A similar approach of combining the predictions from rare pathogenic variants and PRS was successfully applied previously, for BMI [168] and prostate cancer [169]. Here, the methodology differs in the fact that we are not manually gathering rare pathogenic variants from the literature, but trait-associated variants from public databases, both approaches having pros-and-cons. Our approach is more straightforward to implement and gather more variants but may lack the finesse of manually selecting variants.

For future studies, this method would benefit from being tested and validated against other cohorts and traits.

4.4.5 Limitations

Limitations pertaining to PRS models in general have been described in 0, here we will focus on the limitations specific to this analysis.

The utilisation of the *VarPhen PRS* carries two limitations. First, the variants retrieved are limited to those present in the public databases, thus this approach works best for well-studied diseases. Second, VarPhen does not return the effect size of the SNP, which constraints us to an unweighted PRS. Moreover, all SNPs are considered pathogenic, even those who are protective. *Note*: the last point can be partially avoided by removing SNPs annotated with a ‘protective’ clinical significance.

There is another limitation spawning from the base set, which was performed mostly on individuals of European ancestry. Since the causal variant is unlikely to be directly genotyped, GWAS only identify variants that are in Linkage Disequilibrium with it (see 2.1.3.5). And as this pattern of associations between variants is population specific, this prevents the generalisability of this PRS model to other ethnicities. This can lead to exacerbated inequalities in care [170]. Indeed, when the PRS model was applied to the other ethnicities present in the UK Biobank, there were a lot of discrepancies between the risk score and the actual BMI (see Figure C.2-1). Only GWAS performed on a variety of ethnic backgrounds or using Whole Genome Sequencing can alleviate this issue.

The sex chromosomes were ignored in this analysis, as (i) they are particularly challenging to analyse and might lead to misinterpretations [171]; and (ii) the GWAS meta-analysis used as the base data only contained six variants on the X chromosome. However, the PRS models presented here would benefit from including variants from the sex chromosome, according that proper care is given to the statistical analysis and interpretation of their effect on BMI.

4.5 Conclusion

Here, a PRS model was developed which successfully identified the genetic risk associated with BMI in the UK Biobank population. The model was expanded with the use of a second, unweighted PRS, which helped in refining the risk of developing a higher or lower BMI in a subset of individuals.

This approach can be easily translated to any phenotype queried using VarPhen, as demonstrated with diabetes mellitus type 2. And as our knowledge about different traits will grow, so will the content of public databases, which will further strengthen this two-step approach.

The release of Whole Genome Sequencing (WGS) data for the UK Biobank participants, planned for 2022, will allow to overcome most of the limitations described previously. For now, the PRS model is limited by the incomplete overlap between the variants obtained from the GWAS and the UK Biobank arrays. With good quality WGS, it will be possible to account for every variant and the PRS will only be bounded by our understanding of the disease. If future GWAS are also based on WGS, then the benefits would add up to allow for a virtually perfect estimation of the genetic risk associated with BMI, for any population.

5 Conclusion and thoughts on the use of genetics for personalised nutrition

5.1 Conclusion

Personalised nutrition currently remains at its early stages of development but shows the potential of improving the health of the general population, at a time when diabetes and obesity are becoming worldwide epidemics. However, it will need to be based on rigorous scientific research, as well as being accompanied by public policies and ethical considerations.

Understanding the impact of the diet on health involve diverse biological pathways, which can be studied at different levels. Certain risk factors originate from the genome, others appear in the metabolome, microbiome or at the protein level. Sometimes, these different layers also interact and influence each other, rendering the picture even more complex. Even when considering a single knowledge level, such as genomics, the data and findings are often stored in different formats, making their interpretation challenging. Creating standards and tools to merge these datasets will be necessary if we want to harness the full potential of what is, and will be, available in public databases. We hope that VarGen will be useful as an easy-to-use package to merge data from OMIM, GTEx, GWAS and FANTOM5, facilitating the study of the impact of variants on disorders.

Polygenic risk scores (PRS) have a high potential for both improving prevention and furthering our understanding of diseases. The models will become stronger and more accurate as our molecular knowledge about diseases will increase. As shown in this thesis, PRS models can be improved by including variants from different sources, not only GWAS results, and tools such as VarGen can help in this regard. The concept of PRS can be further improved for the study of endotypes. Indeed, some studies are developing ‘partitioned PRS’, where SNPs are categorised by mechanisms, which helps in categorising the risk according to different facets of each disease. A similar concept was elaborated by McCarthy in his ‘palette’ of diabetes [172]. Instead of compartmentalising diabetes into

discrete endotypes, the idea is to list pathways affecting diabetes pathogenesis and assess each patient according to each pathway. Some patients will have a high risk in one category, akin to a monogenic form of diabetes, while others will have moderate risks across a range of different pathways. Indeed, some endotypes of diabetes are sharing common risk factors, and in such cases the boundaries between discrete endotypes become blurred. Prevention and treatment can be tailored per affected pathway instead of per diabetes type, moreover, it is easier to design a drug targeting a precise malfunctioning biological process than a whole endotype. The main challenge with this approach is to identify and define the exhaustive list of pathways that will compose such a ‘palette’.

Fortunately, ambitious projects, such as GTEx, UK Biobank or large GWAS are really pushing our genetic knowledge of diseases forward. These projects allow for research on a scale that was never seen before and give researchers a strong foundation to test hypotheses and obtain statistically solid results.

5.2 Limitations

VarGen is limited by the content of public databases. The output of the main pipeline can also be overwhelming, as hundreds of thousands of variants are retrieved for complex diseases. There is also a need for experimental validation of the variants of interest detected by VarGen, as the link between a variant and a disease is hypothetical.

Future versions of VarGen could benefit from novel functionalities and improvements to existing ones. Adding an SQLite database would allow VarGen to manage complex data types, instead of just outputting a table of variants and annotation. For example, variants retrieved from the GWAS Catalog could have information about the study they originate from, their associated p-values and effect size. The annotation of variants could also benefit from the SQLite database, with detailed information about each field, for example the ClinVar significance could be accompanied by information about supporting evidence (e.g.: publications, submitters).

Other datasets could be added to VarGen's pipeline, for example the Genome Aggregation Database (gnomAD) [173] is a project aiming at aggregating and harmonising genome sequencing data from an assortment of studies and consortiums. GnomAD contains a short variant data set, with the latest version, v3.1, containing 76,156 genomes aligned against GRch38. VarGen could use this database to obtain Allele Frequencies for SNPs in different populations. This could help to study the impact of loss-of-function mutations on genes, as these mutations are often deleterious and thus sustained at a low-frequency in human populations [173].

VarGen would benefit from integrating information from developed PRS models. More specifically, for a certain trait, it would be interesting to compare the variants retrieved by VarGen and those used in one or more PRS models for the same trait. The variants retrieved by VarGen would have the added information about their predictive relevance (as part of one or more PRS models) and the variants effect size, while the variants from the PRS models would have added information from the annotation retrieved by VarGen. This could facilitate in-depth analyses of PRS models to identify the most relevant variants and/or help in identifying the variants that classified an individual in a high-risk group. Indeed, we can suppose that some high-risk individuals are in this category because of the accumulation of a lot of small impact variants, while others might have less but more impactful variants (and all the other combinations in-between). Integrating data from genotyping, annotation from VarGen and PRS models could allow this kind of in-depth analysis of genetic risk. The PGS Catalog, a database of more than 2,000 PRS models for 535 traits, is accessible via a REST API and could be integrated within VarGen's pipeline [174].

There are several limitations associated with PRS models. First, they are not self-sufficient as clinical predictors, their strength is in complementing other risk factors. Secondly, since most of the base datasets available are based on genotyping chips, the variants detected are not causal. This hinders both the interpretation of the PRS models to understand the genetic roots of diseases and the generalisation of the models to other populations. The estimated risk is also

relative within the target set; all these limitations does not allow the use of PRS in a clinical setting yet.

For both VarGen and the two-step PRS approach, one of the limitations is the lack of consideration of structural variants, such as copy number variations or large deletion, insertion, or inversions. There is increasing evidence of the impact of these type of variants in gene regulation and the pathogenesis of certain diseases. But, as with non-coding SNPs, the difficulty lies in the interpretation of the variants' consequences.

5.3 Thoughts on the future personalised medicine/nutrition

The field of precision medicine is moving forward rapidly. Diseases are more finely described, notably via the identification of endotypes. This will translate into better understanding of these diseases, which in turn, will translate into better management. Going further, research should focus on considering complex diseases as continuous, as there are often overlaps between endotypes and such categorisation is sometimes not enough to accurately represents the reality of the underlying biological processes. In addition, there is a growing interest in finding shared pathways between diseases, with the aim of finding shared targets, allowing the repurposing of existing drugs.

PRS models will play an important role in the future of personalised medicine. Especially as sequencing will become more accessible and less expensive. Despite their limitations, presented above, they remain a powerful tool to both understand diseases and performing informed prevention. The approach of partitioned models will increase the potential of PRS to be used as tools to comprehend disease endotypes. Moreover, the use of whole genome sequencing to perform GWAS will solve most of the issues mentioned above, as it will allow the detection of causal variants, instead of just variants in LD with the causal variants. Knowing the causal variants will make the PRS models more accurate and generalisable across populations, which will increase their clinical usefulness.

However, PRS models on their own will probably still be not enough to make a reliable estimation of risk. There will be a need to merge data from multiple sources, such as clinical data, microbiome sequencing, or multi-omics measurements. Multi-omics approaches are increasingly being used in the field of personalised medicine, and tools are designed to handle this heterogeneous input, such as the Multi-Omics Factor Analysis tool [175]. This increasing complexity and heterogeneity in data and analysis highlight the need of a cross-disciplinary approach to science, and the importance of bioinformatics to bridge the gap between the different fields. As data will become more diverse, tools such as VarGen will be needed to merge information from different sources and draw a clear picture of the links between the different layers of -omics and the phenotype.

Projects, such as Nutrishield, which include a wide variety of data, will be necessary to advance the field. The original aim behind the development of VarGen and the PRS models, was to apply and validate them through the Nutrishield project's clinical trials. Unfortunately, due to Covid-19 and the resulting delays in the project, it was not possible to do it throughout the course of this thesis. Applying and validating VarGen and the PRS models through the clinical trials remains a core objective of future work based on this thesis.

More precisely, one important point to move forward is the interpretability of non-coding and structural variants. Personalised medicine will greatly improve if tools and methods are developed to reliably assess the impact of such variants on biological pathways. This, for example, will avoid the misinterpretation of variants impacting distal genes, such as *rs1421085 T-to-C* on *FTO* which over-expresses *IRX3* and *IRX5* in obesity (see 2.3.4.2).

As the diabetes and obesity epidemics will continue to progress, the necessity for personalised nutrition will increase. It can be a key element of the solution to curb the rise in cases and keep the general population healthy. However, it will be not enough on its own, and will need to be accompanied by public policies to render our societies less obesogenic, even if more research is needed to fully understand the intricacies of the impact of our environments on obesity [176].

There will also be ethical questions to debate. Should we sequence everyone, or just individuals suspected with genetic diseases? WGS is showing promises towards better care, especially for rare and acute diseases. A study from the *NICUSeq Study Group* demonstrated the impact of WGS in getting a change of management faster for acutely ill infants [177]. Moreover, rare diseases can sometimes leave families in a ‘diagnostic limbo’ for years, and sequencing has the potential to shorten this period of uncertainty, via genetic diagnostics. While, using WGS for rare diseases is generally accepted and show great potential, the same cannot be said for ‘general prevention’, especially for complex diseases. First, the genetic background of most complex diseases is not well defined yet. Indeed, they are often due to many variants, providing varying degrees of risk, including protective variants. The results of broad genetic tests can therefore be difficult to interpret. Second, the genotype is not the only factor affecting disease risk, the environment, lifestyle, microbiome, often play an equal or even greater role than the genotype, which add another layer of potential misinterpretation of the results. Third, there is the ethical question of shared family risk, i.e., “when we obtain a positive result, should family members at risk of developing the same disease, be warned?”. This is a complex question involving the patient’s rights to privacy, the family members’ rights to know, and the burden/benefice ratio of informing people about their genetic risks, especially for debilitating diseases [178]. Finally, genetic data are very sensitive by nature, and it is necessary to protect the confidentiality of genetic results, to avoid discrimination by insurance companies, employers, and society in general [178].

Thus, precision medicine and nutrition will become very important in the next decades, but they should be rigorously implemented, need collaboration from multiple scientific fields and should be accompanied by societal and ethical considerations, to be exploited to their full potential.

6 MicroRNA differential expression analysis for Hypoxic-Ischemic Encephalopathy

Throughout the course of my PhD, I performed a differential expression analysis in collaboration with one of the Nutrishield partners, *Hospital Universitari i Politècnic La Fe*, located in Valencia, Spain. The aim of this study was to identify microRNAs that could serve as biomarkers for early detection of severe cases of Hypoxic-Ischemic Encephalopathy (HIE). This chapter will provide a short literature review about miRNAs and a description of the differential expression analysis.

6.1 Background on microRNAs

6.1.1 Definition and discovery

MicroRNAs (miRNAs) are small non-coding RNAs, about 20~24 nucleotides long. They are too short to code for proteins but have an important role in gene expression. It is estimated that they regulate at least 60% of the human genes [179] [180]. The first miRNA was discovered in the nematode *Caenorhabditis elegans* in 1993. Named *lin-4*, it regulates the *lin-14* messenger RNA, which is involved in the transition from the larval stage 1 to the larval stage 2 [181]. The second one was found seven years later, also in *C. elegans*, labelled *let-7*, it regulates the *lin-41* messenger RNA (mRNA) which is involved in the transition from the larval stage 4 to the adult stage [182]. The same year, *let-7* was also found in other animals (flies, humans, zebrafish…) which suggested a conserved role of this miRNA across animal phylogeny [183]. In 2001, many more of these small RNAs were found, in different species [184] [185] [186], and the term ‘microRNA’ was coined to label them as a different class of non-coding RNAs.

6.1.2 The miRNA biogenesis in animals

In animals, the miRNA biogenesis starts in the nucleus. The DNA locus containing the miRNA is transcribed into a primary transcript. Then the *Drosha* protein cleaves the primary transcript to generate the pre-miRNA hairpin structure, an example of hairpin is available in Figure 25. The hairpin is transported outside the nucleus, where the *Dicer* protein removes the loop from

the hairpin. At this stage, the miRNA consists of a miRNA duplex containing two strands: 5' and 3' (alternatively called *mature* and *star*). One of these strands will associate with the *Argonaute* protein to form the *RNA-induced silencing* complex, which will be guided to its target mRNA and influence its expression [187]. Hence, the miRNAs can be studied in two different forms: the pre-miRNA hairpin (or precursor) and the mature sequence.

Figure 25: Example of a miRNA hairpin (hsa-miR-3134). After the cleavage by Dicer, the mature sequence, highlighted in red, will form the silencing complex with the Argonaute protein. Figure generated with miRDeep2.

6.1.3 The regulation of mRNAs by miRNAs in humans

First, the *silencing complex*, formed by the miRNA mature sequence and the *Argonaute* protein, pairs with the mRNA. The pairing is based on the 'seed' of the mature sequence (nucleotides 2 to 7) and the target site on the mRNA, often located in the 3' untranslated region. As mentioned in Section 6.1.1, more than 60% of the human genes are regulated by miRNAs, and the mRNAs are under selective pressure to conserve these pairing sites [179], which denotes their importance. In animals, the miRNAs are affecting post-transcriptional regulation of gene expression via two main processes: (i) reducing the stability of mRNAs and (ii) hindering the translational machinery.

(i) Once the *silencing complex* is attached, it recruits the *TNRC6* (also called *GW182*) protein which interact with the *poly-A binding protein (PABP)* from the mRNA poly(A)-tail, resulting in the destabilisation of the mRNA and decapping of the 5', thus enabling the degradation of the mRNA. In addition, the *silencing complex* will recruit the *deadenylase complex* to speed up the degradation [188].

(ii) Translation inhibition by the *silencing complex* involves different mechanisms: *TNRC6* competes with *eIF4G,* thus disturbing the circularisation of the mRNA, which is an important component for an efficient translation; in parallel the *silencing complex* associates with *eIF6* to prevents the formation of 80S ribosomes by blocking the combination between the 60S and 40S ribosomes [189] [190].

MiRNAs can also alter the regulation of genes via another process, called *site-specific cleavage*, however this requires a perfect match with the target mRNA and is much less common in mammals.

The number and efficiency of the different target sites of a specific mRNA, coupled with the fact that some miRNAs are specific to certain cells, allow for complex and nuanced patterns of gene regulation by miRNAs [191].

6.1.4 The impact of miRNAs on development and health

MiRNAs are regulating a wide range of genes, which makes them involved in many different processes [192]. miRNAs are essential to the development of organisms. As mentioned before (see Section 6.1.1), *lin-4* and *let-7* are regulating the transition between larval stages in *C. elegans* [181] [182]. Removing the *Dicer* protein from zebrafish, thus altering the processing of miRNAs precursors, stopped brain morphogenesis in embryos [193]. MiRNAs play a role in certain diseases, non-exhaustive examples include, mutations in the seed region of *hsa-miR-96* causing progressive hearing loss in both human and mouse [194] [195] and cancer related miRNAs, either promoting tumour growth [196] or acting as tumour suppressors [197].

6.1.5 IsomiRs

Many miRNAs have alternative forms, called isomiRs, which differs by one or more nucleotides at the 5’ or 3’ end of their sequence. These isomiRs are generated by a variability in the cleavage performed by *Dicer* or *Drosha* [198]. The human *trans-activation response RNA binding protein* has been shown to generate isomiRs that are one base longer than the canonical sequence, which can affect guide strand selection and thus mRNA targeting [199].

6.1.6 The bioinformatics of miRNA: tools and challenges

Many bioinformatics tools have been developed in recent years to answer for the growing interest in the study of miRNAs [200]. This part will focus on the tools used for differential expression analysis.

Several databases have been created to store the knowledge gained on miRNAs. One of the most comprehensive is *miRBase* [201], which contains ~50,000 miRNAs for 271 organisms. Each entry corresponds to a predicted precursor, with information about the locations and sequences of the mature forms. The database is available in fasta format, which facilitate its inclusion in analysis pipelines.

Several tools have been developed to quantify miRNAs from high-throughput sequencing data, such as *miRDeep2* [202]. One should also keep in mind that different precursors can result in the same mature sequence (see Figure 26), in which case the strategy to use for quantification differs. If the interest lies in the precursor sequences, then they should be considered as separate entities, however if one is interested in the mature sequences then their read counts from all the corresponding precursors can be averaged.

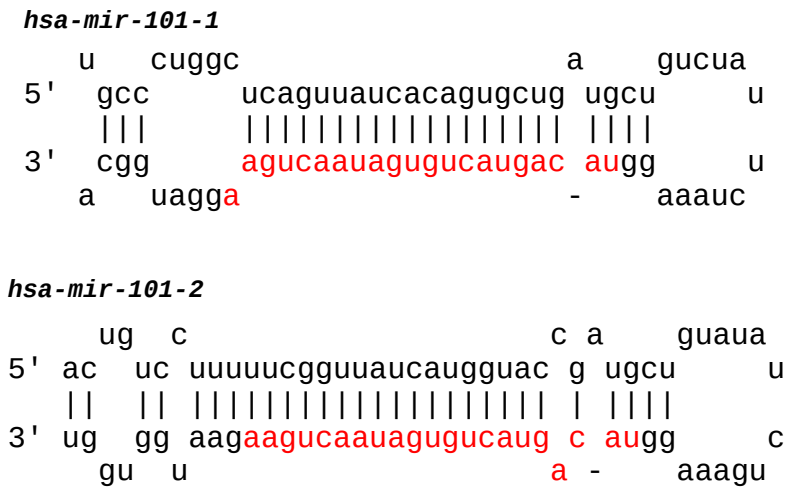


***Figure 26: Sequences of miRNA precursors hsa-mir-101-1 and hsa-mir-101-2*. The mature sequences of both, here highlighted in red, are the same.**

It is becoming standard to identify novel miRNAs before running the quantification step. The tool *miRDeep2* can also be used for that purpose. To understand the

biological implications of novel miRNAs, it is possible to perform target prediction based on the novel miRNA sequence, for example with *miRDB* [203]. For the differential expression analysis itself, the same tools as used for standard RNA-seq are suitable, notably the popular R packages *EdgeR* [204], *limma* [205] and *DESeq2* [206].

One of the major bioinformatics challenges in miRNA analysis is the alignment of reads against a reference. Indeed, aligners are not designed to align such short sequences, and in addition, many miRNAs are part of families with very similar sequences. This can be alleviated with the use of specific aligners and alignment parameters [207].

6.2 Differential expression analysis of miRNA in neonates with Hypoxic-Ischemic Encephalopathy

6.2.1 Introduction

Hypoxic-Ischemic Encephalopathy (HIE) is a brain injury caused by a lack of oxygen and glucose delivery to the brain. The prevalence of HIE is around 2-3 per 1000 live births, rising up to 26 per 1000 in developing countries [208]. The consequences of HIE are serious, with a quarter of the affected neonates dying in intensive care units, this is the leading cause of death and morbidity in newborns worldwide. Severe HIE can also have long term effects on the motor-sensory functions, inducing visual, behavioural and auditory problems as well as seizures and cerebral palsy [209]. Infants with moderate HIE will develop normally, until they reach 2 years of age, where half of them will develop a disability. Beside, all children with HIE, notwithstanding the severity, are at risk of academic problems [210]. In preterms, HIE progression is more complex and tends to lead to worse outcomes [211].

The aetiologies of HIE are multiple but can be arranged in three main categories: *repeated hypoxia,* representing 65% of the cases, *acute catastrophic asphyxia,* representing 25% of the cases and *chronic hypoxia*, representing 10% of the cases (see Figure 27). The first two happen at birth, while chronic hypoxia is antenatal; in some cases HIE is due to a combination of these insults [210].

The brain injury does not occur at the onset of the insult but happens after a cascade of events happening in distinct phases, spanning days or even weeks. Table 13 describes the different phases of HIE with their own time frame and consequences [208] [210]. Interestingly, when the hypoxia is not severe, the brain can reduce its energy consumption to avoid energy depletion and the resulting brain cell damage.

Table 13: Description of the Hypoxic-Ischemic Encephalopathy phases

Phase name	Time frame	Consequences
Acute phase	At the time of the insult	Decreased cerebral blood flow, inducing a decrease in adenine triphosphate, (i.e.: energy failure), leading to cellular damage and ultimately cell death.
Latent phase	1 - 6 hours after the insult	Brain cells undergo partial recovery. This is the phase where therapeutic hypothermia is most effective. In parallel, inflammation appears and there is a continuation of the apoptotic cascades initiated during the acute phase.
Secondary phase	6 – 48 hours after the insult	Happens in neonates with moderate to severe injury. Induced by failure of mitochondrial activity, which lead to further cell death, cytotoxic edema, excitotoxicity and clinical deterioration (e.g.: seizures).
Tertiary phase	Months after the insult	Characterised by latent cell death, remodelling of the brain and astrocytosis.

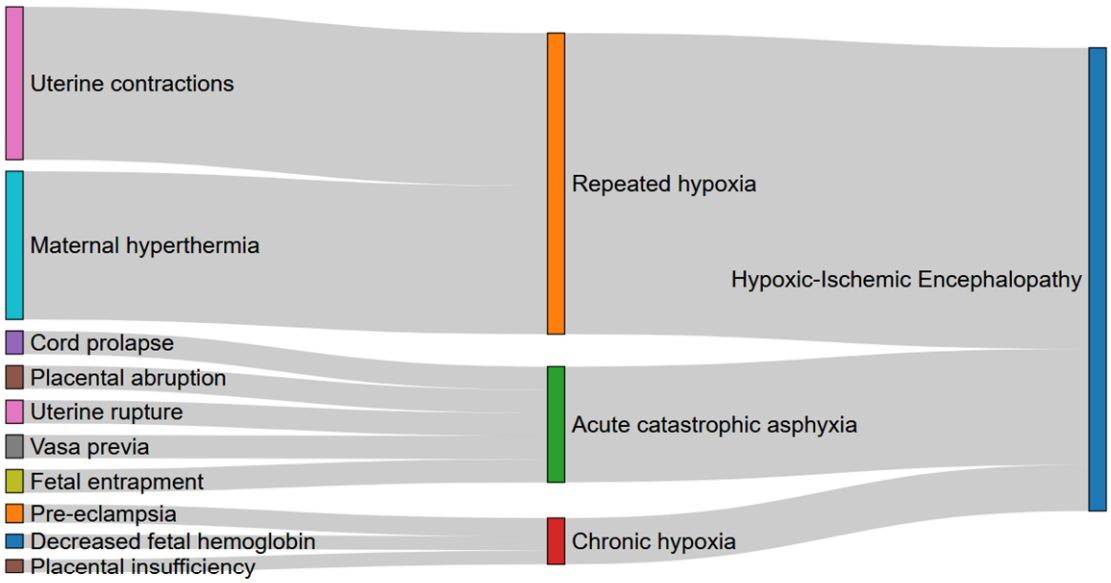


***Figure 27: Sankey plot representing the aetiology of neonatal Hypoxic-Ischemic Encephalopathy. The first column represents the events that can lead to hypoxia (without relevant proportion). The second column represents the three main type of hypoxia that can result in HIE (with relevant proportion). Data derived from Gunn et al.* [210].**

When HIE is suspected, the standard of care is therapeutic hypothermia. It offers neuroprotection during the secondary phase of HIE, thus limiting the mortality and disability rates by 18 months of age and might even offers long term protection [210]. However, this is merely a supportive treatment, and it only offers partial protection. Despite the recent progress in care, the main hurdle is still the identification of infants that would benefit from a treatment [208] [212]. Furthermore, another challenge is the limited 'window of opportunity' for the treatment, as it should be administrated before the secondary phase of HIE. That is why a reliable biomarker of brain injury is needed. The current biomarkers for HIE are the signs of exposure to hypoxia-ischemia, i.e. foetal heart rate changes, oxygen debt on cord blood gases and the Apgar score [212]. An MRI scan can confirm the brain damage. Unfortunately, these biomarkers have limitations; first, they mostly detect severe cases, which are already obvious to identify, second they are efficient after the latent phase, which is too late to provide an efficient treatment [212].

An review from 2019 analysed 323 articles to review the role of miRNAs in newborn brain development and HIE [213]. They identified miRNAs involved in four major processes, *brain development, HIE, neuronal cell death* and *neuroinflammation.*

6.2.2 Materials and methods

6.2.2.1 Library preparation and sequencing

A total of 120 samples were sequenced for the differential expression analysis. These were split into three groups:

- Normal: 12 HIE patients with **normal** Magnetic Resonance Imaging (MRI) outcomes, sequenced over 4 time points (0, 24, 48 and 72 hours).
- Pathological: 12 HIE patients with **pathological** MRI outcomes, sequenced over 4 time points (0, 24, 48 and 72 hours).
- Control: 12 **control** samples, sequenced over 2 time points (0 and 48 hours).

The RNA for time 0 was extracted from umbilical cord blood, the rest of the time points were extracted from plasma.

The Ethics Committee for Biomedical Research of the Health Research Institute La Fe (Valencia, Spain) approved the protocol involving the recruitment of the control group of healthy term infants (2019/0312) as well as the clinical trial, which is registered under the acronym HYPOTOP (EudraCT 2011-005696-17). The HYPOTOP trial is a randomized, controlled, multicentre, double-blinded clinical trial for assessing the efficacy of topiramate vs placebo as an adjuvant therapy in newborns with neonatal encephalopathy undergoing therapeutic hypothermia. A stringent study protocol and written standard operating procedures were followed at all 13 participating sites. All methods were performed in accordance with relevant guidelines and regulations and informed consent was obtained from legal representatives of infants. A detailed description of the study design and inclusion and exclusion criteria of the HYPOTOP trial can be found elsewhere [214].

DNA extraction was performed at the *Hospital Universitari i Politècnic La Fe* in Spain, using the commercial kit *miRNeasy*. First, 250 µL of *QIAzol Lysis Reagent* was added to the blood/plasma sample, and the collection tube was left at room temperature (15-25°C) for 5 minutes. Then, 3.5 µL of *miRNeasy Serum/Plasma Spike-In Control* (1.6 x 108 copies/µl working solution) was mixed with the lysate. 50 µL of chloroform was added to the tube, which was then vortexed for 15 seconds. The tube was left at room temperature for 2-3 minutes before being centrifuged at 12,000 x g at 4°C for 15 min. The aqueous phase was transferred to another collection tube. Then, 150 µL of 100% ethanol was added and mixed to the sample. Finally, the sample preparation was pipetted to a *RNeasy MinElute* spin column and centrifuged at 8000 x g for 15 seconds at room temperature.

The library preparation and sequencing were performed by *Novogene*, on an Illumina NovaSeq™ platform. First, the RNA underwent quality control; *Nanodrop* was used to measure preliminary RNA, followed by an agarose gel electrophoresis to test for degradation and potential contamination and finally, integrity and quantitation were measured with *Agilent 2100*. Then, the library was

constructed with the *Small RNA Sample Pre-Kit*; the final cDNA library was ready after a round of sequencing adaptor ligation, reverse transcription, PCR enrichment, purification, and size selection. The cDNA library also underwent quality control, in three steps. First, the preliminary library concentration was measured with *Qubit 2.0*, second, the insert size was tested with *Agilent 2100* and finally the library concentration was precisely measured with Q-PCR. Due to limited RNA weight, 18 samples did not pass the QC thresholds and had to be re-sequenced in a different batch.

The filtering of the raw reads was also performed by *Novogene*. The reads were removed if one of the following rules applied: more than 50% of bases have a quality score lower than 5; Ns are accounting for more than 10% of the bases; the read has 5' primer contaminants; the read does not have 3' primer or insert tag; the read has a polyA/T/G/C tail. For all the remaining reads, the 3' primer sequence was trimmed. The number of reads available for each sample, before and after filtering is described in Table D1-2.

6.2.2.2 Bioinformatics analysis

The filtered FASTQ files were quality controlled with FASTQC [17]. Some samples had their FASTQ files filled with a PCR primer instead of reads. The PCR primer sequence was as follow:

CGCGACCTCAGATCAGACGTAGATCGGAAGAGCACACGTCTGAACTCCAG

The samples with >90% of their reads corresponding to this primer were removed, as they would not provide any useful information and might even bias the results. The list of samples removed are list in Table 14.

Table 14: ***List of samples removed from the analysis due to a high number of PCR primer contamination.***

Group	Time point	Number of samples (after filtering)	Filtered samples
Control	0h	10	AC1, AC5
	48h	12	/
HIE - normal	0h	9	A41, A56, A74
	24h	10	A41, A56
	48h	9	A41, A56, A158
	72h	10	A95, A158
HIE - pathological	0h	12	/
	24h	12	/
	48h	12	/
	72h	12	/

Prior to the alignment, the latest human reference genome, *GRch38*, was indexed using *bowtie* v1.1.1. The reads were aligned with the *mapper.pl* module from *miRDeep2* v2.0.1.2. The input to *mapper.pl* were given as a config file (option `-d`) listing all the reads as fasta files (option `-c`), each read was allowed to map to up to 5 positions in the genome (option `-r`), the input reads were collapsed (option `-m`) and the output was written to an *.arf* file (option `-t`). The reads are collapsed to save disk space and computational time during the analysis, each read is written only once per sample in the fasta file, with the occurrence of the read available in the read ID, e.g., for a read present 1000 times in sample *N01*, the read ID would be: *>N01_123_x1000*.

The collapsed reads and alignment generated with *mapper.pl* were given as input to the *miRDeep2.pl* script to perform the identification of novel miRNAs. As before, GRch38 was used as the reference genome. This script also required a list of mature and precursor sequences for the species under study as well as mature sequences from closely related species. Here, the mature sequences and

precursors for *Homo sapiens* were extracted from *miRBase* v22.1. For the related species, the mature sequences of chimpanzee and mouse were chosen, as they are close to human and well annotated. *MiRDeep2.pl* provide a scoring system to identify reliable novel miRNAs, here a cut-off of 5 was chosen as it provided the best signal-to-noise ratio. Target prediction for the novel miRNAs identified as differentially expressed was also carried out with *miRDB,* which implements a support vector machines model trained on high-throughput datasets [203].

For the quantification step, the mature sequences from *miRBase* and the novel miRNAs detected previously were merged into a single fasta file; the same was done for the precursors. The *quantifier.pl* script from *miRDeep2* only recognises miRNAs identifiers with a certain format, three words separated by dashes. The format from *miRBase* was compatible but not the one generated for the novel miRNAs, hence *sed* was used to translate the novel miRNAs identifiers into a compatible format, i.e., 'hsa-mature-[number]' for both the mature and precursor sequences. Then, *quantifier.pl* was run, using as input the fasta files containing the mature (option `-m`) and precursor (option `-p`) sequences, the collapsed reads generated with *mapper.pl* (option `-r`). In addition, the species (option `-t`) was set to *hsa,* both the number of nucleotides to consider upstream (option `-e`) and downstream (option `-f`) of the mature sequence were set to 3 and the read counts were weighted by their number of mappings (option `-W`).

The differential expression was performed with *R* v4.0.2, using the *DESeq2* package v1.28.1. First, as the samples were sequenced in two different batches, batch effect correction was performed with *ComBat_seq* on the raw counts [215]. Then, a *DESeqDataSet* object was created from the batch corrected counts. The design was set to the condition (normal and pathological) and time. The miRNAs with less than 10 counts across all samples were discarded. As the aim was to find biomarkers to explain the difference between *normal* and *pathological* samples, the *results* function was used to compare the two conditions across the different time points. *DESeq2* automatically adjusted the p-values with the Benjamin-Hochberg method. The *lfcShrink* function was then applied to the results to adjust the log-fold change to a value more conservative, which usually

follow what would be observed with a larger sample size. *BiomaRt* annotated the differentially expressed miRNAs with Gene Ontology (GO) terms, based on their *miRBase* identifier. Then, *OmicsBox* v1.3.11 performed an enrichment analysis with Fisher's Exact Test, for each time point, comparing the GO terms of the DE miRNAs against those of the complete list of human miRNAs. This list was obtained from *miRBase* via *biomaRt*, using the *Ensembl Gene 101* Mart.

6.2.3 Results

The sequencing generated 2,657 million raw reads. After filtering, 1,065 million reads remained, with an average of 10 million reads per sample.

The *identifier.pl* script from *miRDeep2* predicted 1,349 novel miRNAs with a score between 0 and 10, a higher score meaning a better probability of the miRNA being a true positive. To keep the best signal to noise ratio, only the 128 novel miRNAs with a score equal or higher than 5 were kept. The *quantifier.pl* script from *miRDeep2* aligned the collapsed reads against the novel miRNAs merged with the sequences from *miRBase*. From the 3,141 miRNAs, 793 were not expressed at all, with 0 read count. The results obtained with the four contrasts from *DESeq* model are described in Table 15.

Table 15: Results of the four different contrasts obtained with DESeq2. Only significant miRNAs are represented (adjusted p-value < 0.05)

Contrast	Up-regulated miRNAs	Down-regulated miRNAs
Pathological vs Normal 0h	37	50
Pathological vs Normal 24h	20	4
Pathological vs Normal 48h	1	7
Pathological vs Normal 72h	45	44

It was noted that very few miRNAs were differentially expressed across different time points, as shown on Figure 28. This could be explained by the low number of differentially expressed miRNAs at time points 24h and 48h. Furthermore, the

two conditions represent the same disease, but with varying impacts, so one might expect some similarities in the expression of miRNAs.

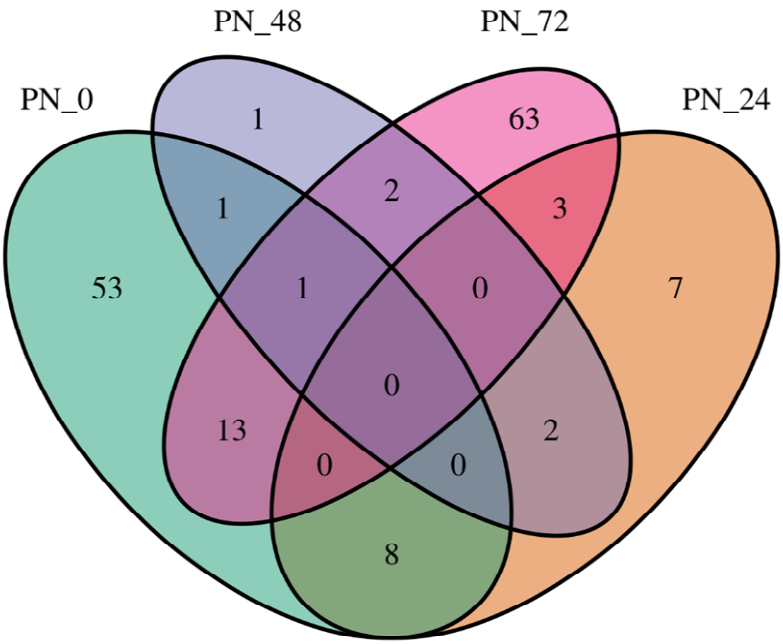


Figure 28: Veen diagram of the differentially expressed miRNAs across the different time points (0h, 24h, 48h and 72h). The miRNAs are filtered by their adjusted p-values (< 0.05). 'PN' represents the contrasts between the two conditions 'Pathological vs Normal'.

Five of the novel miRNAs identified previously were also differentially expressed at time points 0, 48 and 72h. As they might reveal interesting insight about the pathophysiology of HIE, target prediction was performed on them, the results are available as Table 16. Detailed lists of target prediction for each one of the differentially expressed novel miRNAs are available as Table D.1-2 to Table D.1-6.

Table 16: List of the differentially expressed novel miRNAs for the 'pathological vs normal' contrast at the different time points, with their Log-Fold Change (LFC) and targets prediction from miRDB. The score from miRDB is between 50 and 100.

Sequence	Time	LFC	Top three targets (score)
GAGTGTGCTAGAGTCCTCGAAG	0	-2.5	MBD5 (99), FUT9 (97), IMPFH1 (97)
CGTGGTCTTCGGGGAGAGAG	0	-2.1	TSC1 (97), FAM120C (94), GCN1 (92)
CACTGCGCTCCAGCCTGGGCAC	0	-3.5	PYROXD1 (92), SPEN (80), STAG2 (75)
GTGTGTGCACCTGTGTCTGTC	48	-2.5	IGF2 (100), SCARA3 (100), ATP11A (100)
TGGTCCAACGACAGGAGTAGG	72	- 2.9	DCUN1D1 (98), PPP4R3A (98), UBA2 (97)

The enrichment analysis identified the GO terms enriched at the different time points. Interestingly, over-enriched terms in the differentially expressed miRNAs include *cellular response to amyloid-beta* and regulations of *inflammatory response, neuron projection development, angiogenesis, endothelial cell migration, gliogenesis* and *apoptotic process*. The bar charts representing the complete list of enriched terms are available as Figure D.1-1, Figure D.1-2 and Figure D.1-3.

6.2.4 Discussion

miRNAs play an important role in brain development [216] and angiogenesis, hence they are of interest to study brain injury in HIE.

In this discussion, we will go through the seven clusters from Figure 29, from top to bottom and highlighting interesting processes and miRNAs from each group. To keep the discussion concise and for clarity's sake, only up to 30 of the top DE miRNAs were selected for each time point.

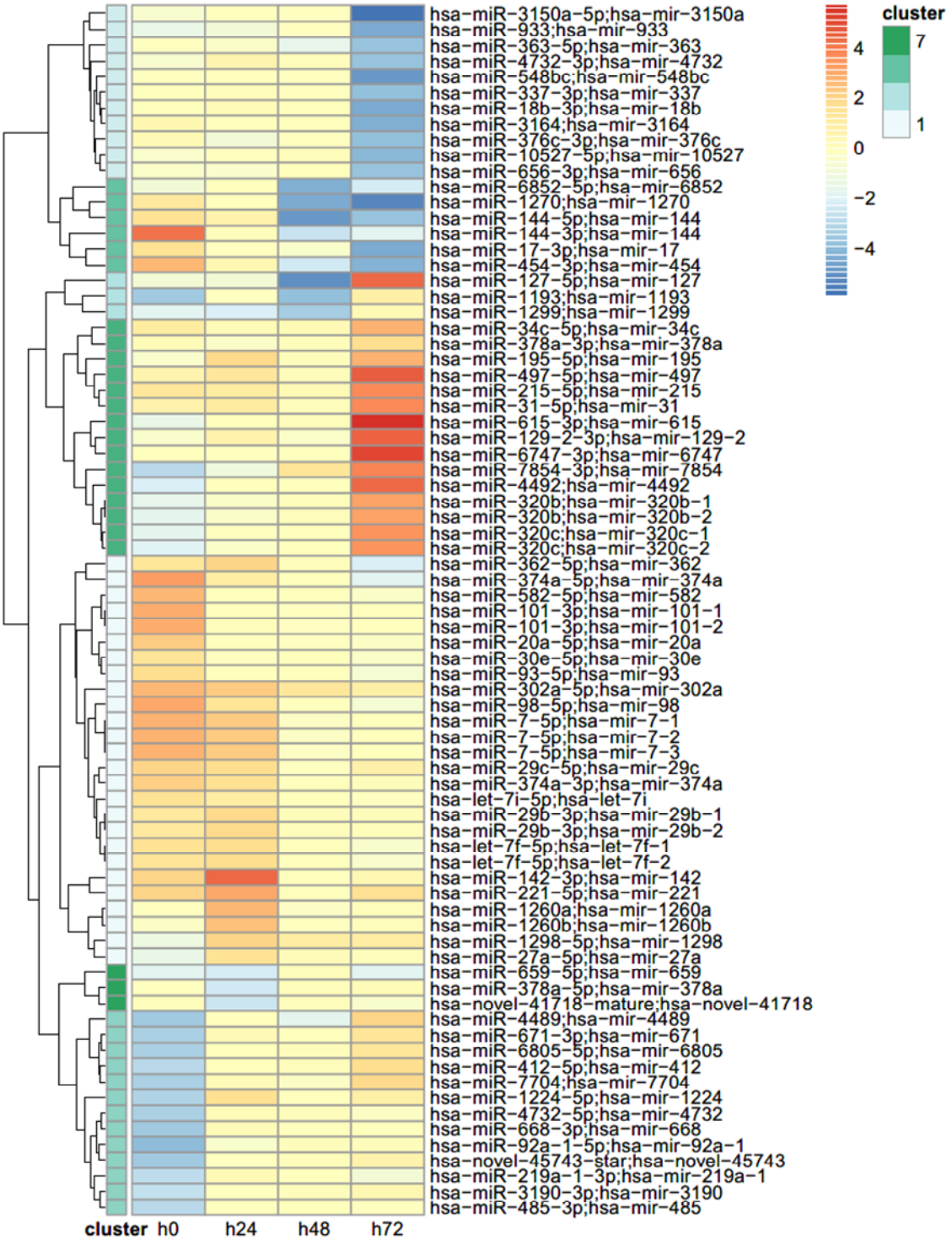


Figure 29: Heatmap of the Log-Fold-Change of the 'pathological vs normal' contrast. For clarity, up to 30 of the most differentially expressed miRNAs were selected at each time point (ordered by adjusted p-value).

6.2.4.1 Cluster 1

This cluster is composed of eleven miRNAs, which are down-regulated at 72h. The miRNAs in this cluster inhibit angiogenesis and epithelial cell differentiation; this is the case of *hsa-miR-363* and *hsa-miR-18b.* Interestingly, *hsa-miR*-376c is down-regulated in infants with HIE, and up-regulation of this miRNA diminishes cell injury from oxygen-glucose deprivation [217]. Hence, this might explain the severe injury suffered by the pathological group.

6.2.4.2 Cluster 2

This cluster is composed of six miRNAs, which are up-regulated at 0h and down-regulated at 48h and 72h. The miRNAs in this group inhibit angiogenesis; this is the case of *hsa-miR-144, hsa-miR-17* and *hsa-miR-454*. The only miRNA significantly differentially expressed at three different time points (0, 48 and 72h) is *hsa-miR-144*. This miRNA has been showed to reduce hypoxia induced autophagy in prostate cancer cells [218], to promote abnormal angiogenesis and hematoma absorption in rats which aggravated neurological deficiencies [219]. Moreover, it is annotated with negative regulation of cholesterol efflux and cholesterol homeostasis is crucial for brain development. Another interesting miRNA from this cluster is *hsa-miR-17,* when up-regulated it attenuates injury from ischemia [220].

6.2.4.3 Cluster 3

This cluster is composed of *hsa-miR-127, hsa-miR1193* and *hsa-miR-1299*. *hsa-miR-127* was found to protect certain cells against ischemia [221] and is down-regulated in the pathological group at different time points.

6.2.4.4 Cluster 4

This cluster is composed of fifteen miRNAs, which are down-regulated at time 0h, while some are up-regulated at 72h. On one hand, some of these miRNAs are annotated in the GO database with a negative regulation of angiogenesis, namely, *hsa-miR-34c, hsa-miR-497* and *hsa-miR-615* [222], on the other hand there is a positive regulation of angiogenesis by *hsa-miR-378a-3p, hsa-miR-31.* This might suggest a dysregulation of the balance between the positive and

negative regulation of angiogenesis in the pathological group. The following miRNAs are also involved in apoptotic processes, *hsa-miR-34c* [223]*, hsa-miR-378a-3p* [224] and *hsa-miR-195*. Some of these miRNAs negatively regulate inflammatory or neuro-inflammatory response; this is the case of *hsa-miR-378a-3p, hsa-miR-195* and *hsa-miR-31.* Interestingly, some of these miRNAs are also involved in dementia; *hsa-miR-34c, hsa-miR-31* are involved in Alzheimer's disease [225] [226]; *hsa-miR-195* have a role in dementia induced by chronic brain hypoperfusion [227] and schizophrenia [228]. Finally, *hsa-miR-129-2* is associated with risk of ischaemic stroke [229], *hsa-miR-6747* was up-regulated in brain arteriovenous malformations [230] and *hsa-miR-497* was identified as a biomarker for acute cerebral infarction [231].

6.2.4.5 Cluster 5

This cluster is composed of twenty-six miRNAs, mostly up-regulated at 0h or 24h. Some miRNAs in this cluster are annotated with negative regulation of angiogenesis and endothelial cell proliferation: *hsa-miR-20a, hsa-miR-29c, hsa-miR-30e, hsa-miR-101* and *hsa-miR-221*. Other miRNAs are annotated with negative regulation of inflammatory response and interleukin production: *hsa-miR-20a, hsa-miR-27a, hsa-miR-93, hsa-miR-98, hsa-miR-101, hsa-miR-142* and *hsa-miR-221*. There are also miRNAs annotated with apoptotic processes, including neuron apoptotic process: *hsa-miR-29c, hsa-miR-30e, hsa-miR-98, hsa-miR-101* and *hsa-miR-221.* Interestingly, a few miRNAs in this cluster are involved in the regulation of amyloid-beta formation, which has a role in Alzheimer's disease, namely *hsa-miR-20a*, *hsa-miR-29c* and *hsa-miR-98.* Moreover, dysregulation of *hsa-miR-142* was associated with Alzheimer's pathogenesis, as its target genes are related to neuronal function and synapse plasticity [232].

Looney et al. previously identified *hsa-miR-374a* as a biomarker in neonatal HIE [233], however as being down-regulated in infants with HIE. In another study, *hsa-miR-374a* was up-regulated in piglet models soon after hypoxic-ischemia with changes in expression specific to moderate and severe cases of HIE, which

is consistent with our measures. This miRNA might be of interest to serve as an early biomarker of severe HIE.

One of the most up-regulated miRNAs at early time points is *hsa-miR-221;* it is annotated with positive regulation of axon regeneration, wound healing and response to glucose, which might suggest a role in the latent phase, when the brain is recovering from the injury.

hsa-miR-142 has interesting annotations, such as positive regulation of astrocyte activation, neuroinflammatory response and regulation of synaptic transmission, which might suggest a potential role in HIE.

Overexpression of *hsa-miR-7* suppresses cell proliferation and promotes apoptosis [234], which is the main source of brain damage in HIE; it was also differentially expressed in the blood of patients with brain arteriovenous malformations [230].

hsa-miR-101 has been previously identified as a hypoxia-responsive miRNA, promoting angiogenesis [235].

Finally, *hsa-miR-27a* and *hsa-miR-302a* are both annotated with negative regulation of cholesterol efflux, as mentioned before, cholesterol homeostasis is crucial for proper brain development. A study demonstrated the protective effect of *hsa-miR-27a* overexpression in hippocampal neurons after hypoxic injury [236].

6.2.4.6 Cluster 6

This cluster is composed of *hsa-miR-659, hsa-miR-378a-5p* and *hsa-miR-41718*, which are down-regulated in the pathological group at 24h. *hsa-miR-659* has been correlated with Progranulin increase in hypoxic conditions, which confers neuroprotection against injury [237].

6.2.4.7 Cluster 7

This cluster is composed of thirteen miRNAs, which are all down-regulated at time 0h. *hsa-miR-671* negatively affects the levels of *CDR1* [238], dysregulation of this gene affects Alzheimer's [239] and Huntington disease [240]. *hsa-miR-668*

preserve mitochondrial activity in ischemic kidney injury [241] and inhibition of this miRNA protects against neuronal apoptosis in cerebral ischemic stroke [242]. *hsa-miR-1224* is annotated with positive regulation of sprouting angiogenesis, while *hsa-miR-92a-1* is annotated with negative regulation of sprouting angiogenesis, blood vessel diameter and inflammatory response, as well as positive regulation of acute inflammatory response and apoptotic process. Interestingly *hsa-miR-219a-1* is annotated with negative regulation of neuron projection development and have a role in regulation of neuronal apoptosis [243] which is a component of brain cell injury in HIE. Low expression of *hsa-miR-219-1* was also linked to epilepsy [244], which is a symptom of severe cases in HIE.

6.2.4.8 Novel miRNAs and the roles of their target genes

All the novel miRNAs that were significantly DE are down regulated (see Table 16). The top target identified for *hsa-novel-27395-mature* is *MBD5*, which has been linked to mental retardation and epileptic encephalopathy [245] [246]. Similarly, the top target for *hsa-novel-3327-mature* is *TSC1*, which was linked to epilepsy [247] and neuroprotection against ischemia [248]. The third target for *hsa-novel-45743-sta*r is *STAG2* and has been linked to *Mullegama-Klein-Martinez Syndrome* [249] and *holoprosencephaly* [250]*.* However, caution is required, as the score for this target prediction was 75 out of 100, while miRDB suggests that scores higher than 80 are reliable*.* The top target of *hsa-novel-41718-mature* is *IFG2* and is expressed in epithelial cells lining the surface of the brain in adults [251]. These findings need to be confirmed experimentally but might provide new understanding of the underlying mechanisms of HIE.

6.2.4.9 Alzheimer's disease

Swarbrick et al. identified miRNAs biomarkers for Alzheimer's disease [225], interestingly some of the miRNAs overlap with our list of DE genes, namely *hsa-miR-26b, hsa-mir-30e, hsa-miR-34c, hsa-miR-200c* and *hsa-miR-485*. Moreover, as mentioned in this discussion, several of the miRNAs are annotated with amyloid beta response, which is a component of Alzheimer's disease, namely *hsa-let-7f, hsa-miR-98, hsa-miR-106b, hsa-miR-200a* and *hsa-miR-200c*. This

might suggest parallel common mechanisms behind the pathogenesis of Alzheimer's disease and severe cases of HIE.

6.2.5 Conclusion

miRNAs are a promising avenue as biomarkers for severity in HIE. While no stable biomarkers over all time points were identified, this analysis provided insights into the pathogenesis of HIE, notably with the discovery of novel miRNAs whose targets are related to brain functions and diseases. There were also some interesting parallels between the miRNAs identified in the differential expression analysis and those involved in the pathogenesis of Alzheimer's. Further research would be needed to refine and confirm these discoveries.

REFERENCES

[1] International Human Genome Sequencing Consortium, ‘Initial sequencing and analysis of the human genome’, *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001, doi: 10.1038/35057062.

[2] International Human Genome Sequencing Consortium, ‘Finishing the euchromatic sequence of the human genome’, *Nature*, vol. 431, no. 7011, pp. 931–945, Oct. 2004, doi: 10.1038/nature03001.

[3] V. A. Schneider *et al.*, ‘Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly’, *bioRxiv*, p. 072116, Aug. 2016, doi: 10.1101/072116.

[4] Bronwen, ‘Accessing alternate sequences in human’, *Ensembl Blog*, May 20, 2011. http://www.ensembl.info/2011/05/20/accessing-non-reference-sequences-in-human/ (accessed Feb. 20, 2020).

[5] R. M. Sherman *et al.*, ‘Assembly of a pan-genome from deep sequencing of 910 humans of African descent’, *Nat Genet*, vol. 51, no. 1, pp. 30–35, Jan. 2019, doi: 10.1038/s41588-018-0273-y.

[6] S. Nurk *et al.*, ‘The complete sequence of a human genome’, p. 2021.05.26.445798, May 2021, doi: 10.1101/2021.05.26.445798.

[7] T. Hon *et al.*, ‘Highly accurate long-read HiFi sequencing data for five complex genomes’, *Sci Data*, vol. 7, no. 1, p. 399, Nov. 2020, doi: 10.1038/s41597-020-00743-4.

[8] J. Shendure *et al.*, ‘DNA sequencing at 40: past, present and future’, *Nature*, vol. 550, no. 7676, pp. 345–353, Oct. 2017, doi: 10.1038/nature24286.

[9] A. Auton *et al.*, ‘A global reference for human genetic variation’, *Nature*, vol. 526, no. 7571, Art. no. 7571, Oct. 2015, doi: 10.1038/nature15393.

[10] S. T. Sherry *et al.*, ‘dbSNP: the NCBI database of genetic variation.’, *Nucleic acids research*, vol. 29, no. 1, pp. 308–11, Jan. 2001, doi: 10.1093/nar/29.1.308.

[11] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, ‘Copy Number Variation in Human Health, Disease, and Evolution’, *Annual Review of Genomics and Human Genetics*, vol. 10, no. 1, pp. 451–481, 2009, doi: 10.1146/annurev.genom.9.081307.164217.

[12] J. Lejeune, ‘Etude des chromosomes somatiques de neuf enfants mongoliens’, *C R Acad Sci (Paris)*, vol. 248, pp. 1721–1722, 1959.

[13] H. Li, ‘A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.’, *Bioinformatics (Oxford, England)*, vol. 27, no. 21, pp. 2987–93, Nov. 2011, doi: 10.1093/bioinformatics/btr509.

[14] R. Poplin *et al.*, ‘A universal SNP and small-indel variant caller using deep neural networks’, *Nature Biotechnology*, vol. 36, no. 10, Art. no. 10, Nov. 2018, doi: 10.1038/nbt.4235.

[15] A. McKenna *et al.*, ‘The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.’, *Genome research*, vol. 20, no. 9, pp. 1297–303, Sep. 2010, doi: 10.1101/gr.107524.110.

[16] M. A. DePristo *et al.*, ‘A framework for variation discovery and genotyping using next-generation DNA sequencing data’, *Nature genetics*, vol. 43, no. 5, p. 491, 2011, doi: 10.1038/NG.806.

[17] S. Andrews, ‘FastQC A Quality Control tool for High Throughput Sequence Data’, 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed Apr. 30, 2018).

[18] J. Xin *et al.*, ‘High-performance web services for querying gene and variant annotation’, *Genome Biology*, vol. 17, no. 1, p. 91, Dec. 2016, doi: 10.1186/s13059-016-0953-9.

[19] W. McLaren *et al.*, ‘The Ensembl Variant Effect Predictor’, *Genome Biology*, vol. 17, no. 1, p. 122, Dec. 2016, doi: 10.1186/s13059-016-0974-4.

[20] P. Cingolani *et al.*, ‘A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.’, *Fly*, vol. 6, no. 2, pp. 80–92, 2012, doi: 10.4161/fly.19695.

[21] D. J. McCarthy *et al.*, ‘Choice of transcripts and software has a large effect on variant annotation.’, *Genome medicine*, vol. 6, no. 3, p. 26, 2014, doi: 10.1186/gm543.

[22] L. D. Ward and M. Kellis, ‘Interpreting noncoding genetic variation in complex traits and human disease’, *Nature Biotechnology*, vol. 30, no. 11, pp. 1095–1106, Nov. 2012, doi: 10.1038/nbt.2422.

[23] GTEx Consortium, ‘The GTEx Consortium atlas of genetic regulatory effects across human tissues’, *Science*, vol. 369, no. 6509, pp. 1318–1330, Sep. 2020, doi: 10.1126/science.aaz1776.

[24] R. Ottman, ‘Gene–Environment Interaction: Definitions and Study Designs’, *Prev Med*, vol. 25, no. 6, pp. 764–770, 1996.

[25] M. Garaulet *et al.*, ‘PERIOD2 Variants Are Associated with Abdominal Obesity, Psycho-Behavioral Factors, and Attrition in the Dietary Treatment of Obesity’, *Journal of the American Dietetic Association*, vol. 110, no. 6, pp. 917–921, Jun. 2010, doi: 10.1016/j.jada.2010.03.017.

[26] J. Westfall and T. Yarkoni, ‘Statistically Controlling for Confounding Constructs Is Harder than You Think’, *PLOS ONE*, vol. 11, no. 3, p. e0152719, Mar. 2016, doi: 10.1371/journal.pone.0152719.

[27] D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. Davey Smith, ‘Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology’, *Statistics in Medicine*, vol. 27, no. 8, pp. 1133–1163, 2008, doi: 10.1002/sim.3034.

[28] A. Xue *et al.*, ‘Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes’, *Nat Commun*, vol. 9, no. 1, pp. 1–14, Jul. 2018, doi: 10.1038/s41467-018-04951-w.

[29] J. Erdmann, T. Kessler, L. Munoz Venegas, and H. Schunkert, ‘A decade of genome-wide association studies for coronary artery disease: the challenges ahead’, *Cardiovasc Res*, vol. 114, no. 9, pp. 1241–1257, Jul. 2018, doi: 10.1093/cvr/cvy084.

[30] L. Yengo *et al.*, ‘Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry’, *Hum Mol Genet*, vol. 27, no. 20, pp. 3641–3649, Oct. 2018, doi: 10.1093/hmg/ddy271.

[31] P. M. Visscher *et al.*, ‘10 Years of GWAS Discovery: Biology, Function, and Translation’, *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, Jul. 2017, doi: 10.1016/j.ajhg.2017.06.005.

[32] E. Lander and L. Kruglyak, ‘Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results’, *Nat Genet*, vol. 11, no. 3, pp. 241–247, Nov. 1995, doi: 10.1038/ng1195-241.

[33] C. Xu, I. Tachmazidou, K. Walter, A. Ciampi, E. Zeggini, and C. M. T. Greenwood, ‘Estimating Genome-Wide Significance for Whole-Genome Sequencing Studies’, *Genet Epidemiol*, vol. 38, no. 4, pp. 281–290, Apr. 2014, doi: 10.1002/gepi.21797.

[34] J. Fadista, A. K. Manning, J. C. Florez, and L. Groop, ‘The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants’, *Eur J Hum Genet*, vol. 24, no. 8, pp. 1202–1205, Aug. 2016, doi: 10.1038/ejhg.2015.269.

[35] T. A. Manolio *et al.*, ‘Finding the missing heritability of complex diseases’, *Nature*, vol. 461, no. 7265, Art. no. 7265, Oct. 2009, doi: 10.1038/nature08494.

[36] C. S. Carlson *et al.*, ‘Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study’, *PLoS Biol*, vol. 11, no. 9, Sep. 2013, doi: 10.1371/journal.pbio.1001661.

[37] L. A. Hindorff *et al.*, ‘Potential etiologic and functional implications of genome-wide association loci for human diseases and traits’, *Proc Natl Acad Sci U S A*, vol. 106, no. 23, pp. 9362–9367, Jun. 2009, doi: 10.1073/pnas.0903103106.

[38] A. Buniello *et al.*, ‘The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019’, *Nucleic Acids Res*, vol. 47, no. Database issue, pp. D1005–D1012, Jan. 2019, doi: 10.1093/nar/gky1120.

[39] M. J. Li *et al.*, ‘GWASdb: a database for human genetic variants identified by genome-wide association studies’, *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D1047–D1054, Jan. 2012, doi: 10.1093/nar/gkr1182.

[40] T. Beck, T. Shorter, and A. J. Brookes, ‘GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies’, *Nucleic Acids Res*, vol. 48, no. D1, pp. D933–D940, Jan. 2020, doi: 10.1093/nar/gkz895.

[41] A. V. Khera *et al.*, ‘Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations’, *Nat Genet*, vol. 50, no. 9, pp. 1219–1224, Sep. 2018, doi: 10.1038/s41588-018-0183-z.

[42] S. W. Choi, T. S.-H. Mak, and P. F. O’Reilly, ‘Tutorial: a guide to performing polygenic risk score analyses’, *Nature Protocols*, vol. 15, no. 9, Art. no. 9, Sep. 2020, doi: 10.1038/s41596-020-0353-1.

[43] S. Purcell *et al.*, ‘PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses’, *Am J Hum Genet*, vol. 81, no. 3, pp. 559–575, Sep. 2007.

[44] S. W. Choi and P. F. O’Reilly, ‘PRSice-2: Polygenic Risk Score software for biobank-scale data’, *GigaScience*, vol. 8, no. 7, Jul. 2019, doi: 10.1093/gigascience/giz082.

[45] F. Privé, J. Arbel, and B. J. Vilhjálmsson, ‘LDpred2: better, faster, stronger’, *Bioinformatics*, vol. 36, no. 22–23, Dec. 2020, doi: 10.1093/bioinformatics/btaa1029.

[46] T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham, ‘Polygenic scores via penalized regression on summary statistics’, *Genetic Epidemiology*, vol. 41, no. 6, pp. 469–480, 2017, doi: 10.1002/gepi.22050.

[47] N. J. Wald and R. Old, ‘The illusion of polygenic disease risk prediction’, *Genetics in Medicine*, vol. 21, no. 8, Art. no. 8, Aug. 2019, doi: 10.1038/s41436-018-0418-5.

[48] G. Lázaro-Muñoz, S. Pereira, S. Carmi, and T. Lencz, ‘Screening embryos for polygenic conditions and traits: ethical considerations for an emerging technology’, *Genet Med*, vol. 23, no. 3, pp. 432–434, Mar. 2021, doi: 10.1038/s41436-020-01019-3.

[49] A. Torkamani, N. E. Wineinger, and E. J. Topol, ‘The personal and clinical utility of polygenic risk scores’, *Nat Rev Genet*, vol. 19, no. 9, Art. no. 9, Sep. 2018, doi: 10.1038/s41576-018-0018-x.

[50] R. Lakhtakia, ‘The history of diabetes mellitus.’, *Sultan Qaboos University medical journal*, vol. 13, no. 3, pp. 368–70, Aug. 2013.

[51] M. Karamanou, A. Protogerou, G. Tsoucalas, G. Androutsos, and E. Poulakou-Rebelakou, ‘Milestones in the history of diabetes mellitus: The main contributors.’, *World journal of diabetes*, vol. 7, no. 1, pp. 1–7, Jan. 2016, doi: 10.4239/wjd.v7.i1.1.

[52] K. Laios, M. Karamanou, Z. Saridaki, and G. Androutsos, ‘Aretaeus of Cappadocia and the first description of diabetes’, *Hormones*, vol. 11, no. 1, pp. 109–113, Jan. 2012, doi: 10.1007/BF03401545.

[53] B. Claude, ‘Du suc pancréatique, et de son role dans les phénomènes de la digestion.’, Société de biologie, 1849.

[54] F. J. G. Ebling, ‘Homage to Paul Langerhans’, *Journal of Investigative Dermatology*, vol. 75, no. 1, pp. 3–5, Jul. 1980, doi: 10.1111/1523-1747.ep12521014.

[55] J. Mering and O. Minkowski, ‘Diabetes mellitus nach Pankreasexstirpation’, *Archiv für Experimentelle Pathologie und Pharmakologie*, vol. 26, no. 5–6, pp. 371–387, Jan. 1890, doi: 10.1007/BF01831214.

[56] J. De Meyer, ‘Action de la sécrétion interne du pancréas sur différents organes et en particulier sur la sécrétion rénale.’, *Arch Fisiol*, 1909.

[57] E. Schafer, *An introduction to the study of the endocrine glands and internal secretions.* Palo Alto, California: Stanford University, 1914.

[58] F. G. Banting, C. H. Best, J. B. Collip, W. R. Campbell, and A. A. Fletcher, ‘Pancreatic Extracts in the Treatment of Diabetes Mellitus.’, *Canadian Medical Association journal*, vol. 12, no. 3, pp. 141–6, Mar. 1922.

[59] D. V. Goeddel *et al.*, ‘Expression in Escherichia coli of chemically synthesized genes for human insulin.’, *Proc Natl Acad Sci U S A*, vol. 76, no. 1, pp. 106–110, Jan. 1979.

[60] C. C. Quianzon and I. Cheikh, ‘History of insulin’, *J Community Hosp Intern Med Perspect*, vol. 2, no. 2, p. 10.3402/jchimp.v2i2.18701, Jul. 2012, doi: 10.3402/jchimp.v2i2.18701.

[61] World Health Organisation, ‘Diabetes fact-sheet’, 2018. https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed Jul. 31, 2019).

[62] P. Leete *et al.*, ‘Studies of insulin and proinsulin in pancreas and serum support the existence of aetiopathological endotypes of type 1 diabetes associated with age at diagnosis’, *Diabetologia*, Mar. 2020, doi: 10.1007/s00125-020-05115-6.

[63] J. Ilonen, J. Lempainen, and R. Veijola, ‘The heterogeneous pathogenesis of type 1 diabetes mellitus’, *Nat Rev Endocrinol*, vol. 15, no. 11, Art. no. 11, Nov. 2019, doi: 10.1038/s41574-019-0254-y.

[64] M. S. Udler *et al.*, ‘Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis’, *PLOS Medicine*, vol. 15, no. 9, p. e1002654, Sep. 2018, doi: 10.1371/journal.pmed.1002654.

[65] L. H. Philipson, ‘Harnessing heterogeneity in type 2 diabetes mellitus’, *Nat Rev Endocrinol*, vol. 16, no. 2, Art. no. 2, Feb. 2020, doi: 10.1038/s41574-019-0308-1.

[66] E. Ahlqvist *et al.*, ‘Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables’, *The Lancet Diabetes & Endocrinology*, vol. 6, no. 5, pp. 361–369, May 2018, doi: 10.1016/S2213-8587(18)30051-2.

[67] G. Roglic and World Health Organization, Eds., *Global report on diabetes*. Geneva, Switzerland: World Health Organization, 2016.

[68] M. J. Redondo, A. K. Steck, and A. Pugliese, ‘Genetics of type 1 diabetes’, *Pediatr Diabetes*, vol. 19, no. 3, pp. 346–353, May 2018, doi: 10.1111/pedi.12597.

[69] S. Y. Choo, ‘The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications’, *Yonsei Medical Journal*, vol. 48, no. 1, p. 11, Feb. 2007, doi: 10.3349/ymj.2007.48.1.11.

[70] J. Kaprio *et al.*, ‘Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland’, *Diabetologia*, vol. 35, no. 11, pp. 1060–1067, Nov. 1992, doi: 10.1007/BF02221682.

[71] J. Köbberling and H. Tillil, ‘Empirical risk figures for first degree relatives of non-insulin dependent diabetics.’, in *The genetics of diabetes mellitus.*, Academic Press, 1982.

[72] A. Stančáková and M. Laakso, ‘Genetics of Type 2 Diabetes’, *Novelties in Diabetes*, vol. 31, pp. 203–220, 2016, doi: 10.1159/000439418.

[73] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, ‘Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.’, *Nucleic acids research*, vol. 33, no. Database issue, pp. D514-7, Jan. 2005, doi: 10.1093/nar/gki033.

[74] P. Sneha, D. Thirumal Kumar, J. Lijo, M. Megha, R. Siva, and C. George Priya Doss, ‘Chapter Six - Probing the Protein–Protein Interaction Network of Proteins Causing Maturity Onset Diabetes of the Young’, in *Advances in Protein Chemistry and Structural Biology*, vol. 110, R. Donev, Ed. Academic Press, 2018. doi: 10.1016/bs.apcsb.2017.07.004.

[75] N. Bottini *et al.*, ‘A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes’, *Nat Genet*, vol. 36, no. 4, pp. 337–338, Apr. 2004, doi: 10.1038/ng1323.

[76] S. Kang, T. Tanaka, M. Narazaki, and T. Kishimoto, ‘Targeting Interleukin-6 Signaling in Clinic’, *Immunity*, vol. 50, no. 4, pp. 1007–1023, Apr. 2019, doi: 10.1016/j.immuni.2019.03.026.

[77] J. C. Roach *et al.*, ‘Genetic Mapping at 3-Kilobase Resolution Reveals Inositol 1,4,5-Triphosphate Receptor 3 as a Risk Factor for Type 1 Diabetes in Sweden’, *The American Journal of Human Genetics*, vol. 79, no. 4, pp. 614–627, Oct. 2006, doi: 10.1086/507876.

[78] P. D. Thomas *et al.*, ‘PANTHER: A Library of Protein Families and Subfamilies Indexed by Function’, *Genome Res.*, vol. 13, no. 9, pp. 2129–2141, Jan. 2003, doi: 10.1101/gr.772403.

[79] S. Carbon and C. Mungall, ‘Gene Ontology Data Archive’. Zenodo, Jul. 02, 2018. doi: 10.5281/zenodo.5080993.

[80] International Diabetes Federation, ‘IDF Diabetes Atlas (8th edition)’, International Diabetes Federation, Brussels, Belgium, 2017. [Online]. Available: http://www.diabetesatlas.org

[81] L. Chen, D. J. Magliano, and P. Z. Zimmet, ‘The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives’, *Nature Reviews Endocrinology*, vol. 8, no. 4, pp. 228–236, Apr. 2012, doi: 10.1038/nrendo.2011.183.

[82] A. B. Olokoba, O. A. Obateru, and L. B. Olokoba, ‘Type 2 diabetes mellitus: a review of current trends.’, *Oman medical journal*, vol. 27, no. 4, pp. 269–73, Jul. 2012, doi: 10.5001/omj.2012.68.

[83] World Health Organisation, ‘Obesity and overweight fact sheet’, 2018. https://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight (accessed Jul. 21, 2019).

[84] M. J. Müller *et al.*, ‘The case of GWAS of obesity: does body weight control play by the rules?’, *International Journal of Obesity*, vol. 42, no. 8, Art. no. 8, Aug. 2018, doi: 10.1038/s41366-018-0081-6.

[85] E. E. Kershaw and J. S. Flier, ‘Adipose Tissue as an Endocrine Organ’, *J Clin Endocrinol Metab*, vol. 89, no. 6, pp. 2548–2556, Jun. 2004, doi: 10.1210/jc.2004-0395.

[86] C. H. Saely, K. Geiger, and H. Drexel, ‘Brown versus White Adipose Tissue: A Mini-Review’, *Gerontology*, vol. 58, no. 1, pp. 15–23, 2012, doi: 10.1159/000321319.

[87] S. Maurer, M. Harms, and J. Boucher, ‘The colorful versatility of adipocytes: white-to-brown transdifferentiation and its therapeutic potential in humans’, *The FEBS Journal*, vol. 288, no. 12, pp. 3628–3646, 2021, doi: 10.1111/febs.15470.

[88] D. C. Berry, D. Stenesen, D. Zeve, and J. M. Graff, ‘The developmental origins of adipose tissue’, *Development*, vol. 140, no. 19, pp. 3939–3949, Oct. 2013, doi: 10.1242/dev.080549.

[89] J.-P. Bastard *et al.*, ‘Recent advances in the relationship between obesity, inflammation, and insulin resistance’, *Eur Cytokine Netw*, vol. 17, no. 1, pp. 4–12, Mar. 2006.

[90] M. Lafontan, ‘Differences Between Subcutaneous and Visceral Adipose Tissues’, in *Physiology and Physiopathology of Adipose Tissue*, J.-P. Bastard and B. Fève, Eds. Paris: Springer, 2013, pp. 329–349. doi: 10.1007/978-2-8178-0343-2_23.

[91] G. Blackburn, ‘Effect of Degree of Weight Loss on Health Benefits’, *Obesity Research*, vol. 3, no. S2, pp. 211s–216s, Sep. 1995, doi: 10.1002/j.1550-8528.1995.tb00466.x.

[92] U. Peters and A. E. Dixon, ‘The effect of obesity on lung function’, *Expert Rev Respir Med*, vol. 12, no. 9, pp. 755–767, Sep. 2018, doi: 10.1080/17476348.2018.1506331.

[93] M. Gao *et al.*, ‘Associations between body-mass index and COVID-19 severity in 6·9 million people in England: a prospective, community-based, cohort study’, *Lancet Diabetes Endocrinol*, vol. 9, no. 6, Jun. 2021, doi: 10.1016/S2213-8587(21)00089-9.

[94] G. S. Barsh, I. S. Farooqi, and S. O’Rahilly, ‘Genetics of body-weight regulation’, *Nature*, vol. 404, no. 6778, Art. no. 6778, Apr. 2000, doi: 10.1038/35007519.

[95] H. H. Maes, M. C. Neale, and L. J. Eaves, ‘Genetic and environmental factors in relative body weight and human adiposity’, *Behav Genet*, vol. 27, no. 4, pp. 325–351, Jul. 1997, doi: 10.1023/a:1025635913927.

[96] D. Albuquerque, C. Nóbrega, L. Manco, and C. Padez, ‘The contribution of genetics and environment to obesity’, *British Medical Bulletin*, vol. 123, no. 1, pp. 159–173, Sep. 2017, doi: 10.1093/bmb/ldx022.

[97] J. M. Friedman, ‘Obesity in the new millennium’, *Nature*, vol. 404, no. 6778, Art. no. 6778, Apr. 2000, doi: 10.1038/35007504.

[98] Y. Yang and Y. Xu, ‘The central melanocortin system and human obesity’, *Journal of Molecular Cell Biology*, vol. 12, no. 10, pp. 785–797, Oct. 2020, doi: 10.1093/jmcb/mjaa048.

[99] A. E. Locke *et al.*, ‘Genetic studies of body mass index yield new insights for obesity biology’, *Nature*, vol. 518, no. 7538, Art. no. 7538, Feb. 2015, doi: 10.1038/nature14177.

[100] M. Claussnitzer *et al.*, ‘FTO Obesity Variant Circuitry and Adipocyte Browning in Humans’, *http://dx.doi.org/10.1056/NEJMoa1502214*, Sep. 02,

2015. https://www.nejm.org/doi/10.1056/NEJMoa1502214 (accessed Sep. 20, 2021).

[101] T. O. Kilpeläinen *et al.*, 'Physical Activity Attenuates the Influence of FTO Variants on Obesity Risk: A Meta-Analysis of 218,166 Adults and 19,268 Children', *PLOS Medicine*, vol. 8, no. 11, p. e1001116, Nov. 2011, doi: 10.1371/journal.pmed.1001116.

[102] L. Abarca-Gómez *et al.*, 'Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults', *The Lancet*, vol. 390, no. 10113, pp. 2627–2642, Dec. 2017, doi: 10.1016/S0140-6736(17)32129-3.

[103] N. J. Schork, 'Personalized medicine: Time for one-person trials', *Nature News*, vol. 520, no. 7549, p. 609, Apr. 2015, doi: 10.1038/520609a.

[104] V. Gambardella *et al.*, 'Personalized Medicine: Recent Progress in Cancer Therapy', *Cancers (Basel)*, vol. 12, no. 4, p. E1009, Apr. 2020, doi: 10.3390/cancers12041009.

[105] C. Sudlow *et al.*, 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age', *PLOS Medicine*, vol. 12, no. 3, p. e1001779, Mar. 2015, doi: 10.1371/journal.pmed.1001779.

[106] S. Senn, 'Statistical pitfalls of personalized medicine', *Nature*, vol. 563, no. 7733, pp. 619–621, Nov. 2018, doi: 10.1038/d41586-018-07535-2.

[107] J. de Toro-Martín, B. J. Arsenault, J.-P. Després, and M.-C. Vohl, 'Precision Nutrition: A Review of Personalized Nutritional Approaches for the Prevention and Management of Metabolic Syndrome', *Nutrients*, vol. 9, no. 8, p. 913, Aug. 2017, doi: 10.3390/nu9080913.

[108] N. V. Dhurandhar *et al.*, 'Energy balance measurement: when something is not better than nothing', *Int J Obes*, vol. 39, no. 7, pp. 1109–1113, Jul. 2015, doi: 10.1038/ijo.2014.199.

[109] F. Sofi, R. Abbate, G. F. Gensini, and A. Casini, 'Accruing evidence on benefits of adherence to the Mediterranean diet on health: an updated systematic review and meta-analysis', *Am J Clin Nutr*, vol. 92, no. 5, pp. 1189–1196, Nov. 2010, doi: 10.3945/ajcn.2010.29673.

[110] H. Schröder *et al.*, 'A Short Screener Is Valid for Assessing Mediterranean Diet Adherence among Older Spanish Men and Women', *J Nutr*, vol. 141, no. 6, pp. 1140–1145, Jun. 2011, doi: 10.3945/jn.110.135566.

[111] M. Sotos-Prieto, B. Moreno-Franco, J. M. Ordovás, M. León, J. A. Casasnovas, and J. L. Peñalvo, 'Design and development of an instrument to measure overall lifestyle habits for epidemiological research: the Mediterranean Lifestyle (MEDLIFE) index', *Public Health Nutrition*, vol. 18, no. 6, pp. 959–967, Apr. 2015, doi: 10.1017/S1368980014001360.

[112] M. A. Martínez-González *et al.*, ‘A 14-Item Mediterranean Diet Assessment Tool and Obesity Indexes among High-Risk Subjects: The PREDIMED Trial’, *PLOS ONE*, vol. 7, no. 8, p. e43134, Aug. 2012, doi: 10.1371/journal.pone.0043134.

[113] C. Celis-Morales *et al.*, ‘Effect of personalized nutrition on health-related behaviour change: evidence from the Food4me European randomized controlled trial’, *International Journal of Epidemiology*, vol. 46, no. 2, p. dyw186, Aug. 2016, doi: 10.1093/ije/dyw186.

[114] D. Zeevi *et al.*, ‘Personalized Nutrition by Prediction of Glycemic Responses’, *Cell*, vol. 163, no. 5, pp. 1079–1094, Nov. 2015, doi: 10.1016/j.cell.2015.11.001.

[115] J. H. Friedman, ‘Greedy Function Approximation: A Gradient Boosting Machine’, *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[116] T. M. S. Wolever, ‘Personalized nutrition by prediction of glycaemic responses: fact or fantasy?’, *Eur J Clin Nutr*, vol. 70, no. 4, pp. 411–413, Apr. 2016, doi: 10.1038/ejcn.2016.31.

[117] V. K. Ridaura *et al.*, ‘Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice’, *Science*, vol. 341, no. 6150, p. 1241214, Sep. 2013, doi: 10.1126/science.1241214.

[118] G. Asher and P. Sassone-Corsi, ‘Time for Food: The Intimate Interplay between Nutrition, Metabolism, and the Circadian Clock’, *Cell*, vol. 161, no. 1, pp. 84–92, Mar. 2015, doi: 10.1016/j.cell.2015.03.015.

[119] J. J. Gooley and E. C.-P. Chua, ‘Diurnal Regulation of Lipid Metabolism and Applications of Circadian Lipidomics’, *Journal of Genetics and Genomics*, vol. 41, no. 5, pp. 231–250, May 2014, doi: 10.1016/j.jgg.2014.04.001.

[120] M. Kohlmeier *et al.*, ‘Guide and Position of the International Society of Nutrigenetics/Nutrigenomics on Personalized Nutrition: Part 2 - Ethics, Challenges and Endeavors of Precision Nutrition’, *LFG*, vol. 9, no. 1, pp. 28–46, 2016, doi: 10.1159/000446347.

[121] J. Piñero *et al.*, ‘DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants’, *Nucleic Acids Research*, vol. 45, no. D1, pp. D833–D839, Jan. 2017, doi: 10.1093/nar/gkw943.

[122] J. Y. Zongfu Cao, Lei Wang, Yilu Chen, Ruikun Cai, Jianbo Lu, Yufei Yu, Cuixia Chen, Feng Gu and X. Ma, ‘VarfromPDB: An Automated and Integrated Tool to Mine Disease-Gene-Variant Relations from the Public Databases and Literature’, *Journal of Proteomics & Bioinformatics*, 2017, doi: 10.4172/jpb.1000455.

[123] C. Molitor, M. Brember, and F. Mohareb, ‘VarGen: an R package for disease-associated variant discovery and annotation’, *Bioinformatics*, vol. 36, no. 8, pp. 2626–2627, Apr. 2020, doi: 10.1093/bioinformatics/btz930.

[124] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, ‘CADD: predicting the deleteriousness of variants throughout the human genome’, *Nucleic Acids Research*, vol. 47, no. D1, pp. D886–D894, Jan. 2019, doi: 10.1093/nar/gky1016.

[125] M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, ‘FATHMM-XF: accurate prediction of pathogenic point mutations via extended features’, *Bioinformatics*, vol. 34, no. 3, pp. 511–513, Feb. 2018, doi: 10.1093/bioinformatics/btx536.

[126] M. J. Landrum *et al.*, ‘ClinVar: public archive of relationships among sequence variation and human phenotype.’, *Nucleic acids research*, vol. 42, no. Database issue, pp. D980-5, Jan. 2014, doi: 10.1093/nar/gkt1113.

[127] GTEx Consortium, ‘Genetic effects on gene expression across human tissues’, *Nature*, vol. 550, no. 7675, pp. 204–213, Oct. 2017, doi: 10.1038/nature24277.

[128] GTEx Consortium, ‘The Genotype-Tissue Expression (GTEx) project’, *Nat Genet*, vol. 45, no. 6, pp. 580–585, Jun. 2013, doi: 10.1038/ng.2653.

[129] M. D. Mailman *et al.*, ‘The NCBI dbGaP database of genotypes and phenotypes’, *Nat Genet*, vol. 39, no. 10, pp. 1181–1186, Oct. 2007, doi: 10.1038/ng1007-1181.

[130] A. C. Nica and E. T. Dermitzakis, ‘Expression quantitative trait loci: present and future.’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 368, no. 1620, p. 20120362, 2013, doi: 10.1098/rstb.2012.0362.

[131] F. W. Albert and L. Kruglyak, ‘The role of regulatory variation in complex traits and disease’, *Nature Reviews Genetics*, vol. 16, no. 4, Art. no. 4, Apr. 2015, doi: 10.1038/nrg3891.

[132] R. Andersson *et al.*, ‘An atlas of active enhancers across human cell types and tissues’, *Nature*, vol. 507, no. 7493, Art. no. 7493, Mar. 2014, doi: 10.1038/nature12787.

[133] D. Smedley *et al.*, ‘BioMart – biological queries made easy’, *BMC Genomics*, vol. 10, no. 1, p. 22, Jan. 2009, doi: 10.1186/1471-2164-10-22.

[134] C. S. Richards *et al.*, ‘ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007’, *Genetics in Medicine*, vol. 10, no. 4, Art. no. 4, Apr. 2008, doi: 10.1097/GIM.0b013e31816b5cae.

[135] D. Lamparter, D. Marbach, R. Rueedi, Z. Kutalik, and S. Bergmann, ‘Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based

Summary Statistics', *PLOS Computational Biology*, vol. 12, no. 1, p. e1004714, Jan. 2016, doi: 10.1371/journal.pcbi.1004714.

[136] O. Ukkola *et al.*, 'Mutations in the preproghrelin/ghrelin gene associated with obesity in humans', *J Clin Endocrinol Metab*, vol. 86, no. 8, pp. 3996–3999, Aug. 2001, doi: 10.1210/jcem.86.8.7914.

[137] W. Wang and Y.-X. Tao, 'Chapter Five - Ghrelin Receptor Mutations and Human Obesity', in *Progress in Molecular Biology and Translational Science*, vol. 140, Y.-X. Tao, Ed. Academic Press, 2016, pp. 131–150. doi: 10.1016/bs.pmbts.2016.02.001.

[138] D. Lin, T.-H. Chun, and L. Kang, 'Adipose extracellular matrix remodelling in obesity and insulin resistance', *Biochem Pharmacol*, vol. 119, pp. 8–16, Nov. 2016, doi: 10.1016/j.bcp.2016.05.005.

[139] Y. Deng *et al.*, 'Adipocyte Xbp1s overexpression drives uridine production and reduces obesity', *Molecular Metabolism*, vol. 11, pp. 1–17, May 2018, doi: 10.1016/j.molmet.2018.02.013.

[140] J. Kim *et al.*, 'Grasp55 −/− mice display impaired fat absorption and resistance to high-fat diet-induced obesity', *Nature Communications*, vol. 11, no. 1, Art. no. 1, Mar. 2020, doi: 10.1038/s41467-020-14912-x.

[141] M. C. Sugden and M. J. Holness, 'Role of nuclear receptors in the modulation of insulin secretion in lipid-induced insulin resistance', *Biochemical Society Transactions*, vol. 36, no. 5, pp. 891–900, Sep. 2008, doi: 10.1042/BST0360891.

[142] A. M. D'Amico and K. M. Vasquez, 'The multifaceted roles of DNA repair and replication proteins in aging and obesity', *DNA Repair*, vol. 99, p. 103049, Mar. 2021, doi: 10.1016/j.dnarep.2021.103049.

[143] L. Massip, C. Garand, R. V. N. Turaga, F. Deschênes, E. Thorin, and M. Lebel, 'Increased insulin, triglycerides, reactive oxygen species, and cardiac fibrosis in mice with a mutation in the helicase domain of the Werner syndrome gene homologue', *Experimental Gerontology*, vol. 41, no. 2, pp. 157–168, Feb. 2006, doi: 10.1016/j.exger.2005.10.011.

[144] J. Hj *et al.*, 'Gut-expressed gustducin and taste receptors regulate secretion of glucagon-like peptide-1.', *Proc Natl Acad Sci U S A*, vol. 104, no. 38, pp. 15069–15074, Aug. 2007, doi: 10.1073/pnas.0706890104.

[145] Z. Fu, E. R. Gilbert, and D. Liu, 'Regulation of Insulin Synthesis and Secretion and Pancreatic Beta-Cell Dysfunction in Diabetes', *Curr Diabetes Rev*, vol. 9, no. 1, pp. 25–53, Jan. 2013.

[146] R. Bhushan, A. Rani, A. Ali, V. K. Singh, and P. K. Dubey, 'Bioinformatics enrichment analysis of genes and pathways related to maternal type 1 diabetes associated with adverse fetal outcomes', *J Diabetes Complications*, vol. 34, no. 5, p. 107556, May 2020, doi: 10.1016/j.jdiacomp.2020.107556.

[147] K. M. Standifer and G. W. Pasternak, ‘G Proteins and Opioid Receptor-Mediated Signalling’, *Cellular Signalling*, vol. 9, no. 3, pp. 237–248, May 1997, doi: 10.1016/S0898-6568(96)00174-X.

[148] A. Singh, Y. Gibert, and K. M. Dwyer, ‘The adenosine, adrenergic and opioid pathways in the regulation of insulin secretion, beta cell proliferation and regeneration’, *Pancreatology*, vol. 18, no. 6, pp. 615–623, Sep. 2018, doi: 10.1016/j.pan.2018.06.006.

[149] C. C. Low Wang, I. Gurevich, and B. Draznin, ‘Insulin Affects Vascular Smooth Muscle Cell Phenotype and Migration Via Distinct Signaling Pathways’, *Diabetes*, vol. 52, no. 10, pp. 2562–2569, Oct. 2003, doi: 10.2337/diabetes.52.10.2562.

[150] E. Popugaeva, E. Pchitskaya, and I. Bezprozvanny, ‘Dysregulation of Intracellular Calcium Signaling in Alzheimer’s Disease’, *Antioxid Redox Signal*, vol. 29, no. 12, pp. 1176–1188, Oct. 2018, doi: 10.1089/ars.2018.7506.

[151] S. M. de la Monte and J. R. Wands, ‘Alzheimer’s Disease Is Type 3 Diabetes–Evidence Reviewed’, *J Diabetes Sci Technol*, vol. 2, no. 6, pp. 1101–1113, Nov. 2008.

[152] E. Steen *et al.*, ‘Impaired insulin and insulin-like growth factor expression and signaling mechanisms in Alzheimer’s disease--is this type 3 diabetes?’, *J Alzheimers Dis*, vol. 7, no. 1, pp. 63–80, Feb. 2005, doi: 10.3233/jad-2005-7107.

[153] L. P. van der Heide, G. M. J. Ramakers, and M. P. Smidt, ‘Insulin signaling in the central nervous system: learning to survive’, *Prog Neurobiol*, vol. 79, no. 4, pp. 205–221, Jul. 2006, doi: 10.1016/j.pneurobio.2006.06.003.

[154] K. N. Belosludtsev, N. V. Belosludtseva, and M. V. Dubinin, ‘Diabetes Mellitus, Mitochondrial Dysfunction and Ca2+-Dependent Permeability Transition Pore’, *Int J Mol Sci*, vol. 21, no. 18, Sep. 2020, doi: 10.3390/ijms21186559.

[155] E. Witsch, M. Sela, and Y. Yarden, ‘Roles for Growth Factors in Cancer Progression’, *Physiology (Bethesda)*, vol. 25, no. 2, pp. 85–101, Apr. 2010, doi: 10.1152/physiol.00045.2009.

[156] D. Cannata, Y. Fierz, A. Vijayakumar, and D. LeRoith, ‘Type 2 Diabetes and Cancer: What Is the Connection?’, *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, vol. 77, no. 2, pp. 197–213, 2010, doi: https://doi.org/10.1002/msj.20167.

[157] V. Vella *et al.*, ‘A Novel Autocrine Loop Involving IGF-II and the Insulin Receptor Isoform-A Stimulates Growth of Thyroid Cancer’, *The Journal of Clinical Endocrinology & Metabolism*, vol. 87, no. 1, pp. 245–254, Jan. 2002, doi: 10.1210/jcem.87.1.8142.

[158] M. Trevisan, J. Liu, P. Muti, G. Misciagna, A. Menotti, and F. Fucci, ‘Markers of Insulin Resistance and Colorectal Cancer Mortality’, *Cancer Epidemiol Biomarkers Prev*, vol. 10, no. 9, pp. 937–941, Sep. 2001.

[159] A. F. Tracz, C. Szczylik, C. Porta, and A. M. Czarnecka, ‘Insulin-like growth factor-1 signaling in renal cell carcinoma’, *BMC Cancer*, vol. 16, Jul. 2016, doi: 10.1186/s12885-016-2437-4.

[160] A. E. Contreras-Ferrat *et al.*, ‘An inositol 1,4,5-triphosphate (IP3)-IP3 receptor pathway is required for insulin-stimulated glucose transporter 4 translocation and glucose uptake in cardiomyocytes’, *Endocrinology*, vol. 151, no. 10, pp. 4665–4677, Oct. 2010, doi: 10.1210/en.2010-0116.

[161] J. R. Speakman, R. J. F. Loos, S. O’Rahilly, J. N. Hirschhorn, and D. B. Allison, ‘GWAS for BMI: a treasure trove of fundamental insights into the genetic basis of obesity’, *Int J Obes (Lond)*, vol. 42, no. 8, pp. 1524–1531, Aug. 2018, doi: 10.1038/s41366-018-0147-5.

[162] C. Bycroft *et al.*, ‘The UK Biobank resource with deep phenotyping and genomic data’, *Nature*, vol. 562, no. 7726, pp. 203–209, Oct. 2018, doi: 10.1038/s41586-018-0579-z.

[163] S. Ritchie, *liftOverPlink*. 2021. Accessed: Aug. 09, 2021. [Online]. Available: https://github.com/sritchie73/liftOverPlink

[164] J. R. I. Coleman, E. Krapohl, T. C. Eley, and G. Breen, ‘Individual and shared effects of social environment and polygenic risk scores on adolescent body mass index’, *Sci Rep*, vol. 8, no. 1, p. 6344, Apr. 2018, doi: 10.1038/s41598-018-24774-5.

[165] S. D. M. Brown and H. V. Lad, ‘The dark genome and pleiotropy: challenges for precision medicine’, *Mamm Genome*, vol. 30, no. 7, pp. 212–216, Aug. 2019, doi: 10.1007/s00335-019-09813-4.

[166] S. D. M. Brown and M. W. Moore, ‘The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping’, *Mamm Genome*, vol. 23, no. 9, pp. 632–640, Oct. 2012, doi: 10.1007/s00335-012-9427-x.

[167] D. W. Belsky *et al.*, ‘Polygenic Risk, Rapid Childhood Growth, and the Development of Obesity’, *Arch Pediatr Adolesc Med*, vol. 166, no. 6, pp. 515–521, Jun. 2012, doi: 10.1001/archpediatrics.2012.131.

[168] N. Chami, M. Preuss, R. W. Walker, A. Moscati, and R. J. F. Loos, ‘The role of polygenic susceptibility to obesity among carriers of pathogenic mutations in MC4R in the UK Biobank population’, *PLoS Med*, vol. 17, no. 7, p. e1003196, Jul. 2020, doi: 10.1371/journal.pmed.1003196.

[169] B. F. Darst, X. Sheng, R. A. Eeles, Z. Kote-Jarai, D. V. Conti, and C. A. Haiman, ‘Combined Effect of a Polygenic Risk Score and Rare Genetic Variants on Prostate Cancer Risk’, *Eur Urol*, pp. S0302-2838(21)00253–0, May 2021, doi: 10.1016/j.eururo.2021.04.013.

[170] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly, ‘Current clinical use of polygenic scores will risk exacerbating health disparities’, *Nat Genet*, vol. 51, no. 4, pp. 584–591, Apr. 2019, doi: 10.1038/s41588-019-0379-x.

[171] A. L. Wise, L. Gyi, and T. A. Manolio, ‘eXclusion: Toward Integrating the X Chromosome in Genome-wide Association Analyses’, *The American Journal of Human Genetics*, vol. 92, no. 5, pp. 643–647, May 2013, doi: 10.1016/j.ajhg.2013.03.017.

[172] M. I. McCarthy, ‘Painting a new picture of personalised medicine for diabetes’, *Diabetologia*, vol. 60, no. 5, pp. 793–799, 2017, doi: 10.1007/s00125-017-4210-x.

[173] K. J. Karczewski *et al.*, ‘The mutational constraint spectrum quantified from variation in 141,456 humans’, *Nature*, vol. 581, no. 7809, Art. no. 7809, May 2020, doi: 10.1038/s41586-020-2308-7.

[174] S. A. Lambert *et al.*, ‘The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation’, *Nat Genet*, vol. 53, no. 4, Art. no. 4, Apr. 2021, doi: 10.1038/s41588-021-00783-5.

[175] R. Argelaguet *et al.*, ‘Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets’, *Mol Syst Biol*, vol. 14, no. 6, p. e8124, Jun. 2018, doi: 10.15252/msb.20178124.

[176] T. Townshend and A. Lake, ‘Obesogenic environments: current evidence of the built and food environments’, *Perspect Public Health*, vol. 137, no. 1, pp. 38–44, Jan. 2017, doi: 10.1177/1757913916679860.

[177] The NICUSeq Study Group, ‘Effect of Whole-Genome Sequencing on the Clinical Management of Acutely Ill Infants With Suspected Genetic Disease: A Randomized Clinical Trial’, *JAMA Pediatrics*, Sep. 2021, doi: 10.1001/jamapediatrics.2021.3496.

[178] K. G. Fulda and K. Lykens, ‘Ethical issues in predictive genetic testing: a public health perspective’, *J Med Ethics*, vol. 32, no. 3, pp. 143–147, Mar. 2006, doi: 10.1136/jme.2004.010272.

[179] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel, ‘Most mammalian mRNAs are conserved targets of microRNAs’, *Genome Res*, vol. 19, no. 1, pp. 92–105, Jan. 2009, doi: 10.1101/gr.082701.108.

[180] B. P. Lewis, C. B. Burge, and D. P. Bartel, ‘Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets’, *Cell*, vol. 120, no. 1, pp. 15–20, Jan. 2005, doi: 10.1016/j.cell.2004.12.035.

[181] R. C. Lee, R. L. Feinbaum, and V. Ambros, ‘The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14’, *Cell*, vol. 75, no. 5, pp. 843–854, Dec. 1993, doi: 10.1016/0092-8674(93)90529-y.

[182] B. J. Reinhart *et al.*, ‘The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans’, *Nature*, vol. 403, no. 6772, pp. 901–906, Feb. 2000, doi: 10.1038/35002607.

[183] A. E. Pasquinelli *et al.*, ‘Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA’, *Nature*, vol. 408, no. 6808, pp. 86–89, Nov. 2000, doi: 10.1038/35040556.

[184] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl, ‘Identification of novel genes coding for small expressed RNAs’, *Science*, vol. 294, no. 5543, pp. 853–858, Oct. 2001, doi: 10.1126/science.1064921.

[185] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel, ‘An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans’, *Science*, vol. 294, no. 5543, pp. 858–862, Oct. 2001, doi: 10.1126/science.1065062.

[186] R. C. Lee and V. Ambros, ‘An Extensive Class of Small RNAs in Caenorhabditis elegans’, *Science*, vol. 294, no. 5543, pp. 862–864, Oct. 2001, doi: 10.1126/science.1065329.

[187] V. N. Kim, ‘MicroRNA biogenesis: coordinated cropping and dicing’, *Nature Reviews Molecular Cell Biology*, vol. 6, no. 5, Art. no. 5, May 2005, doi: 10.1038/nrm1644.

[188] O. S. Rissland *et al.*, ‘The influence of microRNAs and poly(A) tail length on endogenous mRNA–protein complexes’, *Genome Biology*, vol. 18, no. 1, p. 211, Oct. 2017, doi: 10.1186/s13059-017-1330-z.

[189] S. Gu and M. A. Kay, ‘How do miRNAs mediate translational repression?’, *Silence*, vol. 1, no. 1, p. 11, May 2010, doi: 10.1186/1758-907X-1-11.

[190] T. P. Chendrimada *et al.*, ‘MicroRNA silencing through RISC recruitment of eIF6’, *Nature*, vol. 447, no. 7146, Art. no. 7146, Jun. 2007, doi: 10.1038/nature05841.

[191] D. P. Bartel, ‘MicroRNA Target Recognition and Regulatory Functions’, *Cell*, vol. 136, no. 2, pp. 215–233, Jan. 2009, doi: 10.1016/j.cell.2009.01.002.

[192] M. S. Ebert and P. A. Sharp, ‘MicroRNA sponges: Progress and possibilities’, *RNA*, vol. 16, no. 11, pp. 2043–2050, Nov. 2010, doi: 10.1261/rna.2414110.

[193] A. J. Giraldez *et al.*, ‘MicroRNAs Regulate Brain Morphogenesis in Zebrafish’, *Science*, vol. 308, no. 5723, pp. 833–838, May 2005, doi: 10.1126/science.1109020.

[194] L. Ma *et al.*, ‘An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice.’, *Nat Genet*, vol. 41, no. 5, pp. 614–618, Apr. 2009, doi: 10.1038/ng.369.

[195] Á. Mencía *et al.*, ‘Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss’, *Nature Genetics*, vol. 41, no. 5, Art. no. 5, May 2009, doi: 10.1038/ng.355.

[196] L. He *et al.*, ‘A microRNA polycistron as a potential human oncogene’, *Nature*, vol. 435, no. 7043, Art. no. 7043, Jun. 2005, doi: 10.1038/nature03552.

[197] G. A. Calin *et al.*, ‘Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers’, *PNAS*, vol. 101, no. 9, pp. 2999–3004, Mar. 2004, doi: 10.1073/pnas.0307323101.

[198] R. D. Morin *et al.*, ‘Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells’, *Genome Res*, vol. 18, no. 4, pp. 610–621, Apr. 2008, doi: 10.1101/gr.7179508.

[199] H. Y. Lee and J. A. Doudna, ‘TRBP alters human precursor microRNA processing in vitro’, *RNA*, vol. 18, no. 11, pp. 2012–2019, Nov. 2012, doi: 10.1261/rna.035501.112.

[200] L. Chen, L. Heikkinen, C. Wang, Y. Yang, H. Sun, and G. Wong, ‘Trends in the development of miRNA bioinformatics tools’, *Brief Bioinform*, vol. 20, no. 5, pp. 1836–1852, 27 2019, doi: 10.1093/bib/bby054.

[201] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, ‘miRBase: from microRNA sequences to function’, *Nucleic Acids Res*, vol. 47, no. Database issue, pp. D155–D162, Jan. 2019, doi: 10.1093/nar/gky1141.

[202] M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky, ‘miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades’, *Nucleic Acids Res*, vol. 40, no. 1, pp. 37–52, Jan. 2012, doi: 10.1093/nar/gkr688.

[203] Y. Chen and X. Wang, ‘miRDB: an online database for prediction of functional microRNA targets’, *Nucleic Acids Res*, vol. 48, no. D1, pp. D127–D131, Jan. 2020, doi: 10.1093/nar/gkz757.

[204] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, ‘edgeR: a Bioconductor package for differential expression analysis of digital gene expression data’, *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.

[205] M. E. Ritchie *et al.*, ‘limma powers differential expression analyses for RNA-sequencing and microarray studies’, *Nucleic Acids Research*, vol. 43, no. 7, pp. e47–e47, Apr. 2015, doi: 10.1093/nar/gkv007.

[206] M. I. Love, W. Huber, and S. Anders, ‘Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2’, *Genome Biol*, vol. 15, no. 12, p. 550, 2014, doi: 10.1186/s13059-014-0550-8.

[207] M. Ziemann, A. Kaspi, and A. El-Osta, ‘Evaluation of microRNA alignment techniques’, *Rna-A Publication of the Rna Society*, vol. 22, no. 8, pp. 1120–1138, Aug. 2016, doi: 10.1261/rna.055509.115.

[208] M. Douglas-Escobar and M. D. Weiss, ‘Hypoxic-ischemic encephalopathy: a review for the clinician’, *JAMA Pediatr*, vol. 169, no. 4, pp. 397–403, Apr. 2015, doi: 10.1001/jamapediatrics.2014.3269.

[209] E. P. Yıldız, B. Ekici, and B. Tatlı, ‘Neonatal hypoxic ischemic encephalopathy: an update on disease pathogenesis and treatment’, *Expert Rev Neurother*, vol. 17, no. 5, pp. 449–459, 2017, doi: 10.1080/14737175.2017.1259567.

[210] A. J. Gunn and M. Thoresen, ‘Chapter 10 - Neonatal encephalopathy and hypoxic–ischemic encephalopathy’, in *Handbook of Clinical Neurology*, vol. 162, L. S. de Vries and H. C. Glass, Eds. Elsevier, 2019, pp. 217–237. doi: 10.1016/B978-0-444-64029-1.00010-2.

[211] K. R. Gopagondanahalli *et al.*, ‘Preterm Hypoxic–Ischemic Encephalopathy’, *Front Pediatr*, vol. 4, Oct. 2016, doi: 10.3389/fped.2016.00114.

[212] L. Bennet, L. Booth, and A. J. Gunn, ‘Potential biomarkers for hypoxic–ischemic encephalopathy’, *Seminars in Fetal and Neonatal Medicine*, vol. 15, no. 5, pp. 253–260, Oct. 2010, doi: 10.1016/j.siny.2010.05.007.

[213] V. Ponnusamy and P. K. Yip, ‘The role of microRNAs in newborn brain development and hypoxic ischaemic encephalopathy’, *Neuropharmacology*, vol. 149, pp. 55–65, May 2019, doi: 10.1016/j.neuropharm.2018.11.041.

[214] A. Nuñez-Ramiro *et al.*, ‘Topiramate plus Cooling for Hypoxic-Ischemic Encephalopathy: A Randomized, Controlled, Multicenter, Double-Blinded Trial’, *Neonatology*, vol. 116, no. 1, pp. 76–84, 2019, doi: 10.1159/000499084.

[215] Y. Zhang, G. Parmigiani, and W. E. Johnson, ‘ComBat-seq: batch effect adjustment for RNA-seq count data’, *NAR Genom Bioinform*, vol. 2, no. 3, Sep. 2020, doi: 10.1093/nargab/lqaa078.

[216] R. Petri, J. Malmevik, L. Fasching, M. Åkerblom, and J. Jakobsson, ‘miRNAs in brain development’, *Experimental Cell Research*, vol. 321, no. 1, pp. 84–89, Feb. 2014, doi: 10.1016/j.yexcr.2013.09.022.

[217] H. Zhang, J. Zhou, M. Zhang, Y. Yi, and B. He, ‘Upregulation of miR-376c-3p alleviates oxygen–glucose deprivation-induced cell injury by targeting ING5’, *Cellular & Molecular Biology Letters*, vol. 24, no. 1, p. 67, Dec. 2019, doi: 10.1186/s11658-019-0189-2.

[218] H. Gu *et al.*, ‘Hypoxia-responsive miR-124 and miR-144 reduce hypoxia-induced autophagy and enhance radiosensitivity of prostate cancer cells via suppressing PIM1’, *Cancer Med*, vol. 5, no. 6, pp. 1174–1182, Mar. 2016, doi: 10.1002/cam4.664.

[219] C. Gao *et al.*, ‘Hematoma-derived exosomes of chronic subdural hematoma promote abnormal angiogenesis and inhibit hematoma absorption through miR-144-5p’, *Aging (Albany NY)*, vol. 11, no. 24, pp. 12147–12164, Dec. 2019, doi: 10.18632/aging.102550.

[220] X. Wang, J. Chen, and X. Huang, ‘Rosuvastatin Attenuates Myocardial Ischemia-Reperfusion Injury via Upregulating miR-17-3p-Mediated Autophagy’, *Cell Reprogram*, vol. 21, no. 6, pp. 323–330, Dec. 2019, doi: 10.1089/cell.2018.0053.

[221] E. Aguado-Fraile *et al.*, ‘miR-127 Protects Proximal Tubule Cells against Ischemia/Reperfusion: Identification of Kinesin Family Member 3B as miR-127 Target’, *PLoS One*, vol. 7, no. 9, Sep. 2012, doi: 10.1371/journal.pone.0044305.

[222] B. Icli *et al.*, ‘MicroRNA-615-5p Regulates Angiogenesis and Tissue Repair by Targeting AKT/eNOS (Protein Kinase B/Endothelial Nitric Oxide Synthase) Signaling in Endothelial Cells’, *Arterioscler Thromb Vasc Biol*, vol. 39, no. 7, pp. 1458–1474, 2019, doi: 10.1161/ATVBAHA.119.312726.

[223] Z. WU *et al.*, ‘Differential effects of miR-34c-3p and miR-34c-5p on the proliferation, apoptosis and invasion of glioma cells’, *Oncol Lett*, vol. 6, no. 5, pp. 1447–1452, Nov. 2013, doi: 10.3892/ol.2013.1579.

[224] X.-B. Guo, X.-C. Zhang, P. Chen, L.-M. Ma, and Z.-Q. Shen, ‘miR-378a-3p inhibits cellular proliferation and migration in glioblastoma multiforme by targeting tetraspanin 17’, *Oncol Rep*, vol. 42, no. 5, pp. 1957–1971, Nov. 2019, doi: 10.3892/or.2019.7283.

[225] S. Swarbrick, N. Wragg, S. Ghosh, and A. Stolzing, ‘Systematic Review of miRNA as Biomarkers in Alzheimer’s Disease’, *Mol Neurobiol*, vol. 56, no. 9, pp. 6156–6167, Sep. 2019, doi: 10.1007/s12035-019-1500-y.

[226] A. T. Barros-Viegas *et al.*, ‘miRNA-31 Improves Cognition and Abolishes Amyloid-β Pathology by Targeting APP and BACE1 in an Animal Model of Alzheimer’s Disease’, *Mol Ther Nucleic Acids*, vol. 19, pp. 1219–1236, Mar. 2020, doi: 10.1016/j.omtn.2020.01.010.

[227] J. Ai *et al.*, ‘MicroRNA-195 Protects Against Dementia Induced by Chronic Brain Hypoperfusion via Its Anti-Amyloidogenic Effect in Rats’, *J Neurosci*, vol. 33, no. 9, pp. 3989–4001, Feb. 2013, doi: 10.1523/JNEUROSCI.1997-12.2013.

[228] A.-Y. Guo, J. Sun, P. Jia, and Z. Zhao, ‘A Novel microRNA and transcription factor mediated regulatory network in schizophrenia’, *BMC Syst Biol*, vol. 4, p. 10, Feb. 2010, doi: 10.1186/1752-0509-4-10.

[229] S. Huang *et al.*, ‘miR-129-2-3p directly targets SYK gene and associates with the risk of ischaemic stroke in a Chinese population’, *J Cell Mol Med*, vol. 23, no. 1, pp. 167–176, Jan. 2019, doi: 10.1111/jcmm.13901.

[230] Y. Chen *et al.*, ‘Deep Sequencing of Small RNAs in Blood of Patients with Brain Arteriovenous Malformations’, *World Neurosurg*, vol. 115, pp. e570–e579, Jul. 2018, doi: 10.1016/j.wneu.2018.04.097.

[231] J. Wang, M. Lin, H. Ren, Z. Yu, T. Guo, and B. Gu, ‘Expression and Clinical Significance of Serum miR-497 in Patients with Acute Cerebral Infarction.’, *Clinical laboratory*, 2019, Accessed: Nov. 24, 2020. [Online]. Available: https://dx.doi.org/10.7754/Clin.Lab.2018.181001

[232] J. Song and Y.-K. Kim, ‘Identification of the Role of miR-142-5p in Alzheimer’s Disease by Comparative Bioinformatics and Cellular Analysis’, *Front. Mol. Neurosci.*, vol. 10, 2017, doi: 10.3389/fnmol.2017.00227.

[233] A.-M. Looney *et al.*, ‘Downregulation of Umbilical Cord Blood Levels of miR-374a in Neonatal Hypoxic Ischemic Encephalopathy’, *J Pediatr*, vol. 167, no. 2, pp. 269-273.e2, Aug. 2015, doi: 10.1016/j.jpeds.2015.04.060.

[234] H. Luo *et al.*, ‘miR-7-5p overexpression suppresses cell proliferation and promotes apoptosis through inhibiting the ability of DNA damage repair of PARP-1 and BRCA1 in TK6 cells exposed to hydroquinone’, *Chemico-Biological Interactions*, vol. 283, pp. 84–90, Mar. 2018, doi: 10.1016/j.cbi.2018.01.019.

[235] J.-H. Kim *et al.*, ‘Hypoxia-responsive microRNA-101 promotes angiogenesis via heme oxygenase-1/vascular endothelial growth factor axis by targeting cullin 3’, *Antioxid Redox Signal*, vol. 21, no. 18, pp. 2469–2482, Dec. 2014, doi: 10.1089/ars.2014.5856.

[236] Q. Cai, T. Wang, W. Yang, and X. Fen, ‘Protective mechanisms of microRNA-27a against oxygen-glucose deprivation-induced injuries in hippocampal neurons’, *Neural Regen Res*, vol. 11, no. 8, pp. 1285–1292, Aug. 2016, doi: 10.4103/1673-5374.189194.

[237] P. Piscopo *et al.*, ‘Reduced miR-659-3p Levels Correlate with Progranulin Increase in Hypoxic Conditions: Implications for Frontotemporal Dementia.’, *Front Mol Neurosci*, vol. 9, pp. 31–31, May 2016, doi: 10.3389/fnmol.2016.00031.

[238] T. B. Hansen *et al.*, ‘miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA’, *EMBO J*, vol. 30, no. 21, pp. 4414–4422, Nov. 2011, doi: 10.1038/emboj.2011.359.

[239] P. Bosco, R. Spada, S. Caniglia, M. G. Salluzzo, and M. Salemi, ‘Cerebellar degeneration-related autoantigen 1 (CDR1) gene expression in Alzheimer’s disease’, *Neurol Sci*, vol. 35, no. 10, pp. 1613–1614, Oct. 2014, doi: 10.1007/s10072-014-1805-6.

[240] A. Hodges *et al.*, ‘Regional and cellular gene expression changes in human Huntington’s disease brain’, *Hum Mol Genet*, vol. 15, no. 6, pp. 965–977, Mar. 2006, doi: 10.1093/hmg/ddl013.

[241] Q. Wei *et al.*, ‘MicroRNA-668 represses MTP18 to preserve mitochondrial dynamics in ischemic acute kidney injury’, *J Clin Invest*, vol. 128, no. 12, pp. 5448–5464, 03 2018, doi: 10.1172/JCI121859.

[242] J. He and X. Zhang, ‘miR-668 inhibitor attenuates mitochondrial membrane potential and protects against neuronal apoptosis in cerebral ischemic stroke’, *Folia Neuropathol*, vol. 58, no. 1, pp. 22–29, 2020, doi: 10.5114/fn.2020.94003.

[243] J. Yan *et al.*, ‘Screening the expression of several miRNAs from TaqMan Low Density Array in traumatic brain injury: miR-219a-5p regulates neuronal apoptosis by modulating CCNA2 and CACUL1’, *J Neurochem*, vol. 150, no. 2, pp. 202–217, 2019, doi: 10.1111/jnc.14717.

[244] O. Hamamoto *et al.*, ‘Modulation of NMDA receptor by miR-219 in the amygdala and hippocampus of patients with mesial temporal lobe epilepsy’, *J Clin Neurosci*, vol. 74, pp. 180–186, Apr. 2020, doi: 10.1016/j.jocn.2020.02.024.

[245] M. E. Talkowski *et al.*, ‘Assessment of 2q23.1 Microdeletion Syndrome Implicates MBD5 as a Single Causal Locus of Intellectual Disability, Epilepsy, and Autism Spectrum Disorder’, *The American Journal of Human Genetics*, vol. 89, no. 4, pp. 551–563, Oct. 2011, doi: 10.1016/j.ajhg.2011.09.011.

[246] G. L. Carvill *et al.*, ‘Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1’, *Nat Genet*, vol. 45, no. 7, pp. 825–830, Jul. 2013, doi: 10.1038/ng.2646.

[247] E. Abs *et al.*, ‘TORC1-dependent epilepsy caused by acute biallelic Tsc1 deletion in adult mice’, *Annals of Neurology*, vol. 74, no. 4, pp. 569–579, 2013, doi: https://doi.org/10.1002/ana.23943.

[248] M. Papadakis *et al.*, ‘Tsc1 (hamartin) confers neuroprotection against ischemia by inducing autophagy’, *Nat Med*, vol. 19, no. 3, pp. 351–357, Mar. 2013, doi: 10.1038/nm.3097.

[249] S. V. Mullegama *et al.*, ‘De novo loss-of-function variants in STAG2 are associated with developmental delay, microcephaly, and congenital anomalies’, *American Journal of Medical Genetics Part A*, vol. 173, no. 5, pp. 1319–1327, 2017, doi: https://doi.org/10.1002/ajmg.a.38207.

[250] P. Kruszka *et al.*, ‘Cohesin complex-associated holoprosencephaly’, *Brain*, vol. 142, no. 9, pp. 2631–2643, Sep. 2019, doi: 10.1093/brain/awz210.

[251] D. Bergman, M. Halje, M. Nordin, and W. Engström, ‘Insulin-Like Growth Factor 2 in Development and Disease: A Mini-Review’, *GER*, vol. 59, no. 3, pp. 240–249, 2013, doi: 10.1159/000343995.

Appendix A

A.1 Enrichment analyses for diabetes type 2 and obesity

Table A.1-1: List of GO terms significantly enriched (FDR < 0.05) in the 27 genes related to type 2 diabetes mellitus (OMIM: 125853) compared to all the Homo sapiens genes in the PANTHER database. The annotation is based on the GO Ontology database (Released 2021-07-02). The list was trimmed to only keep the most specific terms. The enrichment was done with Fisher's Exact Test and the False Discovery Rate (FDR) was calculated to correct for multiple comparisons.

GO biological process	Homo sapiens (20595)	T2DM (27)	expected	Fold Enrichment	+/-	raw P-value	FDR
hepatocyte differentiation	14	2	0.02	> 100	+	2.26E-04	1.49E-02
negative regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway	20	3	0.03	> 100	+	4.34E-06	5.26E-04
positive regulation of glycogen biosynthetic process	16	4	0.02	> 100	+	1.50E-08	4.55E-06
positive regulation of fatty acid beta-oxidation	11	3	0.02	> 100	+	9.00E-07	1.36E-04
negative regulation of long-chain fatty acid import across plasma membrane	4	2	0.01	> 100	+	2.85E-05	2.54E-03
negative regulation of type B pancreatic cell apoptotic process	6	4	0.01	> 100	+	6.58E-10	3.05E-07
detection of glucose	3	2	0	> 100	+	1.90E-05	1.83E-03
insulin secretion	37	5	0.05	96	+	3.14E-09	1.12E-06
negative regulation of endoplasmic reticulum unfolded protein response	16	2	0.02	89	+	2.88E-04	1.85E-02
reverse cholesterol transport	18	2	0.03	79	+	3.57E-04	2.23E-02
negative regulation of insulin secretion	39	4	0.05	73	+	3.74E-07	6.48E-05
nitric oxide mediated signal transduction	20	2	0.03	71	+	4.33E-04	2.61E-02
type B pancreatic cell differentiation	21	2	0.03	68	+	4.74E-04	2.82E-02
glucose 6-phosphate metabolic process	22	2	0.03	65	+	5.17E-04	3.04E-02
signal transduction involved in regulation of gene expression	22	2	0.03	65	+	5.17E-04	3.02E-02
negative regulation of insulin receptor signaling pathway	34	3	0.05	63	+	1.88E-05	1.82E-03

NADH metabolic process	24	2	0.03	59	+	6.08E-04	3.44E-02
negative regulation of receptor signaling pathway via STAT	25	2	0.04	57	+	6.56E-04	3.69E-02
NAD metabolic process	25	2	0.04	57	+	6.56E-04	3.67E-02
positive regulation of glucose import	38	3	0.05	56	+	2.57E-05	2.34E-03
positive regulation of insulin secretion	79	6	0.11	54	+	1.80E-09	7.28E-07
carbohydrate phosphorylation	27	2	0.04	53	+	7.57E-04	4.17E-02
positive regulation of transcription initiation from RNA polymerase II promoter	29	2	0.04	49	+	8.66E-04	4.70E-02
insulin receptor signaling pathway	60	4	0.08	47	+	1.89E-06	2.63E-04
positive regulation of DNA binding	58	3	0.08	37	+	8.52E-05	6.47E-03
regulation of JUN kinase activity	57	3	0.08	37	+	8.10E-05	6.28E-03
negative regulation of MAP kinase activity	57	3	0.08	37	+	8.10E-05	6.25E-03
pyruvate metabolic process	61	3	0.09	35	+	9.83E-05	7.20E-03
glucose metabolic process	101	5	0.14	35	+	3.51E-07	6.21E-05
cellular response to glucose stimulus	74	3	0.1	29	+	1.70E-04	1.16E-02
fat cell differentiation	103	4	0.15	28	+	1.47E-05	1.48E-03
carbohydrate transport	78	3	0.11	27	+	1.98E-04	1.31E-02
regulation of potassium ion transport	101	3	0.14	21	+	4.14E-04	2.53E-02
carbohydrate catabolic process	104	3	0.15	20	+	4.50E-04	2.70E-02
lipid homeostasis	154	4	0.22	18	+	6.78E-05	5.31E-03
regulation of circadian rhythm	122	3	0.17	17	+	7.08E-04	3.91E-02
response to drug	394	7	0.55	13	+	1.08E-06	1.55E-04
generation of precursor metabolites and energy	391	7	0.55	13	+	1.02E-06	1.50E-04
rhythmic process	269	4	0.38	11	+	5.50E-04	3.18E-02
regulation of transporter activity	287	4	0.4	10	+	6.98E-04	3.88E-02
purine ribonucleotide metabolic process	306	4	0.43	9	+	8.84E-04	4.78E-02
regulation of DNA-binding transcription factor activity	424	5	0.6	8	+	2.99E-04	1.88E-02
regulation of cell population proliferation	1654	10	2.33	4	+	5.46E-05	4.37E-03
negative regulation of transcription, DNA-templated	1322	8	1.86	4	+	3.71E-04	2.30E-02

Table A.1-2: List of GO terms significantly enriched (FDR < 0.05) in the 20 genes related to obesity (11 genes related to 'obesity' (OMIM: 601665) and 9 genes related to 'body mass index quantitative trait locus' (OMIMs: 602025, 607447, 607514, 612362, 612460, 614411, 615457, 617885, and 618406)) compared to all the Homo sapiens genes in the PANTHER database. The annotation is based on the GO Ontology database (Released 2021-07-02). The list was trimmed to only keep the most specific terms. The enrichment was done with Fisher's Exact Test and the False Discovery Rate (FDR) was calculated to correct for multiple comparisons.

GO biological process	Homo sapiens (20595)	Obesity (20)	expected	Fold Enrichment	+/-	raw P-value	FDR
norepinephrine-epinephrine-mediated vasodilation involved in regulation of systemic arterial blood pressure	3	2	0	> 100	+	8.93E-06	3.12E-03
response to melanocyte-stimulating hormone	3	2	0	> 100	+	8.93E-06	3.05E-03
adult feeding behavior	10	3	0.01	> 100	+	2.22E-07	1.94E-04
positive regulation of feeding behavior	10	2	0.01	> 100	+	5.87E-05	1.15E-02
regulation of glucagon secretion	10	2	0.01	> 100	+	5.87E-05	1.14E-02
regulation of eating behavior	10	2	0.01	> 100	+	5.87E-05	1.13E-02
regulation of glucocorticoid secretion	11	2	0.01	> 100	+	6.93E-05	1.25E-02
diet induced thermogenesis	12	2	0.01	> 100	+	8.08E-05	1.40E-02
white fat cell differentiation	14	2	0.01	> 100	+	1.06E-04	1.80E-02
regulation of appetite	22	3	0.02	> 100	+	1.77E-06	8.71E-04
regulation of transmission of nerve impulse	16	2	0.02	>100	+	1.36E-04	2.24E-02
negative regulation of behavior	17	2	0.02	> 100	+	1.51E-04	2.41E-02
regulation of response to food	18	2	0.02	> 100	+	1.68E-04	2.49E-02
response to superoxide	18	2	0.02	> 100	+	1.68E-04	2.47E-02
positive regulation of cAMP-mediated signaling	19	2	0.02	> 100	+	1.86E-04	2.66E-02
temperature homeostasis	29	3	0.03	> 100	+	3.80E-06	1.62E-03
adenylate cyclase-activating adrenergic receptor signaling pathway	21	2	0.02	98.07	+	2.23E-04	3.08E-02
regulation of vascular endothelial cell proliferation	21	2	0.02	98.07	+	2.23E-04	3.06E-02
negative regulation of peptide hormone secretion	44	4	0.04	93.61	+	1.22E-07	1.48E-04
regulation of brown fat cell differentiation	23	2	0.02	89.54	+	2.65E-04	3.47E-02

response to cold	50	4	0.05	82.38	+	1.97E-07	1.94E-04
negative regulation of interleukin-1 beta production	29	2	0.03	71.02	+	4.09E-04	4.69E-02
peptide hormone secretion	57	3	0.06	54.2	+	2.58E-05	6.44E-03
positive regulation of cold-induced thermogenesis	97	5	0.09	53.08	+	3.93E-08	6.17E-05
energy reserve metabolic process	63	3	0.06	49.04	+	3.44E-05	7.83E-03
regulation of multicellular organism growth	64	3	0.06	48.27	+	3.60E-05	7.96E-03
circadian regulation of gene expression	68	3	0.07	45.43	+	4.28E-05	9.09E-03
circadian rhythm	136	4	0.13	30.29	+	9.07E-06	3.04E-03
positive regulation of peptide hormone secretion	106	3	0.1	29.14	+	1.54E-04	2.41E-02
neuropeptide signaling pathway	109	3	0.11	28.34	+	1.66E-04	2.52E-02
hormone-mediated signaling pathway	123	3	0.12	25.12	+	2.36E-04	3.14E-02
regulation of insulin secretion	168	4	0.16	24.52	+	2.04E-05	5.35E-03
negative regulation of inflammatory response	141	3	0.14	21.91	+	3.49E-04	4.15E-02
glucose homeostasis	194	4	0.19	21.23	+	3.55E-05	7.97E-03
cellular response to insulin stimulus	149	3	0.14	20.73	+	4.09E-04	4.72E-02
lipid localization	335	4	0.33	12.3	+	2.83E-04	3.64E-02
positive regulation of MAPK cascade	471	5	0.46	10.93	+	7.48E-05	1.32E-02

A.2 Personalised nutrition

Table A.2-1: List of items included in the Mediterranean Diet Adherence Screener

Do you use olive oil as the principal source of fat for cooking?
Do you drink wine? How much do you consume per week?
How many servings (150g) of pulses do you consume per week?
How many servings of vegetables do you consume per day? A full serving is 200g.
How many servings (12g) of butter, margarine, or cream do you consume per day?
How many servings of fish/seafood do you consume per week? (100–150g of fish, 4–5 pieces or 200g of seafood)
Do you prefer to eat chicken, turkey, or rabbit instead of beef, pork, hamburgers, or sausages?
How much olive oil do you consume per day?
How many pieces of fruit do you consume per day?
How many times do you consume nuts per week? (1 serving = 30g)
How many carbonated and/or sugar-sweetened beverages do you consume per day?
How many servings of red meat, hamburger, or sausages do you consume per day? A full serving is 100–150g.
How many times do you consume commercial (not homemade) pastry such as cookies or cake per week?
How many times per week do you consume boiled vegetables, pasta, rice, or other dishes with a sauce of tomato, garlic, onion, or leeks sautéed in olive oil?

Appendix B

B.1 VarGen

Table B.1-1: List of data sources accessed by MyVariant.info to perform the annotation of variants.

dbNSFP	DOCM	Cancer Genome Interpreter
dbSNP	SNPedia	genome Aggregation Database
ClinVar	EMVClass	CIViC
EVS	Wellderly	Geno2MP
CADD	ExAC	GWAS Catalog
MutDB	GRASP	UniProt
COSMIC		

Table B.1-2: List of tissues with significant variant-gene pairs from the GTEx project

Adipose Subcutaneous	Brain Spinal cord cervical c-1
Adipose Visceral Omentum	Brain Substantia nigra
Adrenal Gland	Breast Mammary Tissue
Artery Aorta	Cells Cultured fibroblasts
Artery Coronary	Cells EBV-transformed lymphocytes
Artery Tibial	Colon Sigmoid
Brain Amygdala	Colon Transverse
Brain Anterior cingulate cortex BA24	Esophagus Gastroesophageal Junction
Brain Caudate basal ganglia	Esophagus Mucosa
Brain Cerebellar Hemisphere	Esophagus Muscularis
Brain Cerebellum	Heart_Atrial Appendage
Brain Cortex	Heart Left Ventricle
Brain Frontal Cortex BA9	Kidney Cortex
Brain Hippocampus	Liver
Brain Nucleus accumbens basal ganglia	Lung
Brain Putamen basal ganglia	Minor Salivary Gland
Muscle Skeletal	Small Intestine Terminal Ileum
Nerve Tibial	Spleen
Ovary	Stomach
Pancreas	Testis
Pituitary	Thyroid
Prostate	Uterus
Skin Not Sun Exposed Suprapubic	Vagina
Skin Sun Exposed Lower leg	Whole Blood

Table B.1-3: Example of raw output from VarGen

rsid	chr	pos	ensembl_gene_id	hgnc_symbol	source	trait
rs1044548	chr6	131890623	ENSG00000197594	ENPP1	omim	601665
rs1044737470	chr5	96412463	ENSG00000175426	PCSK1	omim	601665
rs1045328134	chr8	37966251	ENSG00000188778	ADRB3	omim	601665
rs1045550142	chr16	54040160	ENSG00000140718	FTO	omim	601665
rs111340993	chr3	12170213	ENSG00000132170	PPARG	fantom5	601665
rs10195271	chr2	24885781	ENSG00000138031	ADCY3	gtex (Adipose_Subcutaneous)	601665
rs1063429	chr3	10279284	ENSG00000157017	GHRL	gtex (Adipose_Visceral_Omentum)	601665
rs11155053	chr6	139338875	ENSG00000218565, ENSG00000226571	AL592429.1, AL592429.2	gwas	Obesity-related traits
rs12295638	Chr11	26583784	ENSG00000134343	ANO3	gwas	Obesity (extreme)

Table B.1-4: Example of annotation obtained with VarGen.

ref	alt	cadd_phred	*fathmm_xf_score*	*fathmm_xd_pred*	annot_type	consequence	clinical_sign	snpeff_ann
G	C	10.33	*NA*	*NA*	Transcript	3PRIME_UTR	Benign	
A	G	32.00	0.952236	D	CodingTranscript	NON_SYNONYMOUS		MODERATE; MODIFIER
C	G	28.00	0.754983	D	CodingTranscript	NON_SYNONYMOUS		MODERATE; LOW
T	A	11.77	*NA*	*NA*	CodingTranscript	SYNONYMOUS		LOW; LOW
C	A	13.23	*NA*	*NA*	NonCodingTranscript	NONCODING_CHANGE		MODIFIER; MODIFIER
C	T	38.00	0.277287	N	CodingTranscript	STOP_GAINED	Pathogenic	HIGH; HIGH

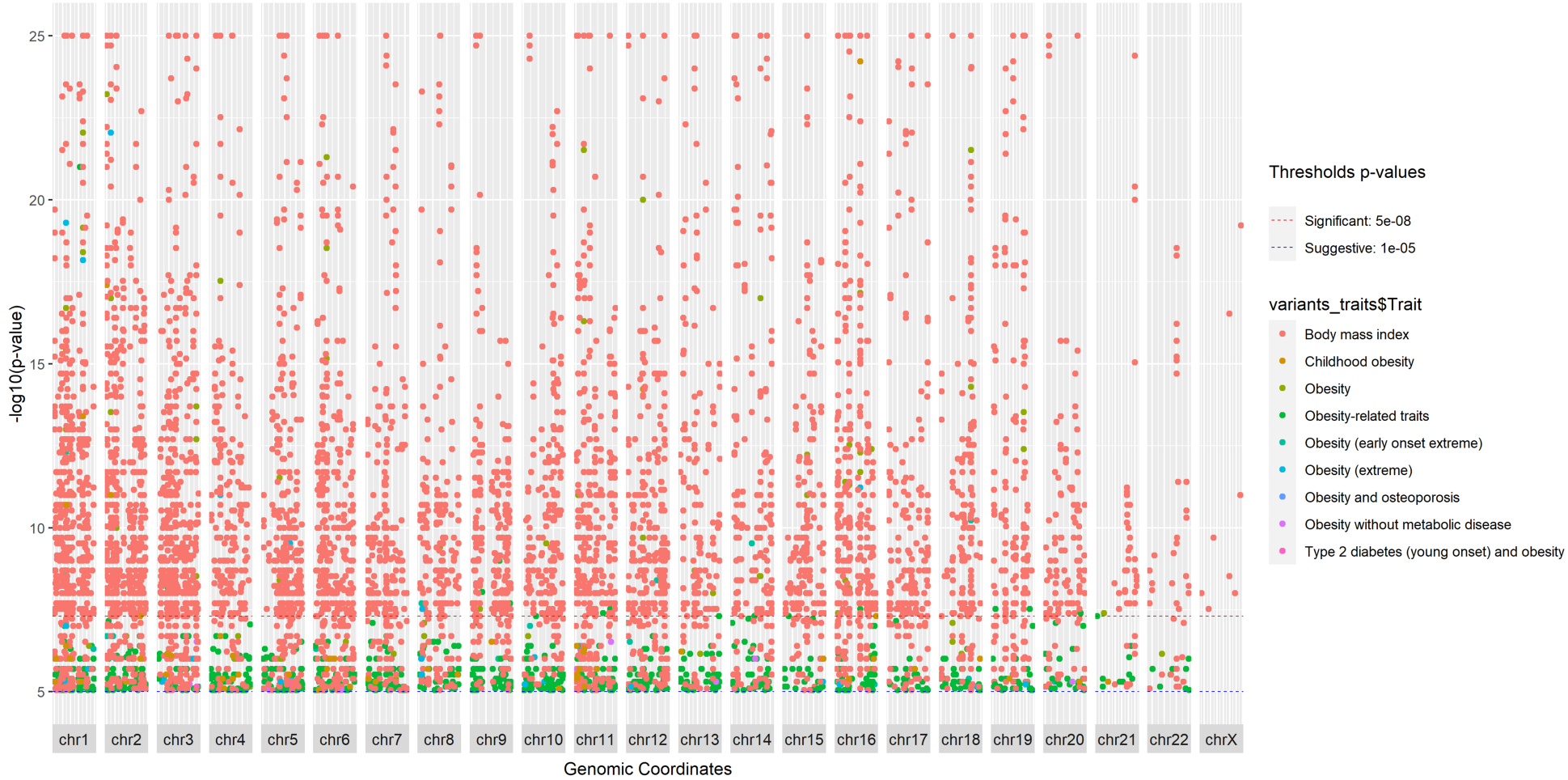


Figure B.1-1: Example of Manhattan plot obtained with VarGen for GWAS traits related to obesity.

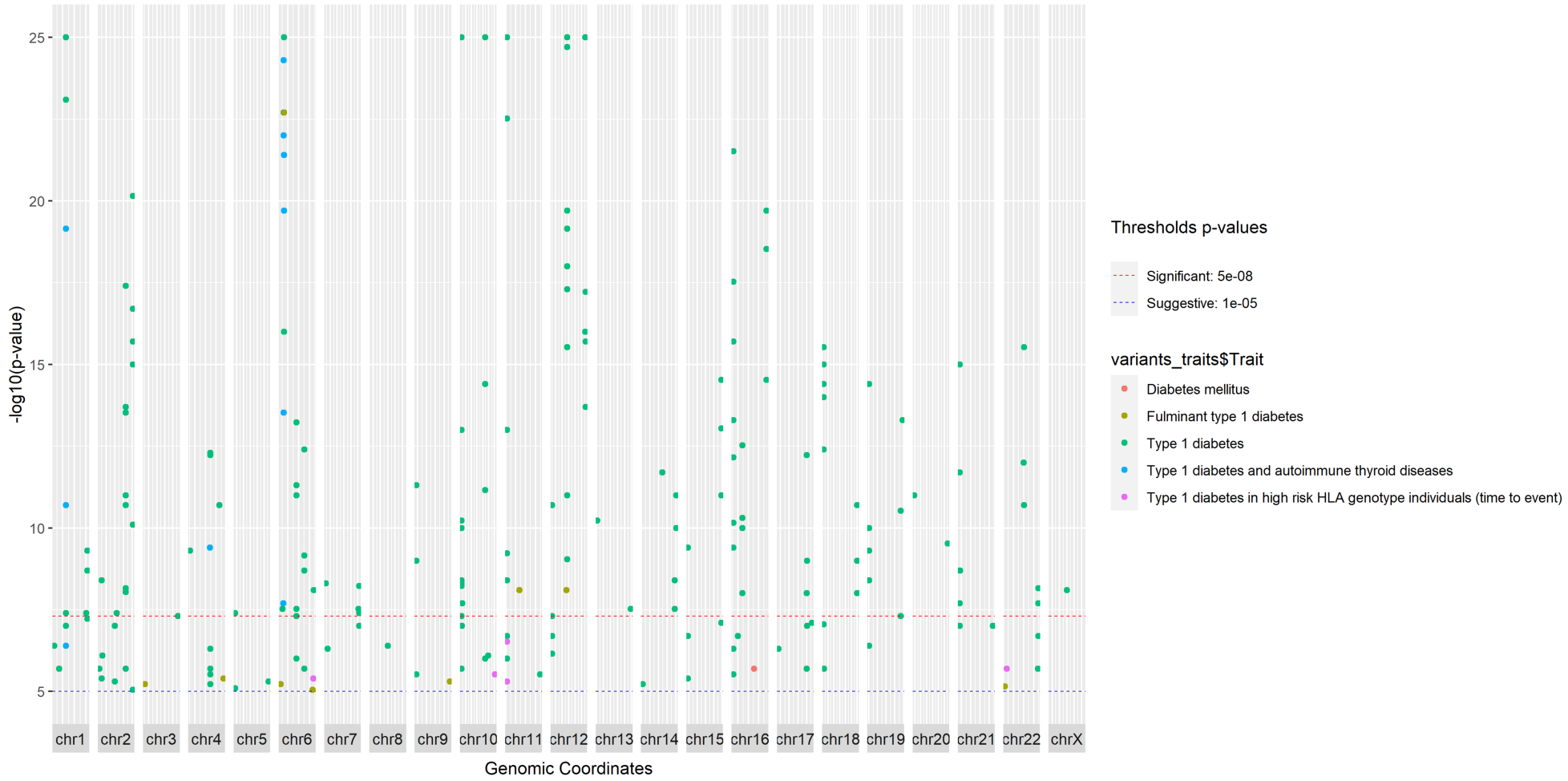


Figure B.1-2: Example of Manhattan plot obtained with VarGen for GWAS traits related to type 1 diabetes.

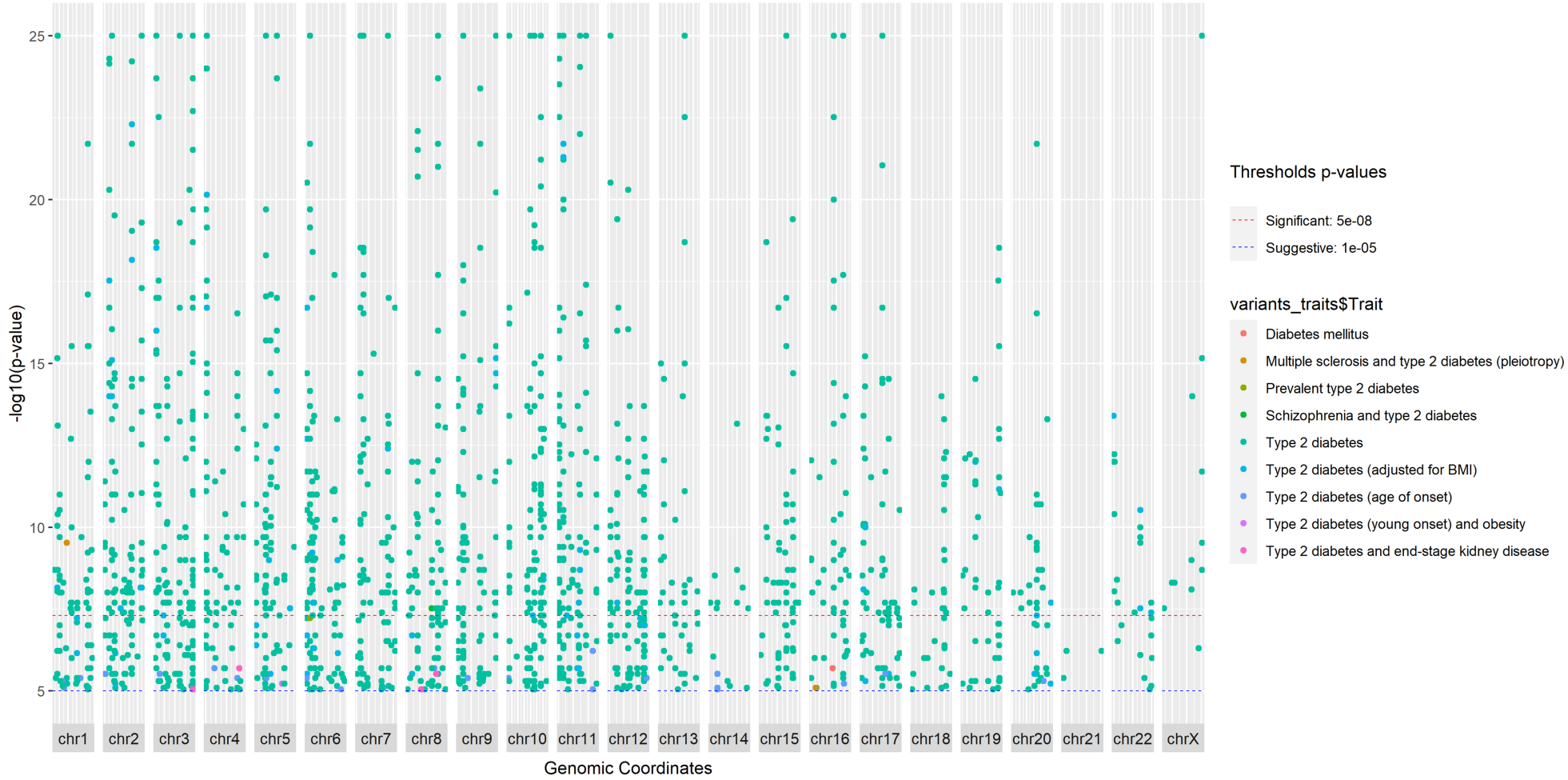


Figure B.1-3: Example of Manhattan plot obtained with VarGen, for GWAS traits related to type 2 diabetes.

Table B.1-5: List of terms obtained with *get_phenotype_terms* to test the VarPhen pipeline for the obesity use case. The key word used was: *obesity*.

Phenotype term
ABDOMINAL OBESITY-METABOLIC SYNDROME 3
Bilirubin levels in extreme obesity
Bitter taste perception 6-n-propylthiouracil in obesity with metabolic syndrome
Bitter taste perception phenylthiocarbamide in obesity with metabolic syndrome
DEVELOPMENTAL DELAY INTELLECTUAL DISABILITY OBESITY AND DYSMORPHISM
Hepatic lipid content in extreme obesity
Hyperinsulinemia in obesity
Monogenic Non-Syndromic Obesity
Morbid obesity
Morbid obesity and spermatogenic failure
Obesity
OBESITY (BMIQ14) SUSCEPTIBILITY TO
Obesity (early onset extreme)
OBESITY AGE AT ONSET OF
Obesity and osteoporosis
OBESITY ASSOCIATION WITH
Obesity autosomal dominant
OBESITY EARLY-ONSET SUSCEPTIBILITY TO
Obesity extreme
Obesity extreme SNP x SNP interaction
OBESITY HYPERPHAGIA AND DEVELOPMENTAL DELAY
Obesity in adult survivors of childhood cancer exposed to cranial radiation
Obesity in adult survivors of childhood cancer not exposed to cranial radiation
OBESITY LATE-ONSET
OBESITY MILD EARLY-ONSET
Obesity modifier of
OBESITY SEVERE AND TYPE II DIABETES
OBESITY VARIATION IN
Obesity without metabolic disease
Obesity-related traits
Retinal dystrophy and obesity
Salty taste perception in obesity with metabolic syndrome
Sour taste perception in obesity with metabolic syndrome
SPASTIC PARAPLEGIA INTELLECTUAL DISABILITY NYSTAGMUS AND OBESITY
Sweet taste perception in obesity with metabolic syndrome
Taste perception total score including 6-n-propylthiouracil in obesity with metabolic syndrome
Taste perception total score including phenylthiocarbamide in obesity with metabolic syndrome
Type 2 diabetes (young onset) and obesity
Umami taste perception in obesity with metabolic syndrome

B.2 VarGen Benchmarking – Alzheimer's disease

B.2.1 Methods

For VarGen, the following input was given to `vargen_pipeline:`

- The OMIM ID *104300.*
- The GTEx tissues corresponding to the brain, obtained using `select_gtex_tissues` with 'brain' as query.
- The GWAS terms related to Alzheimer's disease, obtained using the `list_gwas_traits` with the keyword 'alzheimer'. The GWAS data correspond to the GWAS Catalog version e96.
- The *fantom_corr* parameter was set to 0.20.

For VarPhen, the keyword 'alzheimer' was used to retrieve 70 phenotype terms having a link with Alzheimer's disease; they are listed in Table B.2-1. These phenotypes were given as input in the `get_variants_from_phenotypes` function.

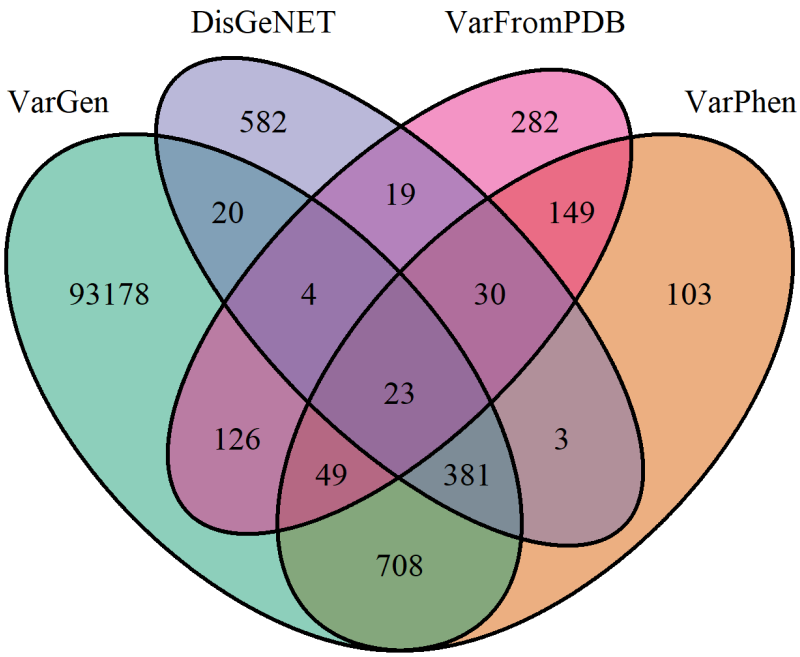
For DiGeNET, the disease identifier *C0002395*, corresponding to Alzheimer's disease, was given as input for the function `disease2variant`. As before, it was run one time with all the databases and one time with the curated ones.

As with obesity, for VarFromPDB, the author's guidelines were used with 'alzheimer' as the keyword to the pipeline. The same error as before appeared during the orphanet step and it was skipped as well.

B.2.2 Results

The results obtained with Alzheimer's disease were very similar to the one obtained with obesity. The overlap between the variant identifiers obtained with the different packages was represented as a Venn diagram (see Figure B.2-1A).

A.



B.

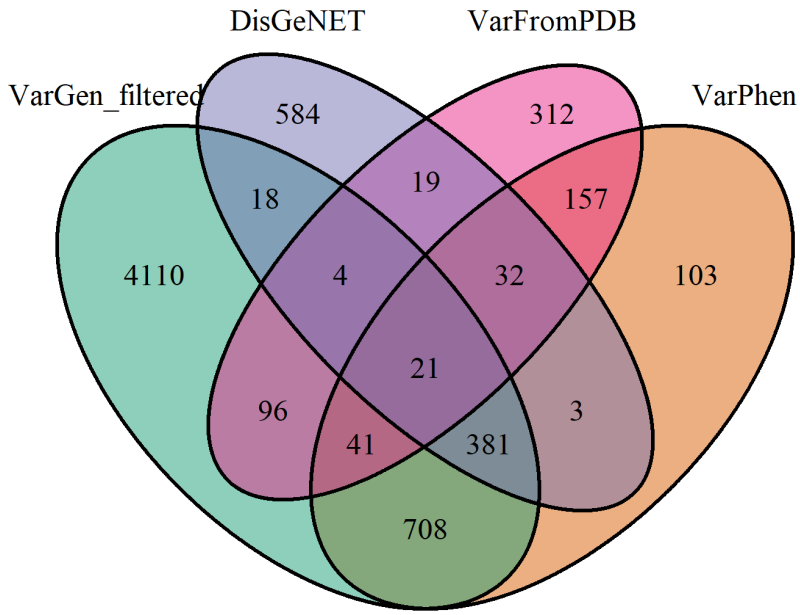


Figure B.2-1: Venn diagrams representing the variants found by the different pipelines: VarGen, DisGeNET, VarFromPDB and VarPhen. Alzheimer's (OMIM: 104300) was chosen as the use case. A. Venn diagram using the raw output for all the pipelines. B. Venn diagram using the filtered VarGen dataset, with the following strategy: all the variants from the GWAS Catalog and with clinical significance were kept, and the rest were filtered if their cadd Phred score was below 10.

The results are very similar to the one obtained with obesity. The highest number of shared variants, 708, are between VarGen and VarPhen.

DisGeNET shares 428 and 437 variants with VarGen and VarPhen respectively. VarFromPDB shares 202 and 251 variants with VarGen and VarPhen respectively. In contrast, VarFromPDB and DisGeNET only share 76 variants. As for obesity, the unique variants from DisGeNET are from literature mining and GwasDB, two resources not implemented in the other packages. Similarly, most of the 312 unique variants found with VarFromPDB do not have ‘alzheimer’ as keywords (Cutaneous photosensitivity, Hemochromatosis type 1, Variegate porphyria etc…). Many variants are found only by VarGen, since it reports variants affecting the genes related to a disease and not variants directly linked to the disease. Hence, some of the 93,178 variants uniquely found by VarGen are potentially false positives. This can be diminished by filtering variants based on their CADD Phred score, source, and clinical significance, while keeping most of the variants found in common with the other databases, as seen in Figure B.2-1B.

In summary, as for obesity, VarGen and VarPhen proved to be more sensitive than current alternatives. Better specificity can be achieved by selecting the VarPhen variants and/or filtering the VarGen results. Only 19 variants were found in common by DisGeNET and VarFromPDB, highlighting that neither VarGen nor VarPhen are missing relevant variants.

Table B.2-1: List of terms obtained with get_phenotype_terms() to test the VarPhen pipeline for the Alzheimer's disease use case. **The keyword used was 'alzheimer'.**

Accelerated cognitive decline after conversion of mild cognitive impairment to Alzheimer's disease (Alzhiemer's diagnosis trajectory interaction)
Alzheimer disease
ALZHEIMER DISEASE 18
Alzheimer disease 19
Alzheimer disease 2
Alzheimer disease 3 protection against due to APOE3-Christchurch
Alzheimer disease and age of onset
Alzheimer disease early-onset susceptibility to
Alzheimer disease familial 3 with spastic paraparesis
ALZHEIMER DISEASE FAMILIAL 3 WITH SPASTIC PARAPARESIS AND APRAXIA
ALZHEIMER DISEASE FAMILIAL 3 WITH UNUSUAL PLAQUES
ALZHEIMER DISEASE FAMILIAL WITH SPASTIC PARAPARESIS AND UNUSUAL PLAQUES
ALZHEIMER DISEASE LATE-ONSET SUSCEPTIBILITY TO
ALZHEIMER DISEASE PROTECTION AGAINST
Alzheimer disease susceptibility to
Alzheimer disease type 1
Alzheimer disease type 3
Alzheimer disease type 4
Alzheimer disease type 9
Alzheimer's disease
Alzheimer's disease (age of onset)
Alzheimer's disease (APOE e4 interaction)
Alzheimer's disease (cognitive decline)
Alzheimer's disease (late onset)
Alzheimer's disease (survival time)
Alzheimer's disease biomarkers
Alzheimer's disease in APOE e4+ carriers
Alzheimer's disease in hypertension
Alzheimer's disease in hypertension-negative individuals
Alzheimer's disease onset at age over 80
Alzheimer's disease onset between ages 58 and 79
Alzheimer's disease or family history of Alzheimer's disease
Alzheimer's disease or fasting glucose levels (pleiotropy)
Alzheimer's disease or fasting insulin levels (pleiotropy)
Alzheimer's disease or HDL levels (pleiotropy)
Alzheimer's disease or small vessel stroke
Alzheimer's disease progression score
Alzheimer's disease SNP x SNP interaction
Alzheimer's disease with language domain impairment
Alzheimer's disease with memory domain impairment
Alzheimer's disease with multiple cognitive domain impairments
Alzheimer's disease with no specific cognitive domain impairment
Alzheimer's disease with visuospatial domain impairment

Alzheimer’s disease and/or vascular dementia clinical subgroup VaD+
Alzheimer’s disease clinical subgroup AD+
Cerebrospinal AB1-42 levels in Alzheimer's disease dementia
Cerebrospinal fluid levels of Alzheimer's disease-related proteins
Cerebrospinal fluid p-tau levels in Alzheimer's disease dementia
Cerebrospinal fluid t-tau levels in Alzheimer's disease dementia
Dementia and core Alzheimer's disease neuropathologic changes
Early onset Alzheimer disease with behavioral disturbance
Early-onset Alzheimers disease
Early-onset autosomal dominant Alzheimer disease
Entorhinal cortical thickness (Alzheimer's disease interaction)
Entorhinal cortical volume (Alzheimer's disease interaction)
Family history of Alzheimer's disease
Hippocampal volume in Alzheimer's disease dementia
Late-onset Alzheimer's disease
Logical memory (delayed recall) in Alzheimer's disease dementia
Logical memory (immediate recall) in Alzheimer's disease dementia
Maternal history of Alzheimer's disease
Paternal history of Alzheimer's disease
Posterior cortical atrophy and Alzheimer's disease
Primary degenerative dementia of the Alzheimer type presenile onset
Psychosis and Alzheimer's disease
Psychosis in Alzheimer's disease
Response to cholinesterase inhibitors in Alzheimer's disease
Total ventricular volume (Alzheimer's disease interaction)
Voxel-wise structural brain imaging measurements in Alzheimer’s disease
Whole-brain volume (Alzheimer's disease interaction)

Appendix C

C.1 Plink format

- *.bim*: contains information about the variants, one line per variant, with the following 6 columns:

Field	Comments
Chromosome code	int or X/Y/XY/MT, 0 indicates unknown
Variant ID	eg: rsID
Position in morgans or centimorgans	0 used as a dummy value
Base-pair coordinates	1-based, maximum is 2^31 - 2
Allele 1	Corresponding to the clear bits in the .bed file
Allele 2	Corresponding to the set bits in the .bed file

- *.fam:* contains information about the participants, one line per sample, with the following 6 columns:

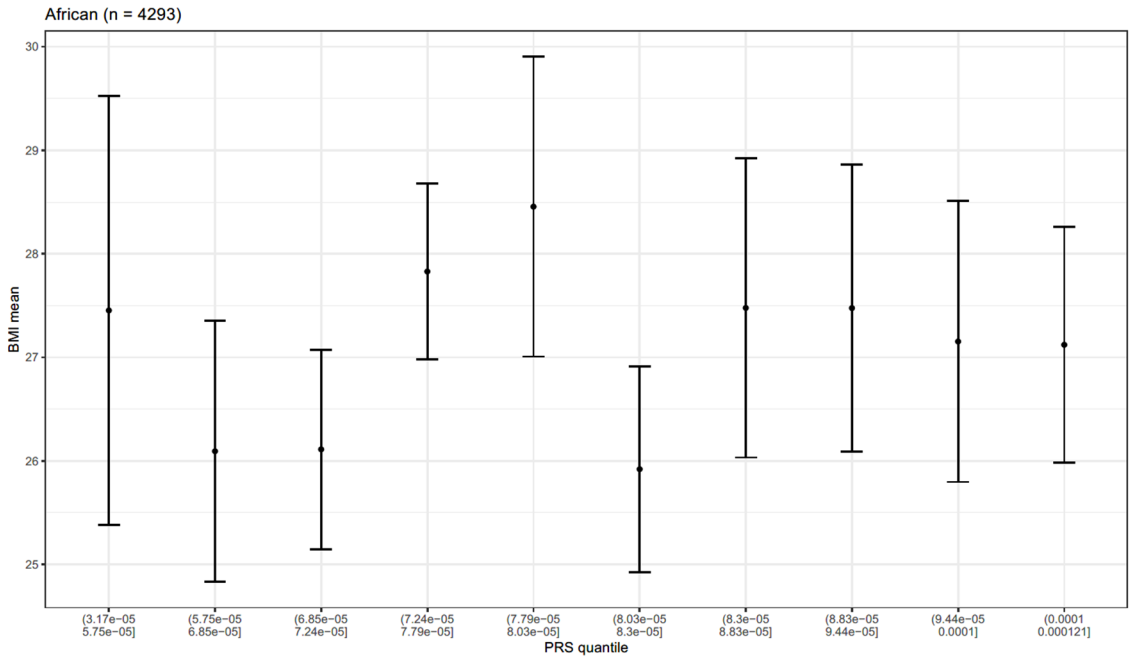
Field	Comments
Family ID	/
Within-family ID	Cannot be 0
Within-family ID or father	0 if not present in the dataset
Within-family ID or mother	0 if not present in the dataset
Sex code	1 = male, 2 = female, 0 = unknown
Phenotype value	1 = control, 2 = case,-9 / 0 / non-numeric = missing data

- *.ped:* contains the genotyping calls. (*.bed* for the binary version). It must be accompanied by a .bim and .fam files. It consists of two-bit genotype codes with the following meaning:

Genotype code	Meaning
00	Homozygous for the first allele described in the .bim file
01	Missing genotype
10	Heterozygous
11	Homozygous for the second allele in the .bim file

C.2 Limitations of the PRS model

A.



B.

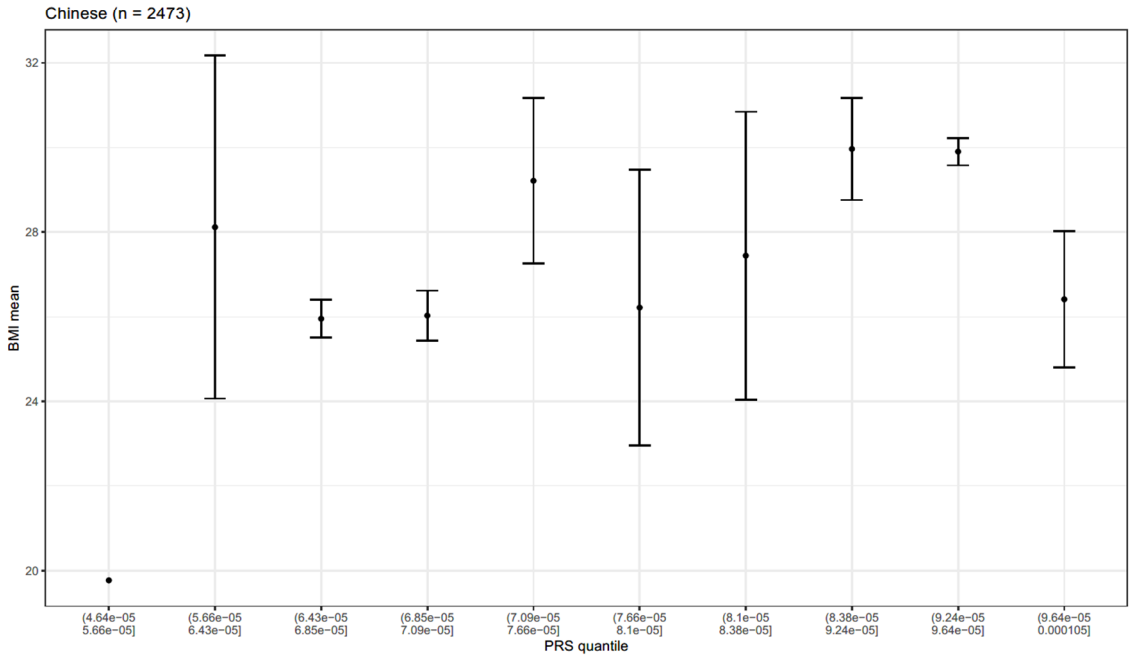


Figure C.2-1: BMI mean for each *backbone* PRS quantile, the bars correspond to the standard error, for individuals in the UK Biobank self-identifying as A. African and B. Chinese. There is no correlation between the PRS score and the BMI. This is because the base set of the model was obtained from individuals of European ancestry, which hinders the generalisability of the model across ethnicities with different Linkage Disequilibrium patterns.

Appendix D

D.1 miRNA analysis

Table D.1-1: Adapters used to generate the libraries for the miRNA sequencing

RNA 5’ Adapter (RA5), part:	5'-GTTCAGAGTTCTACAGTCCGACGATC-3'
RNA 3' Adapter (RA3), part:	5'-AGATCGGAAGAGCACACGTCT-3'

Table D.1-2: Number of reads (in millions) per sample before and after the filtering performed by *Novogene*.

	0h		24h		48h		72h	
Sample	**Raw**	**Filtered**	**Raw**	**Filtered**	**Raw**	**Filtered**	**Raw**	**Filtered**
A141	11.1	10.5	11.8	11.0	10.2	10.1	11.2	10.8
A145	12.0	11.4	12.7	11.9	11.9	11.1	11.0	10.1
A148	10.4	10.2	10.9	10.7	11.9	11.7	11.1	9.2
A151	11.0	9.5	13.8	13.0	12.6	12.1	12.2	10.6
A154	10.9	10.5	*missing*	*missing*	11.4	11.1	19.7	18.3
A157	12.0	11.4	*missing*	*missing*	11.0	10.8	10.0	8.4
A158	14.9	14.1	11.5	9.8	10.8	0.4	15.3	0.3
A163	*missing*	*missing*	11.8	9.6	11.9	11.5	*missing*	*missing*
A166	10.9	10.6	11.7	11.0	12.8	12.3	11.9	10.3
A194	15.4	14.8	10.7	10.0	17.4	16.4	14.8	13.8
A199	11.5	11.4	12.7	12.6	13.1	12.8	11.7	4.5
A36	10.3	9.8	11.4	11.2	*missing*	*missing*	*missing*	*missing*
A40	11.5	10.9	12.9	10.9	12.4	12.1	12.6	10.9
A41	11.8	0.3	11.8	1.6	14.8	0.9	11.4	8.6
A42	11.2	11.1	11.3	10.7	10.8	10.2	11.7	11.1
A43	11.5	5.7	13.6	13.4	13.2	12.2	14.1	11.1
A52	13.7	12.2	11.0	10.5	12.1	11.1	*missing*	*missing*
A54	13.1	11.5	*missing*	*missing*	12.9	11.7	12	10.6
A55	12.1	11.6	12.9	12	12.6	4.1	10.9	8.1
A56	11.7	0.2	12.5	1.1	15.2	1.2	11.3	7.3
A73	12.7	12.6	13.4	13.3	*missing*	*missing*	13.9	13.0
A74	11.3	1.1	*missing*	*missing*	*missing*	*missing*	11.2	10.0
A94	13.1	12.7	12.1	11.6	13.2	11.8	10.5	10.1
A95	16.0	9.3	11.6	9.3	14.1	13.1	10.9	0.4
AC1	10.5	0.6	*NA*	*NA*	11.3	10.1	*NA*	*NA*
AC2	10.5	7.7	*NA*	*NA*	11.5	10.8	*NA*	*NA*
AC3	10.4	5.9	*NA*	*NA*	12.1	6.7	*NA*	*NA*
AC4	12.8	12.2	*NA*	*NA*	12.5	11.9	*NA*	*NA*
AC5	10.3	0.6	*NA*	*NA*	10.4	7	*NA*	*NA*
AC6	10.9	7.6	*NA*	*NA*	13.2	12.9	*NA*	*NA*
AC7	10.7	9.3	*NA*	*NA*	13.4	13.3	*NA*	*NA*
AC8	10.6	9.9	*NA*	*NA*	11.5	11.4	*NA*	*NA*
AC9	12.4	11.8	*NA*	*NA*	11.9	11.6	*NA*	*NA*
AC10	13.3	11.9	*NA*	*NA*	12.4	12.1	*NA*	*NA*
AC11	12.3	11.4	*NA*	*NA*	13.8	13.7	*NA*	*NA*
AC12	10.8	9.7	*NA*	*NA*	11.3	10.0	*NA*	*NA*

Table D.1-2: List of targets predicted by miRDB for the novel miRNA 'hsa-novel-27395-mature', sequence: GAGTGTGCTAGAGTCCTCGAAG

Rank	Target Score	miRNA	Gene Symbol	Gene Description
1	99	hsa-novel-27395-mature	MBD5	methyl-CpG binding domain protein 5
2	97	hsa-novel-27395-mature	FUT9	fucosyltransferase 9
3	97	hsa-novel-27395-mature	IMPDH1	inosine monophosphate dehydrogenase 1
4	96	hsa-novel-27395-mature	CASZ1	castor zinc finger 1
5	95	hsa-novel-27395-mature	SLC35F1	solute carrier family 35 member F1
6	95	hsa-novel-27395-mature	HBP1	HMG-box transcription factor 1
7	95	hsa-novel-27395-mature	RHEBL1	RHEB like 1
8	95	hsa-novel-27395-mature	KIF13A	kinesin family member 13A
9	94	hsa-novel-27395-mature	CDYL	chromodomain Y like
10	94	hsa-novel-27395-mature	BAGE2	BAGE family member 2

Table D.1-3: List of targets predicted by miRDB for the novel miRNA 'hsa-novel-3327-mature', sequence: CGTGGTCTTCGGGGAGAGAG

Rank	Target Score	miRNA	Gene Symbol	Gene Description
1	97	hsa-novel-3327-mature	TSC1	TSC complex subunit 1
2	94	hsa-novel-3327-mature	FAM120C	family with sequence similarity 120C
3	92	hsa-novel-3327-mature	GCN1	GCN1, eIF2 alpha kinase activator homolog
4	90	hsa-novel-3327-mature	LUZP2	leucine zipper protein 2
5	89	hsa-novel-3327-mature	TXNRD3NB	thioredoxin reductase 3 neighbor
6	88	hsa-novel-3327-mature	CNOT2	CCR4-NOT transcription complex subunit 2
7	87	hsa-novel-3327-mature	TIMD4	T cell immunoglobulin and mucin domain containing 4
8	87	hsa-novel-3327-mature	PRKAR2A	protein kinase cAMP-dependent type II regulatory subunit alpha
9	87	hsa-novel-3327-mature	MTRNR2L3	MT-RNR2 like 3
10	87	hsa-novel-3327-mature	EEF1E1	eukaryotic translation elongation factor 1 epsilon 1

Table D.1-4: List of targets predicted by miRDB for the novel miRNA 'hsa-novel-45743-star', sequence: CACTGCGCTCCAGCCTGGGCAC

Rank	Target Score	miRNA Name	Gene Symbol	Gene Description
1	92	submission	PYROXD1	pyridine nucleotide-disulphide oxidoreductase domain 1
2	80	submission	SPEN	spen family transcriptional repressor
3	75	submission	STAG2	stromal antigen 2
4	74	submission	WDR26	WD repeat domain 26
5	73	submission	DNAH14	dynein axonemal heavy chain 14
6	73	submission	AMACR	alpha-methylacyl-CoA racemase
7	72	submission	RTKN	rhotekin
8	72	submission	GPAM	glycerol-3-phosphate acyltransferase, mitochondrial
9	71	submission	TMEM81	transmembrane protein 81
10	68	submission	ETS1	ETS proto-oncogene 1, transcription factor

Table D.1-5: List of targets predicted by miRDB for the novel miRNA 'hsa-novel-41718-mature', sequence: GTGTGTGCACCTGTGTCTGTC

Rank	Target Score	miRNA Name	Gene Symbol	Gene Description
1	100	hsa-novel-41718-mature	IGF2	insulin like growth factor 2
2	100	hsa-novel-41718-mature	SCARA3	scavenger receptor class A member 3
3	100	hsa-novel-41718-mature	ATP11A	ATPase phospholipid transporting 11A
4	100	hsa-novel-41718-mature	ADGRA1	adhesion G protein-coupled receptor A1
5	98	hsa-novel-41718-mature	FHL5	four and a half LIM domains 5
6	98	hsa-novel-41718-mature	TNRC6B	trinucleotide repeat containing 6B
7	98	hsa-novel-41718-mature	F7	coagulation factor VII
8	98	hsa-novel-41718-mature	FGFRL1	fibroblast growth factor receptor like 1
9	98	hsa-novel-41718-mature	KIF20A	kinesin family member 20A
10	97	hsa-novel-41718-mature	ASTN1	astrotactin 1

Table D.1-6: List of targets predicted by miRDB for the novel miRNA 'hsa-novel-44942-mature', sequence: TGGTCCAACGACAGGAGTAGG

Rank	Target Score	miRNA Name	Gene Symbol	Gene Description
1	98	hsa-novel-44942-mature	DCUN1D1	defective in cullin neddylation 1 domain containing 1
2	98	hsa-novel-44942-mature	PPP4R3A	protein phosphatase 4 regulatory subunit 3A
3	97	hsa-novel-44942-mature	UBA2	ubiquitin like modifier activating enzyme 2
4	97	hsa-novel-44942-mature	LHFPL6	LHFPL tetraspan subfamily member 6
5	97	hsa-novel-44942-mature	PRKX	protein kinase X-linked
6	97	hsa-novel-44942-mature	ARHGAP12	Rho GTPase activating protein 12
7	97	hsa-novel-44942-mature	RRAGC	Ras related GTP binding C
8	96	hsa-novel-44942-mature	MTX3	metaxin 3
9	96	hsa-novel-44942-mature	MAML1	mastermind like transcriptional coactivator 1
10	96	hsa-novel-44942-mature	RTL4	retrotransposon Gag like 4

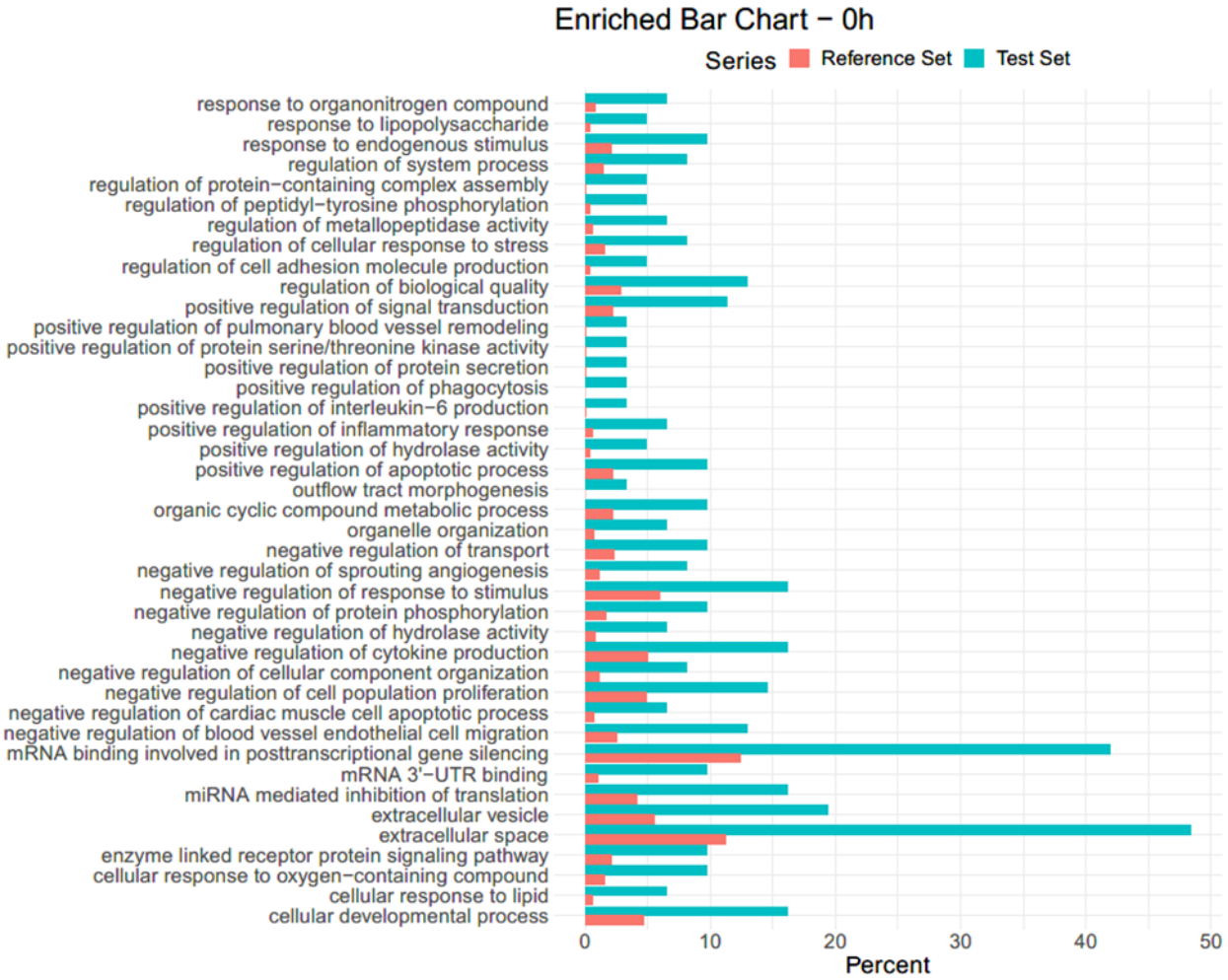


Figure D.1-1: Bar chart of the enriched GO terms at 0h. The test set corresponds to the differentially expressed miRNAs from the contrast 'pathological vs normal'; the reference set corresponds to the human miRNAs from miRBase, obtained with biomaRt.

Figure D.1-2: Bar chart of the enriched GO terms at 24h. The test set corresponds to the differentially expressed miRNAs from the contrast 'pathological vs normal'; the reference set corresponds to the human miRNAs from miRBase, obtained with biomaRt.

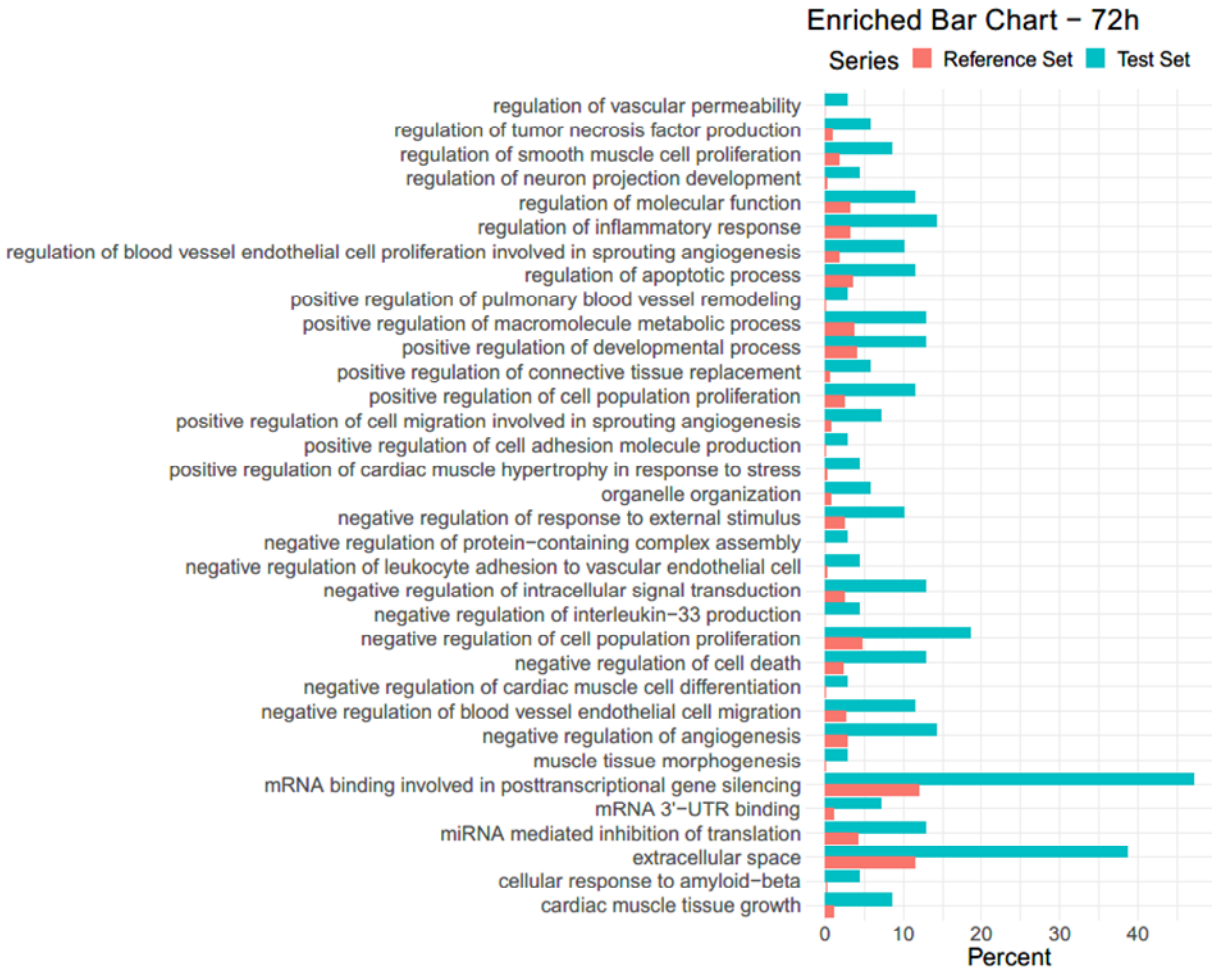


Figure D.1-3: Bar chart of the enriched GO terms at 72h. The test set corresponds to the differentially expressed miRNAs from the contrast 'pathological vs normal'; the reference set corresponds to the human miRNAs from miRBase, obtained with biomaRt.