THE UNIVERSITY *of* EDINBURGH

Development of the PathWAS methodology integrating transcriptomics and proteomics to predict pathway functionality for the association with complex genetic traits.

Sebastian May-Wilson

Doctorate of Philosophy in Precision Medicine

The University of Edinburgh, 2023

Declaration

I declare that this thesis is an original report of my research, has been written by me and has not been submitted for any previous degree. The experimental work is almost entirely my own work; the collaborative contributions have been indicated clearly and acknowledged. Due references have been provided on all supporting literatures and resources.

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

Signed

“There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.” – Donald Rumsfeld

“Had to be me, someone else might have gotten it wrong.” – Mordin Solus

“Do or do not. There is not try.” – Yoda

Acknowledgements

It may be lip service for some, but for me obtaining a PhD in biology and even within the broader field of genetics has been a goal of mine since I was a teenager. It is something I worked towards through my Undergraduate and Masters careers, but also something to which there was a significant road-block following my early University career. It is no exaggeration to say that for a period I believed I would never be able to obtain one.

I would therefore like to start my acknowledgements with my sincere and heartfelt gratitude to both Dr. Nicola Pirastu and Prof. Jim Wilson, both for giving me the chance to work with them and on such a topic that has proved so fascinating to me. Nicola has been unflinchingly kind and patient throughout my time at the University of Edinburgh and I could not ask for a better mentor in the sciences, both in terms of expertise and attitude. Jim has also been extremely encouraging and helpful, providing me with so much opportunity that I never expected to have, both in the lab and outside of it. They are both quite simply the sort of scientists I aspire to be.

I would also like to extend my thanks to the rest of the Wilson group who have been a source of support and assistance in numerous ways throughout my PhD. Paul Timmers for his repeated assistance with GWAS and polygenic scoring and putting up with me asking him the same question once a week. Peter Joshi for his insights into coding and being on time. Erin Macdonald-Dunlop and Linda Repetto who took time out of their own work to provide me with data which became fairly vital to my own project. Marisa Muckian for the commiserations. And also, the rest of the Wilson group who at various stages have been all been just great friends and provided an excellent environment to learn and work in.

I'd also like to extend my thanks to the many academics who I interacted with, however briefly, while working on PathWAS. Both for their ideas but also for their enthusiasm about my work, which was always a source of reassurance and encouragement.

Also, an additional special thanks towards Jeremy Mottram and Tom Van Agtmael, supervisors of mine from prior to my PhD whose references I believe helped actually helped get my foot in the door.

I'd also like to thank my parents who have been my staunchest supporters since day one. This should go without saying, but I love you both.

To all my friends who put up with me when I was just whining and assisted me when I genuinely needed it. Tim Mottram, Jamie Royle and Michael Burke for their realistic attitudes towards completing a PhD as well as commiserations from Sam Ireland and Dale Maxwell. Of course also, Ali Pollock, Lamios and the rest of The Crew™ for just everything. Plus, all of the Precision Medicine PhD Pilgrimage who were just a fantastic laugh for four years. Finally, those I lived over the past four years who, in many ways, made much of my life just much more bearable.

I don't really have the space to thank everyone that probably deserves acknowledgement, but suffice to say this feels like something which has been a long-time coming and feels like I had to be dragged over the finish line by everyone around me.

Thank you all so much!

Abstract

In the field of complex genetics, the use of genome wide association studies (GWAS) to discover genetic variants associated with complex traits and disease has been extremely successful. With sample sizes ranging into the millions, the focus has shifted from discovery of single nucleotide polymorphisms (SNPs) associated with complex traits, to elucidation of the biological mechanisms behind the associations. Traditional biological dogma is that variation in genetics influences phenotype by modulating expression of an associated gene through many possible mechanisms. The expression of the mRNA for this gene is, in theory, tied to expression of an equivalent protein, and it is proteins (and the metabolites they act upon) which are the drivers of phenotype. However, outside of Mendelian inheritance, the influence of individual associations and genes is much subtler and less perceptible. Part of this is likely due to the existence of broader biological pathways and networks in which multiple proteins and gene products act in concert, each contributing a smaller individual effect, which add up to one single large effect. While much work has been done in utilising GWAS results to search for enrichment of pathway terms in various databases, this methodology has the issue of potentially missing interactions due to small effect sizes or incorrect assignment of causality between genetic loci and genes.

In the field of precision medicine, determination of relationships between biological pathways and phenotypes has the potential benefit of allowing more targeted interventions and therapies. Discovery of relationships between pathways and specific diseases could allow prediction of the individuals most at risk as well as of variable response to medication targeting pathways, due to contrasting pathway activity. Therefore, the prospect of being able to determine which pathways are differentially regulated between individuals is an attractive one, both for determining causality behind genetic variation and for therapeutic benefit. Based on this, the aim of this project was the creation of a method, dubbed PathWAS, which could predict pathway functionality in the form of a polygenic risk score (PRS) and to then use these pathway polygenic scores to search for relationships between traits and the pathways.

The methodology of PathWAS involved the creation of PRS for prediction of gene expression, using expression quantitative trait loci (eQTLs). These PRS_{Gene} would provide an estimate for the activity of individual genes and could then be combined using pathway databases into a broader PRS for different pathways. A vital component of the project was an

estimate of pathway function, for which I used proteomics measurements for proteins at the ends of the pathways as a proxy for function, with the assumption that the cumulative effect of the pathway would directly influence expression levels of downstream genes and proteins. Using these measurements, I conduct a multivariable Mendelian randomisation of each gene from within the pathway against SNPs from a GWAS of the protein. This provides a weight for each gene, allowing us to weight each PRS_{Gene} by its effect on "the pathway". These combined and weighted scores were then used in an exploratory PheWAS analysis in the UK Biobank, searching for pathway-phenotype associations

From two different proteomics data sets I obtained a total of ~3,200 pathway-protein models (with some overlapping between the two sets of proteomics). These were then each individually tested against 60 different phenotypes. Following subsequent sensitivity analyses, this resulted in ~2000 significant phenotype-pathway associations, many of which were supported by existing literature.

Overall, this provided a proof of concept for the PathWAS methodology. Other users can apply the method to their own GWAS and pathway data through use of the developed R package, which is made available through GitHub, with further subsequent expansion and refinement of the method still possible. This methodology has the potential to complement GWAS in discovering pathway-phenotype relationships beyond existing enrichment techniques. It also allows for a novel way of expanding the usage of the data which has been generated in the field of genetics, much of which remains under-utilised.

Lay Summary

The central dogma of molecular biology is that the DNA of an individual acts as a blueprint for the creation of mRNA, in the form of genes within the DNA. These mRNA molecules in turn act as instructions for building separate proteins and it is proteins which, theoretically, cause individuals to develop certain traits. This includes variable proteins which cause differing eye colours or mutations within another protein leading to disorders like Huntington's disease. However, in so-called complex traits such as height, cancer and type-2 diabetes, there is no singular protein or molecule acting independently. Instead, these traits are the result of many overlapping effects from multiple different genes and proteins acting cooperatively through biological pathways (in some cases there might be one pathway directly responsible for a trait, whereas in cases like height it is likely influenced by dozens of different overlapping and interconnected pathways).

Variable activity of these pathways could potentially explain an individual's predisposition towards various traits but could also explain the range of responses people with the same disease can have to the same treatment. If one person metabolises a drug faster or more efficiently, it is possible they would see far more or fewer benefits from it than another individual with a subtly different genetic make-up. This is an aspect of precision medicine which potentially has far reaching consequences, even allowing better understanding in drug discovery and clinical trials.

Unfortunately, while it is possible to measure levels of mRNA and proteins within individuals, there is no way of measuring overall pathway activity, primarily because a pathway is a series of biochemical reactions occurring in a known sequence, rather than something with a quantifiable input and output. However, in this study we have attempted to develop a method which circumvents this and estimates and then predicts pathway functionality. With this prediction, we then search for associations between pathways and various complex traits, providing a novel method for the discovery of relationships between traits and biological networks.

To estimate a pathway's overall function, we use a measurement of a protein molecule which is at the end of a given pathway. In doing this we make the assumption that a pathway's activity is a measurement of the cumulative activity of the multiple genes involved within it, and that the molecule measured at the end will therefore have a direct relationship with

overall activity. We also use measurements of the mRNA molecules for the genes within the pathway and perform a statistical analysis which provides an effect for each gene on the selected end-point. In doing so, we weight the effect of each gene on our proxy for pathway function, while also accounting for the effect of every other gene in the pathway.

I combine these weighted effect sizes with the genetic information of individuals into a single score. This score will provide an estimate of the overall pathway functionality in every individual, predicted by genetic information. This newly created score acts as a new trait for examination which can be analysed with complex traits, to characterise any relationships between individual pathways and traits of public health importance or interest.

This method which we dubbed “PathWAS” involves numerous sources of online data, amalgamated into one single score per individual in the population-based cohort UK Biobank. The scores were tested against 60 different traits (including height, BMI, lifespan, bone mineral density, lung function traits and more). From this I discovered several thousand significant associations, refined to approximately 2000 following a sensitivity analysis. Many of the relationships validate those which have either been previously reported in the literature or make logical sense, suggesting that our method works as intended in finding true relationships between biological pathways and complex traits. Future research could use the PathWAS methodology to search for novel associations with additional traits and diseases and also use expanded sources of molecular data to estimate pathway functionality.

Contents

Output and contributions

Output

The methodology described in this thesis has been collated into a publicly available R package available for download and installation from GitHub at the following location: https://github.com/Sabor117/PathWAS

Contributions

Andy Bretherick, David Clark, Peter Joshi and Paul Timmers were responsible for the creation of the association testing pipeline utilised in **Chapters 5** and **6** when conducting elements of the PheWAS analyses (**Section 5.2.4**). Paul Timmers also contributed heavily to the creation of the PRS-creation pipeline used in creating PRS-CS and PRsice-2 polygenic scores (**Section 6.1.4**).

Arianna Landini created a UK Biobank genotype reference panel of 10,000 white British individuals, this was used at many points in the PRS-creation and is described in **Section 3.5.3.1**.

Erin Macdonald-Dunlop generated the GWAS of the ORCADES and Vis proteomics used in **Chapter 5** (described in **Section 5.2.1**). These protein measurements specifically from the ORCADES cohort were also used in **Chapter 6** for the prediction of the SCALLOP PathWAS results (**Section 6.1.3**). She also generated the meta-analysis of the SCALLOP proteomics GWAS for the Olink CVD2 and CVD3 panels used in **Chapter 6** (**Section 6.1.3**).

Linda Repetto generated the meta-analysis of the SCALLOP proteomics GWAS for the Olink NEX and NEU panels used in **Chapter 6** (**Section 6.1.3**).

Xue Li conducted the TreeWAS of the PathWAS version 1 scores in **Chapter 5** (Section 5.3.4).

Chapter 1

Introduction

1.1 Genetics and precision medicine background

The study of the genetic basis of complex traits, those which do not follow the patterns of Mendelian inheritance, is a constantly evolving discipline. Unravelling both the underlying genetic cause and then biological mechanism of these traits has been challenging, with every new discovery unveiling additional levels of complexity. Early twin-studies in genetics suggested that many complex traits were driven in a large degree by inherited genes, for example height where genetic variation was consistently predicted to be responsible for between 70-90% of the variation of the phenotype[1]. These twin-studies were previously considered the gold-standard for determining heritability of a trait[2,3] (the degree to which the phenotype is influenced by inherited genes). However, the results of these studies are now considered to be more contentious with the possibility that twin-studies inflate the heritability estimates of the trait[4]. Subsequent genome-wide association studies (GWAS) revealed that many of the discovered variants discovered could only account for a much smaller proportion of the trait's heritability[5,6], resulting in a discrepancy between the predicted heritability of a trait and that which can be discovered within genetic variation. This problem, known as missing heritability[7,8], is just one of the primary reasons that meant that on an individual level the possibility of predicting phenotypes, such as height and diseases like cancer or cardiovascular disease, from genes, is largely still reserved for the pages of science fiction[9,10]. This is in stark contrast to conditions which show Mendelian inheritance patterns, those which are caused by single variants or mutations with much larger effect sizes (with traditional examples such as phenylketonuria[11] and Huntington's disease[12]).

Being able to predict these traits is a desirable concept in order to help mitigate risk in individuals who are most at danger of developing complex diseases. This is an inherent goal of the field of precision medicine[13], with the broad aim of developing targeted therapies and treatments for individuals based on genetic variation and this starts with being able to accurately predict differences between people. There are numerous obstacles to creating accurate methods of prediction of complex traits from genetic variation, including but not limited to the tiny contributions of effect from individual genetic variants towards any

particular trait of interest[14,15] and the complexity in defining how these variants actually exercise their effect on the phenotype.

To aid with treatment and management of complex disease it is important to understand both the genetic basis of these phenotypes and the underlying biological mechanisms by which these changes in DNA action their effect on phenotype. This is one reason why in recent years there has been a greater focus on not just improving the power and availability of genetic data (with introduction of online repositories of genetic data[16,17]), but also on incorporating additional levels of biological knowledge such as transcriptomics and proteomics. This objective of moving beyond just the discovery of genetic variants associated with phenotypes and instead towards understanding causal variants and their biological mechanisms is a primary aspect of the "post-GWAS era"[18,19].

Incorporation of these omics datasets could facilitate analysis of the functional effects of changes in genetic variation, which in turn allows the possibility of discovering new druggable or actionable targets and biomarkers for disease[20–22]. This is because these various omics datasets are closer to the core functional units in biology, particularly when considering traditional biological dogma which states DNA is used to create RNA which is used to create proteins. Therefore, these molecules arguably have a closer relationship with the phenotype than the genotype and so expanding upon the molecular knowledge of genetic function (with the incorporation of transcriptome, proteome, and metabolome information) increases biological understanding, for both normal and pathological phenotypes.

Study of the molecular make-up of individuals also has the added benefit in that incorporating additional levels of biological understanding theoretically improves the power of discovery for associations through techniques such as transcriptome-wide association studies (TWAS)[23,24]. So, incorporation of multi-omics has the capacity to not only improve biological understanding but also genetic knowledge.

With the aim of improving the understanding of the genetic basis of complex phenotypes and their biological mechanisms there have been great strides in improving the quantity and quality of genetic information available and introduction of large multi-omics datasets. Alongside many new statistical techniques, this data has led to many exciting discoveries[25–27]. An example of how this is already affecting the field of precision medicine is the pilot study for incorporation of predictive scores in cardiovascular disease (CVD) being trialed by the NHS[28]. The significant progress made in recent years in predicting CVD[29,30] certainly

contributes to this new study. However, many of these data sources may even be underutilized. So, being able to further integrate these data sources, incorporating omics with disease prediction, would potentially provide another step towards realizing the goal of precision medicine.

1.2 Advances and issues in genetics and GWAS

The first GWAS was published in 2002 on myocardial infarction by Ozaki *et al.*, relying on an array of fewer than gene-based 100,000 single nucleotide polymorphisms (SNPs)[31]. In this study 65,671 SNPs were successfully genotyped from a database of fewer than 200,000 variants[32] in 94 individuals with 658 controls. Using a P-value threshold of $P < 0.01$ they identified a candidate locus on chromosome 6 associated with susceptibility to myocardial infarction and indicated that variants in the gene for lymphotoxin alpha (LTA) are associated with risk. This also implicates LTA in the pathogenesis of the disorder, thus identifying a previously unrealised possible drug target. This study may have indicated the possible power of GWAS, even despite the small sample size by modern standards.

In the following two decades the quality and quantity of GWAS have increased exponentially[33,34]. In 2008 the GWAS catalog was founded as an online repository for summary statistics from any GWAS experiment[16]. In the 10 years following its advent, the number of GWAS statistics in the catalog increased 10-fold with a similar increase in the sample size of the studies[34]. Modern GWAS examine many millions of SNPs across hundreds of thousands or even millions of individuals, to search for variants associated with the selected complex traits[5].

This massive increase in sample size has led to a vast increase in the power and ability to discover genetic variants associated with complex traits. However, despite the increase in GWAS power, it has still not been possible to discern all variants responsible for contributing to a trait except in rare cases with studies in the many millions of samples. This issue is known as the problem of missing heritability, where all the known genetic variation associated for a given trait does not account for all of the expected variation driven genetic causes[7,8]. Twin-studies, arguably the gold-standard for defining the heritability of a given trait, have demonstrated that approximately 80% of the variance in height is caused by genetic variation[1], even GWAS using sample sizes in the hundreds of thousands did not

reveal every variant associated with the trait, only explaining ~50% of the SNP heritability for height[6]. It was not until 2022 that Yengo *et al.* was able to elucidate much of the remaining missing variance for height in a GWAS with 5.4 million individuals, discovering 80-90% of the SNP heritability[5]. Even with the incorporation of tens of thousands of samples from whole-genome sequencing (WGS), this still only accounted for 40% of the variance of height predicted from twin studies. Similarly, other very large GWAS for intelligence[35] and schizophrenia[36] have also failed to uncover all of the predicted variance for the phenotypes. This means that even through utilising GWAS with sample sizes in the millions and incorporating additional sources of data such as WGS, it is possible that it would not be possible to account for all the genetic variance of a trait and so predicting the phenotype based on genetic information would like necessitate the usage of mixed models.

There are several possible reasons for this discrepancy. One such possible reason is the effect of rare variants which, due to their very low frequency within larger populations, are difficult to identify[37]. Alternatively, it is possible that some elements of the genetic variance of a trait will be driven by larger mutations and changes in the genome, such as copy number variation (CNVs)[38,39]. In many GWAS though, a major factor is variants with extremely small effect sizes, to the extent that even the current sample sizes for most studies are insufficient to identify them.

Alongside the enormous increase in data, this issue of being unable to identify all causal variants for a trait has led to the development of numerous tools to try and improve the power of discovery. This has been coupled with developments in genetic theory such as the omnigenic model, which argues for particularly polygenic traits (such as height and intelligence) the phenotype will be driven not just by dozens or hundreds of associations and genes, but rather by thousands with a multitude of cell-specific effects[40,41].

1.3 Omics as biological functional units

The central dogma of molecular biology is based on the theory that proteins are translated from RNA and RNA is produced from DNA, and that it is the proteins that are, in large part, responsible for phenotype. At the time this concept was introduced, the full complexity of the process was not appreciated, for example, it does not take account of the many non-coding RNAs which are transcribed but not translated, and yet have significant biological functions.

Furthermore, the roles of alternative splicing and epigenetics are now more fully appreciated. Nevertheless, the central dogma remains a fundamental principle in understanding molecular biological mechanisms.

Acceptance of the central dogma then also implies that each subsequent step of the process is more closely associated with the phenotypes it drives. While genes encode the blueprint for the make-up of an organism and its cells, there are also many more layers of regulation and processing between them and any final phenotype.

RNA, measured by transcriptomics, is closer to phenotype as this represents the cell's or organism's usage of the blueprints provided. So, in theory when a particular gene is transcribed at a greater or lower level and this is seen in association with a phenotype, then it is a better indication of the biology at work rather than knowing the changes in one's genetic blueprint.

Similarly, while RNA reflects the initial reading of the genomic blueprint, analysis of the proteome provides the subsequent reading that will impact phenotype. This makes proteomics once-again a step closer to biological function and so studying protein levels in relation to phenotype is also theoretically closer to understanding the biology.

There can be drawbacks to studying the full genome, transcriptome or proteome. Specifically, with proteomics some present major limitations are the associated cost and that it is not currently feasible to analyse an individual's entire proteome.

Traditionally proteomics studies were conducted using mass spectrometry (MS)[42]. MS-based technologies have the advantage of providing an almost complete picture of an individual's proteome however they come with a number of limitations. There are issues such as the reproducibility of the results but also a major issue is that below a certain level of abundance, MS will be unable to detect a protein whatsoever[43,44]. For my project the primary issue is that at a population-level, MS is both a costly and labour-intensive practice and so proteomics GWAS have only been conducted using it in relatively small cohorts[45,46].

More recent technologies rely on studying a set panel of previously defined "interesting" proteins. Companies such as Olink and SomaLogic have produced competing technologies which are now widely used, with numerous studies analysing proteins using their assays[47–49]. Briefly, Olink technologies uses antibodies bound with DNA molecules which bind together and allow quantification via RT-PCR, named as an proximity-extension assay (PEA)[50], while

the SomaScan platform uses single-stranded DNA aptamers which bind the proteins directly and are then measured on a specific DNA-binding chip[51,52].

Recently there have been further developments in proteomics with Quantum SI offering the possibility of analysing entire proteomes with semi-conductors, however this technology is still in its infancy and so has no large-scale results produced (although it promises to be highly scalable)[53].

Beyond transcriptomics and proteomics, there are additional layers of biology to consider as well, such as the metabolome. Like many proteins, the small molecules examined via metabolomics are those which might directly mediate phenotype, with their levels tightly regulated by other proteins.

Therefore, if it were possible to integrate a number of these technologies, to combine genomics and proteomics in order to predict protein levels from an individual's genetic makeup, then it would be one step closer to fully understanding and predicting complex traits as a whole.

1.4 Strategies for discovering biological mechanisms behind genetic variation discovered by GWAS

As a consequence of the increase in power of GWAS there are now many thousands of loci associated with different complex traits, containing tens of thousands of SNPs. However, significant GWAS associations only reveal SNPs which are associated with a trait and not how, and so linking the SNPs to biological function is a current objective in the field of genetics[54,55]. This is important as understanding the mechanism by which changes in the genome affect phenotype is vital in aiding in the search for drug targets and targeted treatment plans for disease.

Numerous approaches have been described in the literature which show biological mechanisms linked to SNPs and loci, however, there are many challenges which have meant that work in this field has lagged behind discovery of significant loci[54]. The two primary challenges for determining biological mechanism of loci identified from a GWAS hit are determining through what mechanism the SNP is affecting phenotype and secondly determining causality within the locus itself.

Assessing the mechanism by which SNPs affect phenotype is challenging as the vast majority of GWAS hits exist in non-coding regions of the genome[18,54]. There are many possibilities for biological mechanism but generally it is assumed that SNPs will affect phenotype by modulating activity of an individual gene or gene-product. SNPs within coding regions may alter protein function moderately or severely, with some SNPs such as missense or nonsense variants being highly deleterious[56,57]. However, these are the minority of SNPs discovered from GWAS. Instead, the mechanisms by which SNPs will affect expression can be as an enhancer or repressor of transcription[54], by modulating splicing[58], altering the local epigenetic and methylation profile[59,60], there are also SNPs which may directly affect protein levels without affecting levels of mRNA expression[18]. Specific examples can include SNPs within enhancers which modulate transcription-factor binding affinities, with these enhancers attached through long-range chromatin folding to a transcription start-site. This can be seen with variable levels of ATF6 binding at different alleles of rs6877329, resulting in differential expression ELL2 and thus variable risk for myeloma[61]. Determining the mechanism by which a SNP alters expression is, however a labour-intensive process and there is also presently no gold-standard for methodology. As well as this, even defining which gene the SNP is affecting is often not entirely straightforward, target genes are usually defined as those closest to the locus in question, however, this is not always accurate[62].

The issue of defining the causal SNP is that within a significant locus there will often be several significant SNPs. These SNPs will likely only be significant because they were within close linkage disequilibrium (LD) with the causal SNP. Unfortunately, significance itself is not enough to define causality, so just taking the sentinel SNP (the SNP with the lowest P-value) is usually not enough to define the causal SNP, and in fact the most significant SNP in many cases is in fact not the causal variant[63].

In order to try and discover both the specific biological mechanism and the causal SNP, various fine-mapping approaches are generally used[63]. Fine-mapping, statistical and otherwise, is the broad name for the approach in attempting to assign causality to specific SNPs from within a genomic locus discovered by GWAS[63,64]. These methodologies assume that of the multiple separate SNPs in each locus discovered to be associated with different complex traits, many of the signals will be caused by the close linkage disequilibrium between variants and the causal SNP. As such there will only be one or a few SNPs actually responsible for the effect on phenotype, while the other signals are discovered due to inheritance of similar patterns of alleles, and fine-mapping attempts to pinpoint which are the

causal variants. These approaches normally incorporate many sources of data to try to demonstrate causality for a candidate SNP, the gene it is acting on and the method through which it causes its effect. Fine-mapping can include many forms of analyses, such as examining open chromatin[65], ATAC-seq[66], histone modification analysis[67], examining whether SNPs are found in known promoter regions, transcription factor colocalization and whether the SNPs alter the sequence of a transcription-factor binding site. Additionally, colocalization of SNPs with those from online sources of expression quantitative trait loci (eQTLs) or protein quantitative trait loci (pQTLs) could be used to prioritise genes when searching for associations with any gene expression[68]. Colocalization of signals are often conducted with the Coloc R package[69], which attempts to assign causality by examining the overlap of two sets of GWAS phenotypes (I.e. a given trait GWaS and the eQTLs from a GWAS of gene expression). Coloc specifically uses a Bayesian methodology to provide putative causal variants from the overlap of the two sets of significant SNPs. Other possibilities for the inclusion in fine-mapping include using scores like Genomic Evolutionary Rate Profiling (GERP) to determine whether a region is evolutionarily conserved, as these regions are often more likely to be functional[70]. On top of this, more costly lab-based techniques such as high-resolution promoter capture (3C and Hi-C) can also be used to link loci with gene promoters[71] while methods such as the massively parallel reporter assay (MPRA) can be used to try and directly associate loci with enhancer and promoter regions[72,73].

As well as manually examining these various forms of data, there are also numerous tools and workflows, such as eCAVIAR, a probabilistic method which incorporate eQTLs along with GWAS summary statistics and LD structure to try and determine which SNPs are most likely to be causal while also identifying causal genes[74].

This is by no means an exhaustive list of options which has the resulting issue that while one can take many avenues for the of fine-mapping of genetic loci to determine causality, there is no widely-used and accepted standard work-flow for producing results. Thus, this makes finding definitive conclusions is often extremely challenging. The challenge of understanding the biological mechanisms behind a given locus stems from the fact that a single genome-wide association study (GWAS) often reveals a large number of significant loci, each containing numerous significant SNPs, many of which may be clustered in regions with multiple potential causal genes. Currently, there are no high-throughput solutions available to

effectively tackle this complex task, and so *in-silico* techniques which can aid in understanding biological function are highly sought after.

1.5 Incorporating transcriptomics to improve power in GWAS

Given the large focus on the discovery of significant associations with complex traits and the subsequent analysis of the biological mechanism of these loci, techniques which aid in the discovery of both can be extremely useful.

TWAS incorporates data from transcriptomics, in the form of eQTLs, with GWAS in order to aid with pinpointing causal genes and loci[23]. eQTLs are usually derived from RNA sequencing (RNA-seq) data, searching for variants which are associated with altered expression of various mRNAs (and also non-coding RNAs) within the genome[75]. Like GWAS, RNA-seq has become a relatively inexpensive technology and so there are already several large eQTL datasets available online such as GTEx (a dataset of RNA-seq performed in multiple tissues)[76,77], eQTLgen (RNA-seq using whole blood samples with a very large sample size)[78], BLUEPRINT (a large study of the epigenome, including transcriptomics, in numerous haemopoietic cell lines)[79] and MESA (a transcriptomics study examined in the context of DNA methylation)[80].

eQTLs discovered by transcriptomics can be incorporated into GWAS analyses during TWAS analyses, combining the search for significant loci associated with the phenotype of interest, with searching for the relevant genes affected by these loci[23]. This can be done through penalized linear regression of the eQTLs with the GWAS loci or through using mendelian randomisation (MR)[81,82].

In theory TWAS offers significant benefits alongside standard GWAS. By combining genetic information with expression data, TWAS incorporates data which is potentially more closely-associated with the phenotype, thus making elucidating function of the loci much easier. Alongside the increase in biological understanding there is also an increase in the power of discovery because of a reduction in the level of multiple testing required. Traditionally, for a locus to be deemed significant in GWAS it has to pass the Bonferoni-corrected threshold of "genome-wide significance" ($P < 5\times10^{-8}$) while for TWAS, as the number of statistical tests made is only the number of genes examined, in humans this means that the number of tests does not exceed 20,000-30,000 meaning a reduced P-value threshold of ~ 2.5×10^{-6}.

Taken together this would suggest that using TWAS allows for not only greater power of discovery of significant loci associated with a phenotype, but also provides the potential for greater depth of biological understanding. Unfortunately while this remains true in theory, in practice TWAS has been less successful at providing this link between GWAS and genes, and has not yet fulfilled its potential[23].

1.5.1 Issues with transcriptomics

Despite the benefits of TWAS there are however still significant obstacles, many of which can be traced to the RNA-seq itself used to generate the eQTL datasets.

A primary issue of utilising data generated by transcriptomics for my purposes is the aforementioned closeness with phenotype. As previously stated there is a broad acceptance that proteins and metabolites are closer to phenotype than genetics and transcriptomics as it is protein abundance which will, in many cases, drive cellular activity and function[83]. Thus, transcriptomics can be less effective for analysing phenotype than the more expensive proteomics or metabolomics. Furthermore, there is substantial evidence demonstrating that eQTLs for a given gene are often poorly predictive for the levels of the associated protein product[83,84].

It is perhaps surprising that variation in RNA levels does not always lead to a corresponding variation in protein levels. eQTLs do correlate with protein levels, but only weakly so[85]. This is indicative of the many tight layers of biological control in place within the cell. As well as varying levels of transcription of mRNA there is also the effects of differential splicing, variable RNA stability, varying polysome attachment affinities to the different RNAs, protein synthesis, stability and degradation as well as the effects of post-translational modifications (PTMs)[83,85,86]. All of which may have an effect on the level of mRNA expression varying significantly from protein expression. It is also possible that the issue may lay with the technology used for measuring protein abundance, for example due to measuring specific isoforms of the protein incorrectly[87].

Beyond these intracellular regulatory aspects there are further mechanisms such as export of protein products from and between various cell and tissue types. It is possible that specific cells in the body will produce high levels of mRNA for a given protein, purely for the purpose of synthesising and then exporting this protein (such as hormones), resulting in

poorly correlated levels of mRNA and protein levels within these cells[88]. This purely mechanical reasoning alone may affect protein levels to a degree such that it is not possible to get an accurate picture of how much effect a change in transcription will have on the final protein abundance.

It is also the case that recently there has been a greater focus on the importance of tissue or cell-type specific effects on phenotypes. This means that some studies may not be appropriate for examining specific phenotypes or may miss important subtleties and information. Fortunately, this is something which has been increasingly recognized and as such larger single-cell transcriptomics studies are already being conducted which may revolutionize many of the concerns raised here. A better understanding of the transcriptome on a cell and tissue-specific scale, may in turn result in better understanding of the high variability between mRNA and protein levels, and thus may make the associations between transcriptome and phenotype clearer.

1.6 Biological pathways and their function

It is commonly accepted that the major way which genes and proteins mediate their function is through large and interconnected biological networks[89]. There are dozens of possible protein-protein interactions which can occur within a pathway, enhancing or inhibiting a downstream effect. These biochemical interactions range from cleavage of proteins to activation of them via attachment of required molecules, binding of constituent parts together into an active complex, catalysis[90]. The effect of proteins on a pathway may also be due to influencing transcriptional control of other proteins which propagate the downstream effect[91]. Many of these pathways will interconnect and exhibit forms of control over each other and it is the accumulated effect of all of these which will result in phenotype.

It has been the focus of many avenues of research therefore to determine the pathways which act within and between cells and to refine understanding of interactions between the constituent elements. This has led to the discovery of many pathways which mediate many phenotypes. Dysregulation and disruption of these pathways through component genes can thus radically alter phenotypes. Well-studied examples include the CDK driven cell-cycle pathway which mediates cell division and mitosis[92,93] and many of the proteins within the pathway have oncogenic potential due to controlling cell division (such as *p53*)[94–96]. Another

example is the Wnt-signaling pathway, which is a highly biologically conserved pathway found in almost all metazoan animals[97,98] and may even be studied in yeast[99]. This pathway has many important effects in cell fate determination and cell migration[97]. Mutations within the Wnt-signaling pathway are associated with canonical colorectal cancer, with numerous separate mutations to genes within the pathway often leading towards a similar effect[100]. Beyond this, disruption and dysregulation of the pathway is also a common factor across many cancers due to the vital importance of the pathway on cell fate[101,102].

There are many other pathways which are also highly evolutionary conserved, like MAPK/ERK signaling which are extremely ubiquitous across many cell types and tissues[103]. The MAPK-pathway in particular is highly involved in cell division due to its signal-transduction effect propagating transcription of many vital downstream genes[103]. Mutations within this pathway are also highly associated with cancer[104,105].

It is important to note that biological pathways do not function as discrete units in reality. They are metaphorical constructs to ease the description of many series of reactions happening in sequence, and often are interconnected and intertwined. However, analysing pathways in a simplistic linear form is helpful in interpreting their effects upon phenotype in that they provide groups of genes which can be studied together in the context of different phenotypes.

1.6.1 Using pathway enrichment to study pathways

A widely-used method of pathway analysis is the pathway-enrichment analysis and is often used to accompany GWAS analyses. Broadly, these analyses are conducted by extracting a list of genes associated with a trait or GWAS and searching for enrichment of pathway terms from online databases[106]. In GWAS specifically the genes selected for the analysis will often be the closest gene product to each significant locus within the GWAS[107].

These gene lists are then used in an enrichment analysis by comparing the list of genes to a background list (which usually comprises every gene in the human genome, but with exceptions based upon the selection criteria for the original list)[108]. The gene list is compared to the background list in the context of genes in biological pathways, often as defined by online databases such as KEGG[109], Reactome[110] an Wikipathways[111], and also resources such as Gene Ontology terms[112]. If then, proportionally there are more genes within the gene list

associated with a given pathway, function, or pathway term than should appear by chance, then the genes for this trait are considered to be enriched for this particular pathway, and thus this suggests a reasonable chance that this pathway is important for the given trait.

There are, however, some limitations with this approach that have to be considered. The first is the selection of the original gene list. As previously described, the genes selected are often those which are physically closest on the genome to significant loci in a GWAS. However, this is not evidence of a genuine relationship between the locus and that gene, and the physically closest gene may be the causal gene in only as few as 70% of cases[113–115]. While this is obviously the majority of loci, this does mean a significant number of selected genes selected will potentially have absolutely no relationship with the GWAS and the phenotype. Significant loci may instead be interacting with further away genes due to chromatin folding and long range enhancers[116] or may even be interacting with multiple different genes in the case of super-enhancers[117], so a gene list extracted this way may not be accurate or may be incomplete. Also, with the exception of the largest GWAS, there is likely insufficient power to find all significant loci as previously discussed. So, once again, genes may be missed from a list because there is simply not enough power to find all the associated loci.

Another limitation in pathway enrichment analyses is the pathway database itself which is used for examining the enrichment. Due to varying methods of curation and definition of pathways it is possible that the genes for a given pathway may differ between the various databases, making it harder to define true pathway-phenotype relationships[118,119]. This issue likely stems from the previously mentioned obstacle of pathways not being true biological constructs with no clear beginnings or ends, and as such defining the constituent elements used within a pathway is potentially subjective.

1.7 Polygenic risk scores

While GWAS and multi-omics are useful for expanding the understanding of the biology of complex traits, utilising them for prediction and applying the knowledge to developing therapeutics is a different matter. A primary goal of studying the proteome is the discovery of novel biomarkers for various traits, with the aim of many proteomics studies and panels being to expand knowledge of important biomarkers for clinical use[120,121]. One such example of a recent discovery driven by proteomics, is a panel of 20 proteins subset from Olink proteomics

which can be used to differentiate between benign tumours and the various stages of ovarian cancer, which could have great clinical utility in terms of analysing and treating the disease[122]. These technologies allow the creation of protein panels for screening, diagnosis and prognosis of disease and conditions[123].

However, while being able to screen for conditions based on novel biomarkers is invaluable, being able to predict disease and complex traits is a major goal in the field of precision medicine. As such being able to accurately predict a trait based on an individual's genetics has been a long-standing goal, allowing for the potential of screening of individuals based purely on their genome and prediction of risk to various complex diseases in the same way one can screen for BRCA mutations[124] to predict risk to breast cancer or screen for CAG repeats in the HTT gene in Huntington's[125].

In order to predict complex traits with many small loci which each provide a small effect towards the final phenotype, a commonly used tool is the creation of polygenic risk scores (PRS, and in some cases PGS). PRS is a summed total estimate of an individual's genetic predisposition to a given trait of interest[126]. In the creation of a PRS, SNP effects are extracted from their association with a phenotype and multiplied by an individual's genotype, usually in the form of dosage for the variant, providing an individual effect for each SNP. These SNPs are then aggregated into one single score, providing an estimate of genetic liability.

PRS are often calculated using many genetic variants across an individual's whole genome, not just those which pass the genome-wide significance. This is to do with the discovery that incorporation of non-significant variants across a whole genome can aid in improving heritability estimates from GWAS (I.e., including SNPs with extremely small effect sizes into the creation of the score)[127]. Thus, it is also predicted that as the power and scale of GWAS improve, so too will the accuracy of PRS in their predictive power[128].

There are numerous examples of creation of PRS for various traits, demonstrating also the improving capacity for PRS as the quality and scale of GWAS have improved. PRS for intelligence were created in 2013 which were able to predict ~2% of the variance of the trait between individuals[129] while a subsequent study in 2019 was able to predict ~11%[130]. PRS created from a GWAS of type-2-diabetes in 2018 had an area-under-curve (AUC) C-statistic of 0.66, meaning it was able to predict the difference between cases and controls in 2 out of 3 individuals[131]. PRS for breast cancer also have been created with an AUC of 0.63[132].

While this means that PRS are not yet capable of perfectly accurately predicting traits from genotype, particularly in extremely polygenic traits such as intelligence and height, the accuracy is rapidly improving. As mentioned, PRS have not yet been used in therapeutic scenarios but are now being trialed by the NHS in the prediction of heart disease[28]. Additionally previous work has already shown the utility of scores with PRS for CVD being able to predict heart disease with a 1.5 increase in accuracy compared to using a traditional biomarker C-reactive protein[29]. With PRS now becoming widely used in large cohorts such as UK Biobank[25], it is increasingly plausible that PRS will become used in conjunction with ordinary risk factors to provide a larger and more accurate risk assessment for complex disease for individuals[132,133].

There are limitations to PRS predictive power. A primary drawback of the methodology is that PRS created in a particular population will be increasingly poorly predictive in other more distantly genetic populations[134].

1.7.1 PRS calculation

There are, however, many possible ways of calculating PRS. Beyond simplistic models of genotype multiplied by effect, there are now increasingly complex and sophisticated methods of creating the scores. Clumping and thresholding (C+T) is a technique by which SNPs are clumped, and so only including SNPs which are loosely correlated and not in LD with each other, and a P-value threshold is imposed upon the SNPs used in creation of the score[135]. This is the least complex method for the generation of PRS and can be performed with tools such as PRsice-2[136,137].

Increasingly more complex forms of PRS creation have been developed in recent years. These algorithmic methods incorporate elements such as LD structure in order to weight SNP effect sizes by LD. There are also methods which utilise Bayesian shrinkage of prior effect sizes, lasso regression (and other forms of penalized regression of effect sizes), and more[138]. Briefly, some of the tools available include LDpred[139], LDpred2[140], SBayesR[141], PRS-CS[142], SBLUP[143], DBSLMM[144] and lassosum[145]. However, it is still a point of contention which method provides the best predictive scores.

1.7.2 PRS for gene expression and pathways

Given the intention of predicting complex traits from disease from genotype, understanding the biological mechanism for variation in traits may aid with improving prediction accuracy and also with prognosis and targeted intervention in those who would benefit the most. By incorporating multi-omics and pathway information it may be possible to more accurately predict complex traits in individuals based on variable gene and pathway activity.

PRS for gene expression have been created previously with examples such as gene-expression risk scores (GeRS) created from eQTLs derived from the GTEx consortium. By combining GeRS with more traditionally derived PRS it was possible to improve the predictive abilities of the score, demonstrating that by incorporating gene-expression of important genes it was possible to improve the power of discovery by improving specificity[146]. Another example is the PredictDB database, which contains polygenic weights for the SNPs for each gene in GTEx, created using PrediXcan[147].

Creation of these scores allows for the possibility of predicting traits based on individual gene levels (with mRNA expression levels used as a proxy for gene activity). It therefore also stands to reason that expansion of this to include pQTLs from proteomics or combining multiple individual genes into a single score for a pathway might also further improve predictive power by stratifying by biological pathways. This is particularly relevant because of the potential that QTLs and GWAS may capture different dimensions of variance[148] and so could be used to complement each in improving predictive capacity of genomic information.

1.8 Study aims: Taking a bottom-up approach to studying pathways

Pathway enrichment analyses are commonly used to try and demonstrate plausible biological mechanisms behind the associations discovered in GWAS, however due the issues described they may not always be accurate or complete. I, therefore, have attempted to take an alternative approach to searching for associations between pathways and phenotypes which I will describe in this thesis. To overcome these issues, the overarching aim of this study was to develop an alternative approach to identifying associations between pathways and phenotypes.

Pathway enrichment analysis could be seen as working from the top-down: following a GWAS for a particular phenotype, genetic loci are discovered associated with the trait, then a

list of genes close to these SNP associations are extracted and finally then this list is used to search for overrepresentation of certain pathway terms. As an alternative, in this study the goal was to develop a methodology which works in the opposite direction, starting with a pre-defined pathway and then examining genetic associations with the pathway and using these associations to test for relationships with phenotypes. The core principle of this method, dubbed PathWAS (**Path**way-**W**ide **A**ssociation **S**tudy), is to find a way to predict the effect of a biological pathway and use that effect then to search for associations with phenotypes.

The creation and testing the utility and efficacy of the PathWAS methodology forms the bulk of this thesis.

1.9 PathWAS theory

The concept behind PathWAS is to try and find complex traits associated with specific pathways, instead of using genes discovered from GWAS of traits and searching for pathways associated with them. In theory this method could provide unique insight into the etiology of complex traits by overcoming some of the issues described in this chapter. The theory of the approach is discussed in **Chapter 2**, but broadly the goal was to generate a PRS for the functionality of a given biological pathway. These scores were created with the assumption that each pathway reflects the sum of the effects of each of its constituent genes. So, by amalgamating these and leveraging a measured end-point of the pathway as a proxy for overall pathway function it should be possible to create an estimate of pathway function.

By combining the effects of multiple genes into a single score, it may be possible to discover new relationships between certain pathways and selected traits. Furthermore, by starting with a selected pathway and a defined gene set, the PathWAS approach could identify loci which have been missed using a GWAS approach. Additionally, the PathWAS approach might incorporate candidate genes and loci that were either ambiguous from GWAS, or even where the wrong gene was assigned in association with a significant SNP. If new pathway/disease associations can be discovered through this novel approach, it has the potential to reveal new therapeutic targets. Moreover, it could take us another step closer to precision or personalized medicine, as the PathWAS approach might reveal if an individual has a disease predisposition centered a single pathway

Through the discovery of novel pathway interactions as well it may then be possible to use downstream analyses to pinpoint druggable and actionable gene targets based on the pathway information, rather than examining individual gene-to-trait associations in a vacuum.

Chapter 2

Approach

2.1 Objective and requirements

The primary objective of the PathWAS methodology is to create a score for pathway functionality predicted by an individual's genetic make-up. We hypothesise that focusing on pathways may provide an increase in biological insight into disease and complex trait aetiology and allow for the discovery of new drug targets, and the possibility of developing targeted interventions based on individual genotype.

Figure 2.1 depicts the approach. Of the causal genetic variants which influence a phenotype, most will do so through regulating expression of a gene or the activity of its product. This variation in expression in turn may affect protein abundance and as a pathway component, will impact flux through the pathway and pathway outcome. We can divide the effect of genetic variation (G_i) on a pathway into three different parts: 1. the effect of the variant on gene transcription levels; 2. the effect of the gene transcription levels on protein levels; and 3. the effect of the protein level on the final pathway functionality.

The intention of the PathWAS methodology is to combine multiple PRS for gene expression into a single overall score that estimates the combined activity of these genes, and thus pathway activity. For this we assume that the abundance of a given protein product will directly correlate with its activity within a pathway.

As such, this necessitates using genetic information which effects gene expression, and we primarily utilise *cis*-eQTLs, which provide effect sizes for given SNPs for the expression of various mRNAs. We limit the selection to only cis-signals in order to maintain specificity of the selected QTL, as *trans*-QTLs will mediate their effect through other unknown means. While transcriptomic data and eQTLs exist for many genes, we have fewer measurements of protein abundance, as such we use the effect of *cis*-eQTLs as an estimate for the genetic effect on protein abundance, combining elements 1 and 2 of the values required (denoted as α in **Equation 2.1**). The final element (referred to as γ in **Equation 2.1**) is then the effect of overall gene expression levels on pathway function, for which there is no direct measure as there is no way of measuring an individual pathway's activity.

Therefore, the effect of G_i in a given pathway will be the product of α multiplied by γ (**Equation 2.1**), the effect of gene expression multiplied by gene activity in a pathway. While α_i can be obtained through direct association analysis (I.e., pQTLs as a direct measure or eQTLs as an estimate), the effect of γ_i is harder to measure. As it is not possible to gain a direct measurement of pathway functionality *in vivo*, as previously described there is nothing specific which can be measured, it is necessary to obtain an estimate for functionality through other means. We thus propose to use a protein which is a product of the pathway as a measure of the pathway functionality. Given that the effect of the pathway functionality on the end-point protein, as defined by genotype, will be constant, using the protein as a proxy of the pathway functionality will generate only a problem of scale (as the pathway function will only be measured in relation to that given protein rather than as a true measurement of functionality)., However, the relative weights of γ_i are still preserved and thus the utility of the score is preserved.

$$PathwayFunctionality_X = \sum_{ij} \alpha_{ij} \times \gamma_j$$

Equation 2.1.

Here, to gain an estimate of the function of pathway X based on genotype for an individual we need two values: α, the measure of effect of SNP i on overall expression of gene j and γ, the effect of the gene j on the pathway. The summed effect of α multiplied by γ is therefore the effect of the pathway based on SNP i.

Therefore, to comply with this equation there are three main components required (**Fig 2.1**, **Equation 2.1**):

1. A list of every gene (j) comprising the pathway and genomic variants associated with each (i).
2. An estimate of the effect of genotype on abundance of each individual gene product within the pathway, estimated primarily using eQTLs transcriptomics and mRNA levels, however pQTLs could also be used for the same purpose (α_{ij}).
3. An estimate of the effect of each gene on the overall pathway functionality, here using a measured protein as a proxy (γ_j).

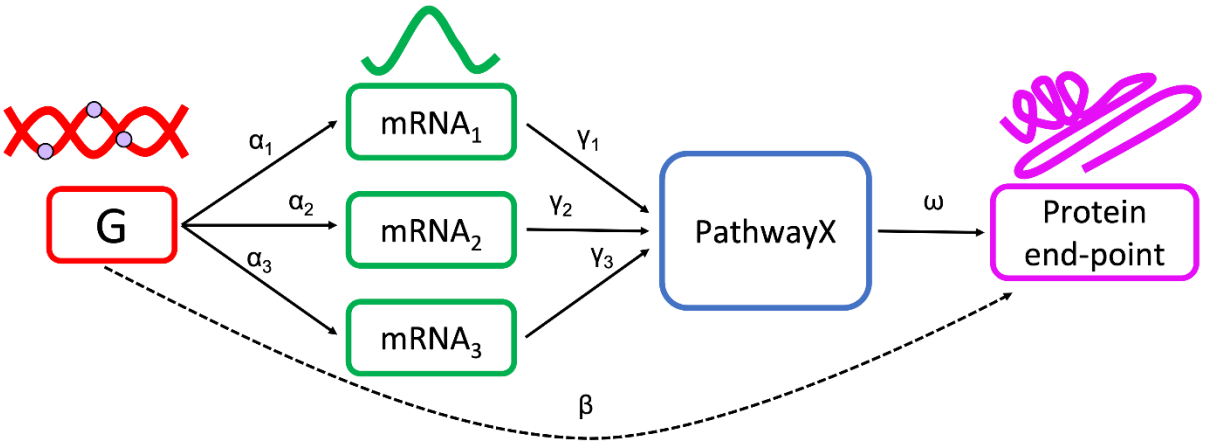


Figure 2.1. PathWAS methodology figure. The overall approach of the pathway method is depicted. The genetic make-up (G, composed of SNPs i) of an individual will lead to the regulation and variable transcription of numerous different genes in the form of mRNA. This effect is denoted as α for each gene (j = 1, 2, 3). Each gene will contribute a certain amount towards the overall function of a biological pathway X, denoted as γ. In order to estimate γ for each gene we use the genetic profile G and their relationship with a protein end-point as a proxy for the overall pathway function (β) as we assume that the relationship between the pathway effect and the protein measurement (ω) will be constant.

The following sections will outline how each of the needed parameters were estimated.

2.2 Analysing biological pathways

The first requirement for the proposed PathWAS method is to have a map of each pathway. In general, biological pathways are perceived as a sequence of biological reactions occurring between proteins (or gene products). This can include various forms of protein-protein interactions such as catalysis, degradation, cleavage, binding and so-on. Through these reactions the proteins will promote, inhibit or simply enable the function of the subsequent proteins or complexes in the chain. Therefore, if we assume that a biological pathway is a series of biochemical reactions occurring in a specific sequence, then theoretically the functionality of the pathway could be defined as the cumulative function of each reaction taking place.

It is worth noting that for this assumption we acknowledge that we treat pathways as discrete biological units with defined starting and ending points. This is obviously not a realistic picture of biology, as the pathways are simplified constructs which do not take into account other interconnected networks and reactions for ease of understanding. Accepting this caveat,

every pathway will have multiple inputs from other pathways feeding into it including feedback loops, with different pathways all modulating and regulating each other. Similarly, there is really no such thing as the "end" of a pathway as, broadly speaking, canonical pathways often end with the regulation of expression of other genes and proteins, therefore feeding into further biological processes (which, again, may in turn feed back into the pathway).

However, with our simplifying assumption in place, we will be able to define any pathway by the individual actors within it. Therefore, a relatively simple but core step of the PathWAS methodology was the creation of a list of genes working in concert within any pathway.

For this we rely on online databases of pathways, where it is possible to extract lists of genes and intermediate products based on prior analyses. There are numerous such online repositories including: KEGG[109,149], Reactome[150], Wikipathways[151], PANTHER[152], MSigDB[153] and more. From these it is possible to extract individual pathways and then a list of genes operating within the pathway. Naturally, however, the methodology could also be used with hand-curated gene lists.

2.3 The effect of genetic variation on gene expression

With a list of genes (or expressed mRNAs) for a pathway the next thing required is a way to predict the activity level of individual genes (**Equation 2.1**, j) from genotype (**Fig 2.1**, α). Again, as with pathways there is no direct measurement of gene function, so it is necessary to use a proxy for this. In this case, we use the expression profiles of the various genes, specifically mRNA or protein levels. Here, the reasoning is that modulation in the level of abundance of a given gene product will result in an equivalent modulation in the level of gene function. I.e., If the level of a protein within a cell is double that of another cell, one might expect the activity to be similarly doubled. This is also likely to be a too simplistic view as it does not take into account aspects like thresholds where increasing the level of a protein takes it over a threshold, below which there is no action and above there is action (or vice versa), then there would not be a linear relationship between levels and activity[154,155].

In order to estimate gene product abundance, we can use SNPs (**Equation 2.1**, i) taken from eQTL databases for assessing RNA levels. There are many studies which detail mRNA levels derived from transcriptomic analyses matched to genomic profiles with results available in

online repositories. So, we can extract the genetically predicted expression levels for each gene within the pathway. The two major examples of this we used throughout the project were the GTEx[76] and eQTLgen[78] databases.

This approach is not without its drawbacks, however. The first of these is that the use of eQTLs are not direct measurements of gene levels, however it is impractical to expect anyone using the method to have access to all elements of raw data. A second concern is that as proteins and gene products are those which conduct the activity of any given pathway. As such, it would likely be better to use proteomics and pQTLs as direct measurements of protein abundance rather than use transcriptomics and eQTLs to merely estimate them. This would allow studying the actual level of the protein product and hence level of function. This argument is particularly compelling as there is substantial evidence suggesting that it is challenging to predict protein levels from eQTLs[83], meaning that direct changes in gene expression at the RNA level do not always result in an equivalent change in the quantity of translated protein. There are numerous possible reasons for this discrepancy, including variable splicing, changing levels of degradation, differing levels of translation from ribosomes and polysomal attachment as well as the consequences of post-translational modifications[88]. Even things like export of proteins from highly expressed locales to the rest of the body may play a role in this[88]. Considering these factors together, it makes the argument for using pQTLs instead of eQTLs as the measurement for gene expression and function, all the stronger.

However, despite these issues, there are good reasons to primarily use eQTLs. First and foremost is the ready availability of many eQTL datasets, from many different tissues and sometimes with very large sample sizes that proteomic datasets have yet to match. A more compelling argument however is that the PathWAS method relies on using multiple gene signals to provide an overall estimate for a pathway, and therefore it is most useful to have access to as many significant QTLs for as many different genes as possible. This is where transcriptomic data retain a significant edge over proteomics as even some of the largest proteomic studies to-date have examined a limited array of proteins, while transcriptomics can feasibly examine expression of every single different gene and thus can produce more complete datasets. For example in a direct comparison of two of the largest datasets of transcriptomics and proteomics there are ~19000 gene signals measured in eQTLgen[78] compared with <5000 in the DeCODE proteomics[156].

To consolidate overall genetic effects on the genes, the QTL SNPs are summed into polygenic risk scores (PRS) for the expression of each gene (PRS_{Gene}) (**Equation 2.2**). Following this equation allows the creation of one single score for each gene in every individual within a dataset.

$$PRS_{Genej} = \sum_{ik}^{N} G_{ik} \times \alpha_i$$

Equation 2.2

Here, to create PRS_{Gene} (for an individual, k, in population N relating to gene j), the total sum of the dosage (G) for SNP i in individual k is multiplied by an effect size for the SNP. For each PRS_{Gene}, we specifically use the effect (α) of each QTL SNP from the expression data (either transcriptomic or proteomic).

This calculation is done using only *cis*-instruments for each gene because *trans*-associations will be associated with the expression levels for a gene through other unknown mediating elements. This could include modulating the level of expression of a different gene which in turn modulates expression of the examined gene, or alternatively may be involved in regulating cell numerosity and thus influencing expression levels of the analysed gene. In either instance this means that incorporating *trans*-QTLs would be including other unknown factors and variables into the creation of the pathway model which might not be involved with the pathway function.

As an addendum, we also use PRS creation software (described in the methods) which weight the effect of SNPs by genetic architecture and LD. As such α is usually a SNP effect weighted by LD structure.

2.4 The effect of genes on pathways

The third and final requirement for the PathWAS method is a measurement of the effects of each gene on the pathway. Here we assume that a certain level of expression will correspond to a certain level activity of a gene, and thus it will contribute a certain amount to the functionality of the pathway. We then assume that the overall pathway functionality is the cumulative effect of each gene within it.

With this in mind, we need some way of measuring the contribution of each gene on the overall pathway. Once again, without direct measurements of gene and pathway effects this is challenging and so we used a novel method to try and estimate it. We selected a gene at the end of a given pathway for which we have a measurement in a proteomics dataset and used this measurement as a proxy for pathway functionality (**Equation 2.3**).

We assume that within a given pathway, the levels of the end-point protein will reflect the combinatorial contributions from the varying levels of each protein at each step of the process, so that the levels of a protein at the end a given pathway will be representative of the total function of the pathway. Therefore, by analysing the relationship of individual genes with this protein "end-point" we are, in effect, regressing genes against a proxy for pathway functionality and can gain an estimate for the effect of each gene on the overall pathway.

Expressing this in mathematical terms:

$$\beta_{ij} = (\alpha_{i1} \times \gamma_1 + \alpha_{i2} \times \gamma_2 + \alpha_{i3} \times \gamma_3 \dots \alpha_{ij} \times \gamma_j) \times \omega$$

Equation 2.3

Where β is the combined effects of individual genetics based on SNP i and gene j has on the end-point protein. α is the effect of SNP i on gene j (as described in **Equation 2.1,** this is taken from eQTL or pQTL data) and γ is the effect of the gene j on overall pathway functionality. ω is the scale or weighting factor linking the pathway functionality to the protein abundance, used as a proxy for measurement of pathway activity. We can disregard ω as it represents only a scaling factor and as such is constant for all parameters. If we have multiple SNPs which are instrumental variables to the expression of at least one gene, then this equation is the same as the regression form of multivariable Mendelian randomisation[157].

Throughout the thesis I use two methods of obtaining γ for each gene. In **Chapter 5**, with PathWAS v1, multivariable linear regression is used to regress an expression PRS for each gene against the end-point, and the coefficient for each gene is used as γ. In Chapter 6 I instead use multivariable Mendelian randomisation (MR), and it is the MR effect of each gene in the model that is used as γ.

One addendum to this is that with the use of specific end-points within pathways we decided to refine the gene lists for each pathway one step further. As many biological pathways have branching steps within the network, it is important that we only use the effects of genes upstream of our selected end-point protein and so the pathways were pruned to exclude branches which did not have an inward effect on the protein. Here an inward effect is defined

based on graph theory as a given node (gene) influencing its neighbours, due to a downstream effect.

2.4.1 Multivariable Mendelian randomisation

In order to estimate the effects of the individual genes on the end-point protein (and thus the pathway function) while accounting for possible interactions between the genes within the pathway we use both linear regression and multivariable Mendelian randomization (MVMR).

MR is a statistical technique regularly used in epidemiology and genetics to investigate the causal relationships between an exposure (often a modifiable risk factor) and an outcome (such as a disease or trait). It leverages genetic variants that are associated with the exposure of interest but are not influenced by the outcome, acting as instrumental variables to aid in establishing causality between the exposure and outcome. As alleles are randomly distributed during reproduction, in theory MR has the advantage over observational studies in that it is potentially less prone to bias from confounding, acting in a way like a randomly controlled trial. In this instance, the exposures analysed would be each gene in a given pathway and the outcome would be the end-point protein. In MVMR then, by conducting the MR of each gene on the end-point simultaneously this provides an estimate for the effect of each gene on pathway function in the context of all the other genes included in the model.

For the MVMR we use the eQTLs for the genes from the pathway, utilising only the significant *cis*-associations. These SNPs were clumped, a method of prioritizing significant signals by removing redundant associations due to LD structure, and these clumped SNPs were then used as the instrumental variables against the SNPs from a GWAS of the end-point proteomic measurements. The result from this is that we are able to obtain an effect for each gene associated with the pathway (γ), either from linear regression or MVMR. We then combine γ this with the effect of variation on gene expression (α) in order to obtain the measure of pathway effect β (**Equation 2.3**).

Utilising MVMR instead of linear regression has the additional benefit that the effect for each gene is created taking into account the effects of every other gene as well, providing a weighted effect score for every actor within the pathway.

2.5 Creation of the pathway score

The final step of the methodology was then to combine each element discussed and use that to create a model for the pathway. The gene list was extracted from the selected pathway database and pruned to include only genes relevant to the end-point (as described in refining the genes to only those with inward facing edges in the pathway graph). *Cis*-eQTLs for each of these genes were used to create PRS_{Gene} in order to create an estimate of the effect of genotype on each gene from the pathway. The eQTLs were then also used to estimate the contribution of each gene on the overall pathway efect by running a MVMR of the QTL SNPs against SNPs from a GWAS for a protein end-point - which is used as a proxy for functionality

With the score for the effect of every gene on the end-point protein it is then possible to multiply this by the PRS_{Gene} for the corresponding gene (**Equation 2.4**). This creates a final score for the effect of each gene in a pathway for every individual. Then, we summed the scores for every gene used and this created an overall pathway PRS ($PRS_{Pathway}$) for every individual. With a $PRS_{Pathway}$ we therefore have an estimated effect of the pathway from genotype and can use this in PheWAS analyses of complex traits to search for associations between each pathway and different phenotypes.

$$PRS_{PathwayX} = \sum_{ij}^{N} PRS_{Genej} \times \gamma_j$$

Equation 2.4

PRS were created for each gene j as described in Equation **2.1**. These PRS were then multiplied by the corresponding effect of the gene on the end-point protein (γ). When each of these weight PRS are summed, this provides a PRS of the estimate of effect of the total pathway from genotype.

Chapter 3

Materials and methods

3.1 eQTL data sources

3.1.1 GTEx

GTEx is an online repository of tissue-specific RNA-seq and eQTL data[76]. This has been downloaded from historical versions V7 and V8. At present (V8 and V9, with V9 including no new data from V8) this incorporates data from up to 838 individuals and 49 tissues used for eQTL analysis. This RNA-seq is conducted from tissue samples extracted post-mortem.

The eQTL data used for multi-tissue scores was from the V7 release, this included fewer tissues and a lower sample size for each respective tissue. The multi-tissue scores and first iteration of PathWAS also utilised the genetic transcripts per million (TPM) file from V7. However, later stages of the project and for the PathWAS package utilised the TPM file from V8.

Only *cis*-signals were used in the creation of expression PRS, with *cis* defined by GTEx as a 1 Mb window around the transcription start site.

3.1.1.1 PredictDB

The PredictDB dataset is a set of polygenic weights created from the GTEx V7 and V8 eQTL datasets using the Predi-X-can methodology[158]. These weights are publicly available here: https://predictdb.org/post/2021/07/21/gtex-v8-models-on-eqtl-and-sqtl/

3.1.2 eQTLgen

The eQTLgen consortium consists of RNA-seq data and eQTL meta-analysis (both in *cis* and *trans*) in whole blood samples from 31,684 individuals[78]. This analysis uses expression profiles of 19,942 genes across 11 million variants. The data includes significant (FDR < 0.05) and non-significant SNPs for all eQTLs. For the purposes of this project only *cis*-signals were used in the creation of expression PRS, with *cis* defined by eQTLgen as being within a 1 Mb window from the centre of the gene. The significant and non-significant *cis*-eQTLs are available here: https://www.eqtlgen.org/cis-eqtls.html

As we were creating PRS_{Gene} using the *cis*-eQTL data, we required betas and standard errors for the effect sizes for each SNP, which is not available in the eQTLgen data. Therefore, we had to estimate both the beta and standard error for each SNP as follows:

$$\beta = \frac{Z}{\sqrt{2 \times freq \times (1 - freq)} \times (N + Z^2)}$$

Equation 3.1: Estimation of SNP effect size (β) in eQTLgen, where Z is the provided Z-score for the SNP, freq is the allele frequency, and N is the sample size.

$$SE = \frac{1}{\sqrt{2 \times freq \times (1 - freq)} \times (N + Z^2)}$$

Equation 3.2: Estimation of SNP effect standard error (SE) in eQTLgen, where Z is the provided Z-score for the SNP, freq is the allele frequency, and N is the sample size.

For these equations we required the allele frequency for each SNP, which was obtained from: https://molgenis26.gcc.rug.nl/downloads/eqtlgen/cis-eqtl/2018-07-18_SNP_AF_for_AlleleB_combined_allele_counts_and_MAF_pos_added.txt.gz

3.2 Population cohorts

3.2.1 ORCADES

The Orkney Complex Disease Study (ORCADES) cohort is a deeply phenotyped and family-based population cohort from the isolated Orkney Isles in Northern Scotland[159]. Comprised of 2078 individuals between 16-100 years of age, who were required to have at least 2 grandparents from Orkney, the cohort demonstrates less genetic diversity in comparison to mainland Scotland due to the increased levels of endogamy. As well as a rich set of phenotypes, the vast majority of the ORCADES cohort has been genotyped and fasting blood was collected for use in omics assays.

The ORCADES proteomics data are frequently used throughout my thesis, with the OLINK proteomics measurements used as for prediction of PRS_{Gene} r2 chapter 4 and, as a training data set before prediction in Vis in chapter 5 and then as a prediction data set based from the SCALLOP consortium in chapter 6.

DNA was extracted and genotyped using three Illumina arrays. Samples were removed if they had a call rate <98%, as were ethnic outliers, duplicates within the data set, gender mismatches or samples presenting identity-by-state (IBS) which was incompatible with the pedigree. Following QC, 854 samples remained on the Hap300, 301 on Omni1 and 1,067 on the OmniX (**Table 3.1**).

	Hap300	Omni1	OmniX
N SNPs pre-QC	293,687	1,016,138	743,427
MAF filter	1%	monomorphic	monomorphic
HWE filter	10-6	10-6	10-6
Call-rate filter	97%	97%	97%
N SNPs post-QC	287,208	843,723	654,651

Table 3.1. Genotype Quality Control in ORCADES including the number of variants prior to and following QC and the filters used within each genotyping array.

The genotype data for ORCADES was previously imputed by colleagues within my group using the following parameters before being passed to me. They were phased using Shapeit2 v2.r873 and duohmm3 software and then imputed with HRC.r1-1 using the Positional Burrows-Wheeler Transform (PBWT) algorithm on the Sanger imputation server. The final ORCADES imputed genotypes used within my project was composed of 12,696,745 SNPs (NCBI Build b37).

3.2.2 CROATIA-Vis

The CROATIA-Vis cohort (henceforth just called "Vis") is another isolated population cohort from the Dalmatian island of Vis[160]. Comprised of 1,008 individuals, between the ages of 18-93, these individuals underwent medical examination and interview (led by the Institute for Anthropological Research and the Andrija Stampar School of Public Health, Zagreb, Croatia). Fasting blood was drawn and used for later omics analyses along with phenotyping and genotyping.

Genotyping for Vis used the Illumina HumanHap300v1 array as previously described by Bretherick *et al.* (2020)[159] with PCAs calculated utilising GenABEL and Plink 1.9 to remove those with outlying ancestry. Imputation was also performed as with ORCADES using the HRC.r1-1 panel.

3.2.3 UK Biobank

The UK Biobank is a deeply phenotyped and genotyped population cohort of individuals from the UK consisting of, initially, 488,377 participants. These individuals were genotyped utilising two similar genotyping arrays as described in Bycroft *et al.* (2018)[161]. Briefly a subset of 49,950 individuals, part of the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study were genotyped using an Applied Biosystems UK BiLEVE Axiom Array (Affymetrix now Thermo Fisher Scientific). This consisted of 807,411 SNPs described by Wain *et al.* (2015)[162]. The remaining 438,427 participants

were genotyped using a similar Applied Biosystems UK Biobank Axiom Array (consisting of 825,927 SNPs) which overlapped heavily (~95%) with the BiLEVE array. The variants selected primarily consisted of SNPs and indels, many selected for pre-existing known associations, as well as including numerous rare markers (<1% MAF). The UK Biobank data was subsequently imputed with the HRC reference panel[161].

3.2.3.1 UK Biobank phenotypes

An initial set of 44 phenotypes from UK Biobank was used in the exploratory PheWAS in **Chapter 5** was used. This was then later expanded to 60 phenotypes which were included in the final association analysis between the PRS created and UK Biobank traits. A full list of these phenotypes and their relating method of measurement is in **Table 3.2**. Of these traits the majority were direct measurements taken from the UK Biobank phenotype catalogue, with the exception of age of death traits (lifespan) and type 2 diabetes.

To identify individuals with Type 2 diabetes, a previously established algorithm was used[163], which included a combination of self-reported illness, hospital admission records, and death records. This algorithm categorized individuals as cases if they had reported having "diabetes" or "type 2 diabetes" were diagnosed with type 2 diabetes during hospital admission using specific codes, or if their death record indicated these diabetes codes as the cause of death. Those who did not meet these criteria were classified as controls.

The survival of participants was determined by comparing their age at the time of assessment with their age at the time of death, as recorded on their death record, or their current age at a specified date, if the participant's status was censored[164]. For determining parent survival, the subject's reports of their parents' age or age at death were taken into account. However, parental deaths before the age of forty were excluded to minimize the possibility of including deaths due to accidents instead of diseases.

Prior to conducting the association analysis, a final step of quality control was conducted to remove outliers from the data by removing individuals with a raw Z-score >99 standard deviations from the mean and a residual Z-score >10 standard deviations from the mean. In both cases these large values for Z-score threshold were used in order to keep the vast majority of all data measurements, specifically including outliers, and only exclude data which had been incorrectly coded with extreme values.

Phenotype	Classification	Form of analysis	Phenotypes derived from	PathWAS v.1 PheWAS
Waist circumference	Anthopometry	Linear regression	UK Biobank field ID: 48	✓
Hip circumference	Anthopometry	Linear regression	UK Biobank field ID: 49	✓
Height	Anthopometry	Linear regression	UK Biobank field ID: 67	✓
Weight	Anthopometry	Linear regression	UK Biobank field ID: 21002	✓
Fitness workload	Anthopometry	Linear regression	UK Biobank field ID: 6032	✓
BMI	Anthopometry	Linear regression	UK Biobank field ID: 21001	✓
Total fat mass	Anthopometry	Linear regression	UK Biobank field ID: 23278	✓
Total fat free mass	Anthopometry	Linear regression	UK Biobank field ID: 23279	✓
Total lean tissue mass	Anthopometry	Linear regression	UK Biobank field ID: 23280	✓
Fat percentage	Anthopometry	Linear regression	UK Biobank field ID: 1343	
Leukocyte count	Blood cell count	Linear regression	UK Biobank field ID: 30000	✓
Lymphocyte count	Blood cell count	Linear regression	UK Biobank field ID: 30120	✓
Monocyte count	Blood cell count	Linear regression	UK Biobank field ID: 30130	✓
Neutrophil count	Blood cell count	Linear regression	UK Biobank field ID: 30140	✓
Eosinophil count	Blood cell count	Linear regression	UK Biobank field ID: 30150	✓
Basophil count	Blood cell count	Linear regression	UK Biobank field ID: 30160	✓
Erythrocyte count	Blood cell count	Linear regression	UK Biobank field ID: 30010	✓
Platelet count	Blood cell count	Linear regression	UK Biobank field ID: 30080	✓

Reticulocyte count	Blood cell count	Linear regression	UK Biobank field ID: 30250	✓
Red blood cell count	Blood cell count	Linear regression	UK Biobank field ID: 30170	✓
Platelet crit	Blood cell count	Linear regression	UK Biobank field ID: 30090	✓
Rib BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23233	✓
Spine BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23234	✓
Heel BMD	Bone mineral density	Linear regression	UK Biobank field ID: 3148	✓
Right Femur BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23212	✓
Left Femur BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23291	✓
Trunk BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23241	✓
Total BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23236	✓
Arm BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23225	✓
Head BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23226	✓
Legs BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23231	✓
Pelvis BMD	Bone mineral density	Linear regression	UK Biobank field ID: 23232	✓
Systolic BP	Heart function traits	Linear regression	UK Biobank field ID: 93	✓
Diastolic BP	Heart function traits	Linear regression	UK Biobank field ID: 94	✓
Heart rate	Heart function traits	Linear regression	UK Biobank field ID: 102	✓
Cardiac output	Heart function traits	Linear regression	UK Biobank field ID: 22424	
Father's lifespan	Lifespan	Survival analysis	As described in text	
Mother's lifespan	Lifespan	Survival analysis	As described in text	

Mother's age at death	Lifespan	Linear regression	UK Biobank field ID: 3526	✓
Father's age at death	Lifespan	Linear regression	UK Biobank field ID: 1807	✓
Age of death	Lifespan	Survival analysis	As described in text	
FVC	Lung function traits	Linear regression	UK Biobank field ID: 3062	✓
FEV1	Lung function traits	Linear regression	UK Biobank field ID: 3063	✓
PEF	Lung function traits	Linear regression	UK Biobank field ID: 3064	✓
Creatine in urine	Metabolite measurements	Linear regression	UK Biobank field ID: 30510	
Potassium in urine	Metabolite measurements	Linear regression	UK Biobank field ID: 30520	
Sodium in urine	Metabolite measurements	Linear regression	UK Biobank field ID: 30530	
Vitamin D blood content	Metabolite measurements	Linear regression	UK Biobank field ID: 100021	
Age serious illness (non-cancer) first occurred	Serious illness	Linear regression	UK Biobank field ID: 87	
Number of serious illnesses (SR)	Serious illness	Linear regression	UK Biobank field ID: 135	
Age of first cancer diagnosis	Serious illness	Linear regression	UK Biobank field ID: 94	✓
Number of cancers (SR)	Serious illness	Linear regression	UK Biobank field ID: 134	✓
Age of cancer diagnosis	Serious illness	Linear regression	UK Biobank field ID: 40008	✓
Number of cancers reported	Serious illness	Linear regression	UK Biobank field ID: 134	✓
Age heart attack diagnosed	Serious illness	Linear regression	UK Biobank field ID: 3894	✓
Interpolated age first cancer diagnosis	Serious illness	Linear regression	UK Biobank field ID: 20007	
Interpolated age first illness	Serious illness	Linear regression	UK Biobank field ID: 20009	✓
Diabetes	Serious illness	Logistic regression	As described in text	

Diagnosed with cancer	Serious illness	Linear regression	UK Biobank field ID: 40009	✓
Mother Alzheimer's	Serious illness	Linear regression	Incidents of "Alzheimer's" extracted from UK Biobank field ID: 20110	

Table 3.2. Table of phenotypes used in UK Biobank PheWAS. Provides the UK Biobank field ID for the majority of phenotypes used. Also lists those phenotypes used in the PathWAS v.1 analysis described in **Chapter 5**.

3.3 Proteomics cohorts

3.3.1 ORCADES

The ORCADES cohort proteomics were measured in 1,048 individuals of the ORCADES cohort, using fasting EDTA plasma samples. Proteins were measured using the OLINK Proseek Multiplex panels with Olink PEA technology. Twelve proteomics panels have been measured in ORCADES: Cardiovascular II and III (CVD2 and CVD3), Cell Regulation (Cell-Reg), Development (DEV), Inflammation (INF), Neurology I (NEU), Neuro-exploratory (NEX), Metabolism (MET), Organ Damage (OD), Oncology II (ONC2), Immune Response (IR) and Cardiometabolic (CVD1). This provided a total of 1068 unique protein measurements.

My project used five of these protein panels across three different phases. During the multi-tissue analysis phase (Chapter 4), we used all of the direct protein measurements to test the predictive power of PRS_{Gene} created. During the first phase of the PathWAS methodology we focused on the proteins overlapping between ORCADES and Vis, comprising a total of 261 proteins in the INF, CVD2 and CVD3 panels (after QC).

Lastly, during the final iteration of the PathWAS methodology, we utilised the ORCADES proteomics as a prediction dataset for those which overlapped with the SCALLOP consortium in the CVD2, CVD3, NEX and NEU panels: a total of 368 proteins.

The measurement, and QC of each of these protein panels was conducted by my colleagues and handed to me.

3.3.2 Vis

Three Olink panels were used for measuring serum proteins in approximately 910 individuals of the Vis cohort. These panels were the CVD2, CVD3 and INF panels. Following QC this left a total of 261

unique proteomics measurements which overlapped between the ORCADES and Vis proteomics sets for the results in Chapter 5.

3.3.3 SCALLOP consortium

The SCALLOP consortium ran a meta-analysis across numerous cohorts for four of the Olink panels, resulting in creation of summary statistics for 368 proteins measured in CVD2, CVD3, NEX and NEU.

The proteins for CVD2 and CVD3 are described in Macdonald-Dunlop *et al*. (2021)[165] while the proteins for NEX and NEU are described in Repetto *et al*. (2023)[166].

The meta-analysis for these proteins was ran as follows:

A second meta-analysis of the SCALLOP proteomics was conducted by my colleagues Linda Repetto and Erin Macdonald-Dunlop, with the same data provided for the papers mentioned above, with the exclusion of the ORCADES cohort. This was performed so that the ORCADES proteomics could be used as an external validation data set for prediction of the pathway models in Chapter 5. The meta-analysis was conducted using the METAL software and the following variables: average frequency on, min and max frequency on, additive sample size, an imputation quality cut off of 0.4 and genomic control off.

Briefly, a GWAS of the proteomics measurements was conducted across several cohorts and then meta-analysed as part of the SCALLOP proteomics consortium. There is a maximum of 26,494 individuals in the CVD2 and CVD3 panels and a maximum of 12,174 and 5,011 for the NEU and NEX panels respectively. For the purposes of the PathWAS analysis, the ORCADES cohort was removed from these meta-analyses and instead used as a separate prediction cohort, meaning the sample size for the four panels is 25,424 (for both CVD panels), 11,188 and 4,034 respectively for NEU and NEX.

3.3.4 DeCODE

The DeCODE proteomics is a data set of 4,907 protein aptamers, measuring 4,719 unique proteins using SomaLogic technology. These proteomics were measured in 35,559 Icelanders and described in Ferkingstad *et al.* (2021)[156] with summary association statistics publicly available from: https://download.decode.is/form/folder/proteomics. The proteomics were measured from EDTA-derived blood plasma samples with the SomaScan version 4 assay and then pQTL data was made available following GWAS of each protein aptamer. These pQTLs were used for the PathWAS

multivariable MR in Chapter 6, although as we only had summary statistics available we could not perform the PathWAS prediction step which requires the individual measurements.

3.3.5 INTERVAL

The INTERVAL study consists of a randomised selection of ~50,000 healthy blood donors from England, initially set up to study the safety of frequency blood donation and examining phenotypes of these individuals[167]. These individuals were subsequently sequenced with 830,000 variants using an Affymetrix Axiom 2.0 Assay array and subjected to standard Affymetrix quality control (a dish QC < 0.82, a call rate < 97%, more than three clusters or had cluster statistics based on Fisher's linear discriminant). Following QC this data was imputed with data from the 1000 Genomes project. This sequencing is described completely in Astle *et al.* (2016)[168].

From these ~50,000 samples, 3,301 (post quality control) were randomly selected in two sub-groups for proteomics analysis. Using the SomaScan v.3 assay, 4,034 aptamers targeted to 3,622 plasma proteins were used to measure protein abundance before subject to GWAS to conduct association testing with each protein[169].

Subsequently INTERVAL has also incorporated OLINK proteomics[170], however for the purposes of this project only the SomaLogic proteomics were used.

3.4 Proteomics technologies

3.4.1 Olink

The results of two separate proteomics technologies were utilised during this project. The first was the SCALLOP proteomics, utilising Olink technology. Olink proteomics uses a proximity extension assay (PEA) methodology for targeting and measurement of proteins. This involves the use of antibodies with attached pairs of oligonucleotides which hybridize only with the opposing member of the pair. Only binding of both antibodies to the target protein allows extension of the sequence via qPCR and thus measurement of the protein in the sample.

3.4.2 Somalogic

The Somalogic SOMAscan assay utilises modified DNA oligonucleotides (aptamers dubbed SOMAmers) which bind to folded proteins. The quantity of these SOMAmers, bound to their

constituent proteins, can then be measured via a custom DNA array, allowing measurement of the target protein via measurement of the DNA bound to the surface.

3.5 PRS creation methods

3.5.1 LDpred

LDpred is a Python package developed by Vilhjálmsson, B.J. *et al.* (2015)[171]. The algorithm inputs summary level data and takes into account LD, from an external reference genotype, between SNP associations to create polygenic weights under a Gaussian distribution. For the purposes of this project, we utilised the 1000 Genomes project as the reference genotypes for establishing LD architecture. The software utilises both an infinitesimal model, which assumes all genetic variants are causal, and a non-infinitesimal model in which only a given fraction (P_r) of variants are causal. In the case of the infinitesimal model the variant weights are calculated under the distribution of Equation 3.3.

$$\beta_i \sim N(0, \frac{h^2}{M})$$

Equation 3.3: LDpred infinitesimal model equation.

Where h^2 is the heritability, estimated from the summary statistics and accounting for sampling noise, and M is the number of genetic variants. In the case of the non-infinitesimal model, a Gaussian mixture prior is assumed with probability p and where $\beta_i \sim 0$ has probability (1-p). This is calculated in the distribution described by **Equation 3.4**.

$$\beta_i \sim N(0, \frac{h^2}{Mp})$$

Equation 3.4: LDpred non-infinitesimal model equation.

The posterior mean effects of the variants are calculated using a Markov chain Monte Carlo (MCMC) Gibbs sampler for the non-infinitesimal sample, this takes the LD structure into account from the reference genotype, and effect sizes calculated for different fractions of causal variants and for the model in which all variants are causal. The software was downloaded from https://github.com/bvilhjal/ldpred.

Usage of LDpred is a two-step process where the first command is as follows:

```
python coord_genotypes.py --gf=PLINK_LD_REF_GENOTYPE_FILE --ssf=SUM_STATS_FILE --
N=SS_SAMPLE_SIZE --out=OUT_COORD_FILE
```

Inputting a reference Plink genotype file (which in my case would be a Plink genotype file for ORCADES, Vis or UK Biobank for the chromosome of the gene in question) and the summary stats (here I used the downloaded summary stats from GTEx). Aside from this the sample size was supplied based on the sample size for the given summary stats (an in the case of GTEx, the sample size of the tissue in question).

This command produces a coordinate file which is used by the second command:

```
python LDpred.py --coord=COORD_DATA_FILE --ld_radius=LD_RADIUS --
PS=FRACTIONS_CAUSAL --N=SAMPLE_SIZE --out=OUTPUT_FILE_PREFIX –H2=HERTIABILITY
```

Here the coordinate file from the first step is input as is a manual value for LD radius, which was defined as 10,000 based on LDpred recommendations, the PS argument was used to define the proportion of causal variants (P_r) for which multiple values were used. Where it was available as well, a heritability estimate was provided from GTEx. This would create the PRS for the given summary stats/gene.

3.5.2 Predi-X-can

The method of Predi-X-can was developed by Gamazon E.R. *et al.* (2015)[158] in the Im lab, designed specifically to create polygenic weights from GTEx data. In theory, like other PRS methods, it allows for the discovery of smaller effect sizes than would normally be discovered in standard association testing. From the GTEx tissue eQTLs, Gamazon *et al.* have also created a data repository of the polygenic weights for each locus, integrated with an additional dataset: the Depression Genes and Networks eQTLs (922 whole-blood samples). This data, called PredictDB, can be accessed freely from http://predictdb.org/.

The package calculates overall effects for *cis*-regulatory variants within 1 Mb of relevant genes using elastic net regression. The weights stored in the PredictDB database are calculated with the Equation 3.5 from the GTEx and DGN summary data:

$$T = \sum_k w_k X_k + \varepsilon$$

Equation 3.5: Predi-X-can model equation.

This formula calculates the estimated level of genetically regulated expression or $\widehat{\mathrm{GReX}}$. Where T is the expression of the gene and X_k is the number of reference alleles for SNP k. From this, the Im lab suggest the data can be applied for the "imputation" of transcriptome data from GWAS summary statistics in a method analogous to classic imputation of genotypes.

3.5.3 LDpred2

LDpred2 represents an expansion of the LDpred methodology, translated into R from Python as part of the bigsnpr R package[140].

Specifically, we primarily utilised the ldpred2_inf and ldpred2_auto functions, the former of which is an extension of the previous version of LDpred and assumes all variants are causal. The latter automatically estimates h^2 (SNP heritability) and p (proportion of causal variants, or in this case sparsity) from the data and is conducted free of hyper-parameters, and as such does not require an external validation dataset to estimate these.

In using LDpred2 we were able to estimate the effect sizes of each SNP in any given dataset conditional to all other SNPs using the LDpred2-auto function. This also relied upon usage of a reference dataset of 10,000 individuals of European descent, extracted from the UK Biobank. PRS for each gene (PRS_{Gene}), for each individual was then estimated as:

$$PRS_j = \sum_i^N \beta_i \times G_{ij}$$

Where β_i is the predicted weight of the SNP estimated by LDpred2 and G_{ij} the dosage for SNP i in person j. In a limited number of instances, LDpred2 was unable to estimate h^2 for a given gene and it was thus not possible to obtain weights for the local PRS. In these cases, PRS_{Gene} were calculated with PRsice-2[137].

3.5.3.1 LDpred2 reference panel

A requirement for LDpred2-auto is the usage of a reference panel for LD structure. The LDpred2 methodology and subsequent iterations have recommended the use of a custom HapMap3 panel of SNPs (named the HapMap3+ panel), and to refine the variants used in the creation of PRS to these.

However, the creation of each PRS_{Gene} utilised an already small number of SNPs as we select only the *cis*-signals from eQTL data (with the creation of many PRS_{Gene} involving the use of fewer than 5,000 SNPs). As such refinement based on HapMap3+ prevented the function of LDpred2 in many cases due to the estimates of heritability being too low, likely due to including too few SNPs in the weight creation. Therefore, in order to be able to create as many PRS_{Gene} as possible which could most accurately reflect the population, I instead used all SNPs from the eQTLs which could be found within a UK Biobank reference panel of Europeans.

For this I used a reference panel of 10,000 randomly selected, unrelated individuals of white British ancestry from the UK Biobank. This includes the fully custom imputed genotypes of all 10,000

individuals and these are then used as the custom LD reference panel. This was created by Arianna Landini.

3.5.3.2 Allele alignment

As part of every PRS creation methodology it is vital to align alleles between the summary statistics and the reference genotype. This is in order to ensure that the beta for the effect allele in the summary statistics corresponds to the correct allele in the reference genotype. While PRS-CS and PRsice-2 both incorporate allele flipping as part of their software, for LDpred1 and LDpred2, a custom allele flipping script was created, which would be incorporated into the PathWAS R package.

For any given SNP, this script confirms that the reference and effect allele corresponds between summary statistics and the reference genotype and if it does not, it inverts the beta. It also checks this with alleles on the opposing strand (e.g., comparing a SNP that has alleles A and G in the summary statistics and alleles T and C in the reference).

3.5.4 PRS-CS

PRS-CS creates PRS using a Bayesian regression algorithm using the prediction formula described in Ge *et al.* (2019)[142] (Equation 3.7).

$$y_{N\times 1} = X_{N\times M}\beta_{M\times 1} + \varepsilon_{N\times 1}$$

Equation 3.7: PRS-CS Bayesian linear regression equation.

Here N is sample size, M is the number of genetic markers, X is a genotype matrix, β is a vector of effect sizes for the genetic variants, and ε is the residual effect for the variants. In this instance though, to calculate the additive PRS, appropriate prior densities can be assigned to the regression coefficients β to control the effect size estimation. The PRS can be calculated using the posterior mean effect sizes.

The prior densities used for β can be expressed as mixtures of normal distributions with varying scales, described by Ge *et al*. as follows:

$$p(\beta_j) = \int N(0, \psi_j) dG(\psi_j), \qquad j = 1,2, \ldots, M,$$

Equation 3.8: PRS-CS Bayesian prior density calculation.

PRS-CS utilizes a continuous Bayesian shrinkage method to apply priors for each variant within the polygenic score. The shrinkage method takes into account the priors of the neighboring SNPs, thus accounting for the genetic architecture and linkage disequilibrium (LD) structure. This multivariate

model is then applied to weight effect sizes using only summary statistics and an external 1000 genomes reference panel with a sparse linear mixed model. Furthermore, PRS-CS implements an iterative procedure that enables efficient estimation of the hyperparameters, including the SNP weights and the parameters of the Bayesian hierarchical model. This approach allows for better prediction accuracy.

The following PRS-CS options were used: parameter A = 1, parameter B = 0.5, n_iter = 1000, n_burnin = 500, thin = 5, beta_std = False, seed = 1.

3.5.5 PRsice-2

In some instances, we used PRsice-2 software for the creation of PRS. PRsice-2 works by taking summary statistics GWAS and combining them with individual-level genotype data to create a PRS with the option of applying LD clumping and P-value pruning to the SNPs selected data to create a pruned set of SNPs. This pruned set of SNPs is then used to calculate the PRS for each individual, performing the basic PRS calculation of summed genotype multiplied by effect size.

For PRsice-2 usage, the following arguments were applied: --score sum --clump-kb 250 --clump-p 1.00000 --model add --fastscore --bar-levels 5e-8, 1e-5, 0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1 --thread 10 --device pdf --ld-type bed

As well as using PRsice-2 to calculate PRS, we also use the software for the final stages of our own PRS calculation pipeline, where the polygenic weight is calculated from the SNP β (using either LDpred2 or PRS-CS) and then PRsice-2 is used to calculate the PRS from this.

3.6 Mendelian randomisation (MR)

3.6.1 Multivariable MR using LASSO

For the PathWAS methodology, we conduct a multivariable Mendelian randomisation (MVMR) analysis to gain weighted estimates of the effect of each gene on pathway function. For this we utilised clumped significant eQTL SNPs as instrumental variables for the various genes, which are the exposures in the MR, regressed against a selected end-point protein as the outcome.

For each eQTL, instrumental variables (IVs) were selected from the significant ones (FDR < 0.05, as defined by eQTLgen) which were independent (r2 < 0.001). To do this in for multiple traits at once, we first merged the list of all SNPs significant for at least one of the required eQTLs. In case SNPs overlapped between multiple genes, the lowest p-value was assigned. Clumping was performed on the overall SNP list obtaining a single list of SNPs for which effect sizes (beta) and standard errors (SE)

were extracted. In case the SNP was not available for one of the genes, beta was set at $1 \times 10{-}7$ and SE at 1. Clumping was performed using the ieugwasr package (which is based on PLINK) and a reference set of 10,000 individuals of European descent from the UK Biobank for LD. These SNPs were then used to run the MVMR.

This is conducted using the MendelianRandomisation R package, using the mv_mrlasso function. This function applies lasso-type penalisation to the effects of the variants against the end-point by performing a multivariable inverse variance-weighted (IVW) MR, with the effects for each exposure being the post-lasso estimates.

3.6.2 Multivariable MR sensitivity analysis

As part of the Leave-One-Out sensitivity analyses conducted for confirmation of the PathWAS pathway-phenotype associations, it was necessary to remove individual gene associations from a pathway and then conduct a second MVMR for each newly created pathway list.

In some cases however, the original pathway only contained two genes connected to the end-point and thus removal of any one gene from the pathway model results in only one exposure being present for the MR. In these instances, standard IVW MR was conducted of the individual gene against the end-point.

3.7 Phenome-wide association study (PheWAS)

Multiple versions of association analysis were performed between UK Biobank phenotypes and different PRS created. The final results included association tests between phenotypes and PRS created for pathways ($PRS_{Pathway}$), for the end-point proteins ($PRS_{Protein}$) and for the leave-one-out analysis ($PRS_{Pathway\text{-}GeneN}$). We also experimented with association tests in ORCADES between phenotypes and PRS for individual genes (PRS_{Gene}), and also between the measured protein levels and the PRS_{Gene}.

These tests involved linear regression of the phenotype and the PRS along with several covariates. In a few instances (such as with the binary phenotype of diabetes) a logistic regression was used instead.

We performed linear and logistic regressions in R using the *speedglm* package and function, depending on whether the trait was quantitative or binary. The covariates included in the analysis were age, sex, batch, the genotyping array, genetic sex (rather than self-reported), and only the first unrelated individual from any familial group (based first on non-missing phenotypes and then on lowest individual identification). In addition, we included the first 40 principal components, and all continuous phenotypes were standardized using the scale function to enable phenotype comparison.

For lifespan and ageing-related traits (such as lifespan, mother's lifespan, and father's lifespan), we conducted Cox survival models using the *coxph* function from the survival R package. For this analysis, we used the same covariates as in the previous analysis, with the exception of excluding age and incorporating only the first 20 principal components instead of the first 40.

3.8 TreeWAS

When analysing the pathway PRS in chapter 5, Xue Li conducted a TreeWAS analysis using the provided $PRS_{Pathway}$ models. A TreeWAS applies a Bayesian analysis framework to PheWAS in order to improve the statistical power of detecting genotype-phenotype associations within the specific sub-phenotypes defined by the ICD-10 coding system. This method models the genetic coefficients across all phenotypes as a set of random variables, and utilizes a Markov process to model the correlations of the hierarchical tree-like structure of ICD-10 codes, referred to as a tree-structured phenotypic model. The tree structure is based on the classification hierarchy of ICD-10 coding system, where each node in the tree represents a clinical term in the classification. This uses methodology described by Cortes *et al.* (2017)[172].

3.9 Pathway databases

3.9.1 KEGG

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) database is an online resource available here: https://www.kegg.jp/kegg/. It provides information on pathways, reaction modules, genes, compounds and more and is a manually curated database of molecular pathways used in this project to define pathway interactions and gene lists.

In order to conduct Pathway-wide association studies (PathWAS), we utilized three R packages to analyse the KEGG database. The first package, KEGGREST, was used to identify pathways that are associated with specific genes. This involves searching for pathways based on a chosen end-point protein.

Next, we used KEGGgraph[173] to generate simplified pathway tables consisting of nodes and vertices, which ensured that the selected gene is an end-point and then created a list of genes connected to this end-point.

Finally, we used KEGGlincs[174] to validate the end-point by confirming that KEGGgraph did not erroneously select a gene that was part of a complex, which would have led to an incorrect labelling of the gene as an end-point.

3.9.2 Other pathway sources

At various points during this project I have attempted to use the Reactome (https://reactome.org/)[150] and WikiPathways (https://www.wikipathways.org/index.php/WikiPathways)[151] databases for pathway information. I have also utilised the FUMA-GWAS pathway enrichment platform (https://fuma.ctglab.nl/)[175].

Chapter 4

Results: Multi-tissue eQTL scores

4.1 Introduction

The broad approach of the PathWAS project and methodology is to use QTLs (specifically eQTLs) to create polygenic scores which can predict pathway functionality, based on weighting of the scores against a measured end-point protein. An important aspect of this then is to create the PRS which are the best at predicting the individual gene expression level. Thus, in turn those PRS which are most representative of, and those which best predict pathway function in an individual. In order to do this, it was important to use the sources of data and methodology which would allow the creation of the best pathway scores. As such, there were two major concerns which I intended to address, namely a potential lack of power and comparing methodologies.

4.1.1 The question of specificity and power

The first problem was the source of the data used for creation of the individual gene expression PRS ($PRS_{Gene)}$ used in the creation of pathway PRS. When considering which data to use to try and create PRS for pathways, one aspect of the problem was that of location-specificity of the expression data. It has become much more accepted recently that in examining phenotypes and disease it is often necessary to directly examine the tissues or even cells specifically involved with said phenotype. As differences between gene and protein expression levels are responsible for variation between cell types and tissues[176], it is likely true that the transcriptome or proteome of one particular organ may have absolutely no bearing on a phenotype while in another cell of the same individual, with the same genotype, it will be intrinsically linked[177–179]. Simply put, it may not be enough to study the broad genotype and omics-profiles of an individual but may be necessary to use cell-type specific populations in order to gain an accurate picture of the cause of phenotype.

Therefore, in order to create the best PRS_{Gene} the ideal source of data would be single-cell eQTLs to specifically examine pathways within individual cell types and gain cell or tissue-specific pathway models. This is particularly true when considering the creation of PRS for biological pathways ($PRS_{Pathway}$) given that, by their very nature, pathways are intracellular functions and the activity of a pathway in a muscle cell may be extremely different to the activity of the same pathway within an immune cell. Examples of this include disease pathways, such as COVID-19, where variable levels of protein expression has been shown across different tissue cell types, as such demonstrating variable

pathway activity and these cells also then display differences in infection levels by COVID-19[180]. Work by Zhang *et al.* (2020) has also shown variable levels of pathway expression by using single-cell RNA (scRNA) expression data[181].

Unfortunately, while such datasets are already available, they presently use small sample sizes and so may be underpowered for the purposes of my own project, although this may begin to change in coming years. A possible alternative to this is through the usage of tissue-specific QTLs. While not as ideal as single-cell sources, tissue-specific QTLs would provide an added layer of specificity and already there are numerous such data sources available.

In terms of specificity, another element to consider is the aforementioned concern of the poor correlating between mRNA and protein levels. As we theorise that protein levels better reflect phenotype than mRNA it follows that genetic variants linked to differences in protein abundance would more accurately [predict changes in pathway function than genetic variants linked to variation in RNA level. As such, a potential alternative to the use of eQTLs at all would have been the usage of pQTLs instead. However, while pQTLs would potentially be more reflective of final phenotype and thus pathway level, they also are susceptible to the same issues in terms of tissue and cell-specificity that I describe with eQTLs. Unlike eQTLs though, sc-pQTLs are not yet readily available.

As well as the question of specificity it is also important to consider the question of power. Despite the theory that by combining multiple PRS_{Gene} into $PRS_{Pathway}$ could improve statistical power of discovery, it is still preferable to incorporate larger expression datasets to maximise power. In terms of the eQTL data used, a dataset can be more powerful both in the sense of having a greater number of individuals sequenced, but also incorporating more gene mRNAs analysed with RNA-seq. The latter is particularly important in the context of PathWAS in order to have the most complete picture of a biological pathway by incorporating as many of its constituent genes as possible. This is where eQTLs hold a significant advantage over pQTLs. The SomaScan v4.1 assay assess over 7,000 individual protein levels and is one of the largest protein panels available to-date[182]. This is in comparison to some of the largest transcriptomics datasets which analyse the levels of tens of thousands of mRNAs[78].

4.1.2 The question of methodology

The second primary concern for creating PRS for pathways is in the methodology used to create the individual PRS_{Gene}. At their most basic, PRS are created using the following formula:

$$PRS_j = \sum_{ij}^{N} \beta_i \times G_{ij}$$

Equation 4.1: Standard PRS creation formula.

Where PRS are created for individual j, it is the cumulative value for each SNP (i) beta or effect (β) multiplied by genotype, in the form of dosage (G), with the effect sizes usually taken from GWAS for traits of interest. SNPs used in the creation of PRS should also be independent, as it is assumed that the cumulative score will not come PRS created in this way are generally trained in one dataset of a trait and then the model can be used to see how well it predicts the same trait in an additional independent dataset.

Over the past years, more sophisticated methods have been developed in the creation of PRS in order to improve their predictive ability, as this is one of their primary purposes and often indicative of quality. This can include usage of Bayesian shrinkage, penalized regression, incorporation of LD structure or usage of techniques like P-value thresholding and clumping. For each of these methods there also exists software which can be used to create PRS from summary statistics.

Therefore, a question which had to be addressed in the PathWAS project was which method of PRS creation would be best to use.

4.1.3 Creating the best PRS_{Gene}

In order to address the concerns of specificity and methodology I first sought to test the creation of PRS_{Gene} by generating PRS from a number of data sources and also testing several available methods. It was intended to define the best PRS_{Gene} based on those which could most accurately predict their associated protein level. Once this had been defined, I would then be able to start creating more PRS_{Gene} using the same method and begin the process of combining them into pathway scores.

4.1.4 Dealing with specificity

A potential issue with eQTLs is that it is known that they are often not predictors of the associated protein product levels, with eQTLs in general only poorly correlating with protein expression. I sought to address this issue by developing a PRS that could account for the possibility of protein transportation across various tissues, with the potential to improve the correlation between the PRS and actual protein levels.

The lack of correlation between mRNA and protein levels is indicative of the many levels of biological control between transcription and translation where a rise in mRNA levels will not always result in an equivalent rise in protein level (as described in **section 1.5.1**). While there are many possible explanations for this, one potential contributing factor is the transport of proteins between cells and tissues, where mRNA and their protein-products may be highly expressed in one group of cells, or in one tissue, but then the protein is exported and so the expression of the mRNA and protein

would be poorly associated if measured only within this tissue. Therefore, I intended to explore the hypothesis of whether by combining the eQTLs from several different tissue sources into one single overall score for an individual gene would improve the power of prediction and relationship with the associated protein product. Our theory was that PRS_{Gene} created from multiple tissue eQTL sources could feasibly better predict measured blood protein levels due to export of the protein from different tissues into the blood for transport around the body.

4.2 Results

4.2.1 Single-tissue PRS_{Gene} in GTEx

In order to first test the creation of PRS_{Gene} from eQTLs I utilised LDpred to create PRS from eQTLs from GTEx v0.7 data. As stated in **section 3.1.1**, GTEx v7 involved the use of 11,688 post-mortem RNA-seq samples from 714 donors across 53 tissues (post quality control this was 10,294 samples, 620 donors and 48 tissues). I sought to create expression PRS for each gene within the dataset, encapsulating over 20,000 mRNAs measured in each tissue, and see how well they would predict an independent measurement of the associated protein product. Here I used measurements of proteins from the ORCADES cohort.

A literature review of PRS creation methods revealed several different competing techniques could be used. These techniques included C+T[135], performed manually or by PRsice (PRS creation software which performs its own thresholding and clumping)[183]. However, several other methods have shown improvements over traditional clumping and thresholding methods by incorporating LD structure into the calculations such as lassosum[145], PrediXcan[147] and LDpred[139] (amongst others).

LDpred was developed by Vilhjálmsson *et al.* (2015)[139], in their study they compared LDpred with unadjusted PRS and those created with C+T. In their study they demonstrated that LDpred had better predictive power than the more simplistic PRS-creation methods, such as pruning and thresholding, due to this incorporation of LD structure into the model. Due to this improvement I decided to utilise this software as our method of creating PRS in favour of PRsice.

LDpred uses Bayesian shrinkage, estimating effect sizes for each SNP based on a provided reference panel for LD structure. The methodology of LDpred relies on the assumption that a given proportion (P_r to distinguish it from a P-value) of SNPs are causal.

I initially only used eQTLs from the GTEx Whole Blood data, to compare the predictive power of blood eQTLs against blood proteins (Olink proteomics are measured in serum). For the initial test I wanted to discover how well individual PRS_{Gene} created using LDpred actually predicted the relevant measured protein so that we could select the best value for P_r for usage in creation of future PRS_{Gene} for future assimilation into $PRS_{Pathway}$.

For this I created PRS using LDpred methodology as described in **section 3.4.1**. For GTEx Whole Blood this involved 280 genes which overlapped between the ~1000 proteins in the ORCADES proteomics with the GTEx expression data, and so creating a PRS_{Gene} for each individual in the ORCADES cohort (2,215 individuals, of which 1,048 had proteomics measured).

The results of this are shown in **Figure 4.1**, demonstrating the r^2 (or predictive ability) of each separate instance of the LDpred-created PRS_{Gene} regressed against their equivalent Olink protein measurement. This demonstrates that the best P_r threshold for use in LDpred was $1x10^{-3}$ (I.e., ~ 1 in 1000 SNPs were causal or $P_r = 0.001$). The fractions used in this experiment were the default fractions used in LDpred.

It is however worth noting that the mean predictive ability of each fraction of LDpred was substantially below 0.1 (with the highest mean r^2 for fraction p1-03 being 0.0273). This emphasises the issue of using eQTLs for the prediction of protein levels and may be due the poor correlation between transcriptomic and proteomic measurements for the same gene and equivalent gene-product.

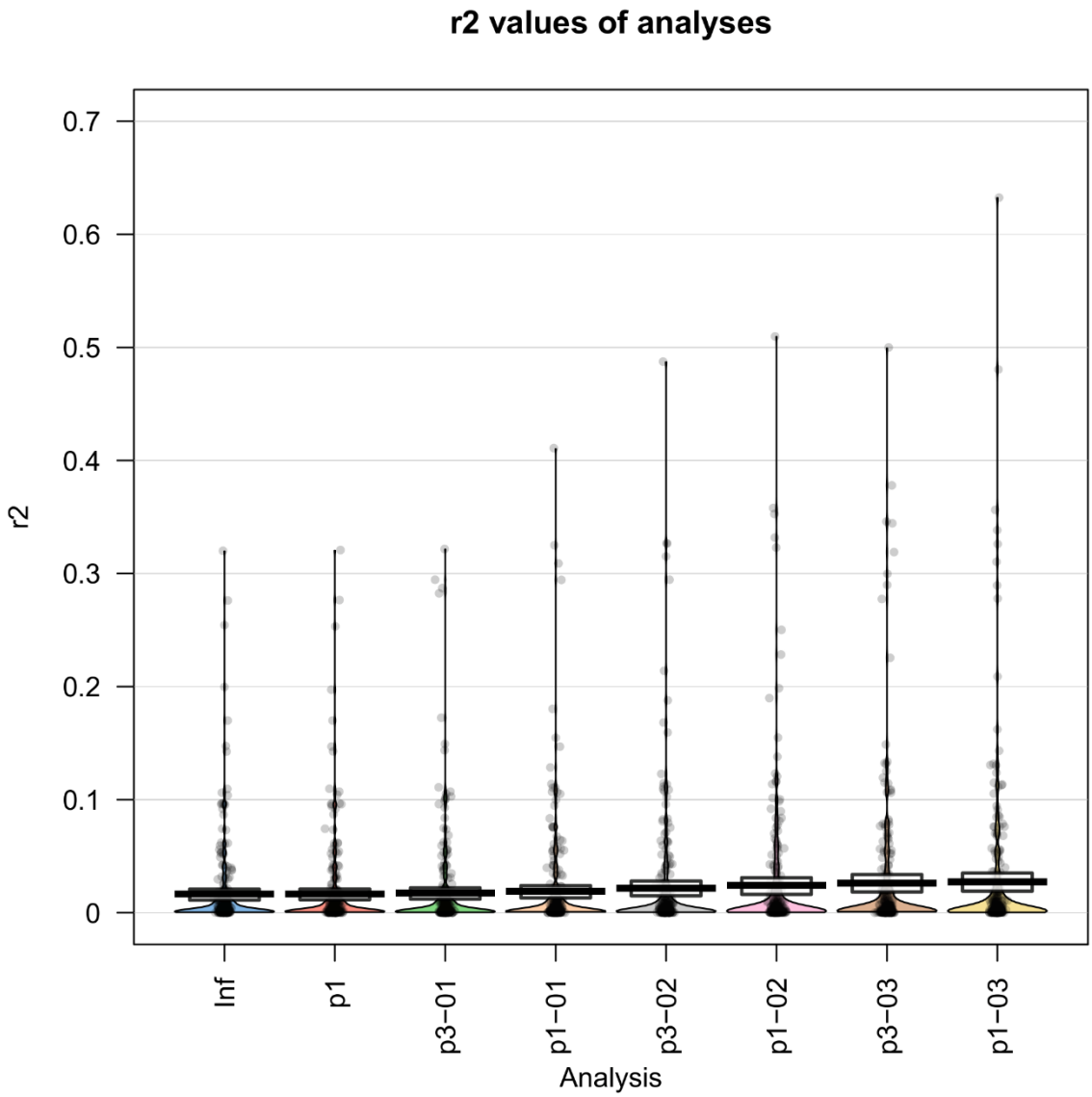


Figure 4.1: Pirate plot of LDpred default fractions. Each column contains the r^2 (predictive ability) for each model of LDpred. The PRS were created using the GTEx v.07 whole blood eQTLs and regressed against the equivalent measured protein in the ORCADES cohort, this accounted for 280 separate proteins. The "Inf" model assumes every variant is causal, each other model assumes a fraction or proportion (P_r) of causal variants (I.e. p1 assumes every instrument is causal, p1-03, P_r = 0.001, assumes 1 in 1000 variants are causal, p3-01 assumes a P_r of 0.3). The boxes and lines show the mean r^2 for each fraction. The values selected for P_r are those which are the default fractions suggested for LDpred.

4.2.2.1 Further examination of LDpred fractions

Given that the LDpred fractions seemed to indicate a general increase in r^2 at smaller and smaller P_r thresholds, a second run of the experiment was conducted including an additional P_r fraction at a much lower level to see if this would cause an increase in r^2. The results of this (**Figure 4.2**) did not show any improvement over the previously best selected fraction, and so using the P_r threshold of $1x10^{-3}$ was maintained.

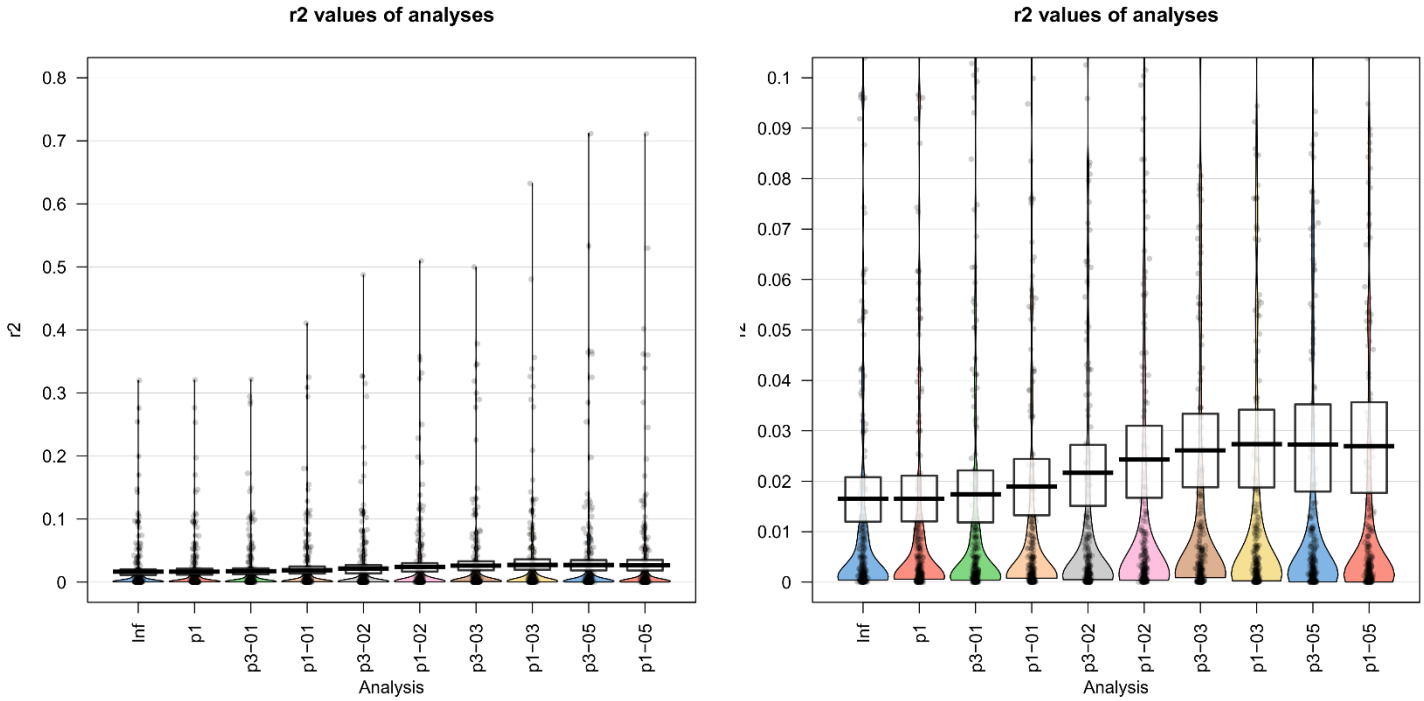


Figure 4.2: Pirate plot of LDpred expanded fractions. As in Figure 4.1 but with the non-default expanded p3-05 fraction. The right-hand panle shows a zoomed-in version of the plot, with an upper r^2 limit of 0.1. This demonstrates that while the expanded, non-default, fractions do push the outlying predictive ability of certain genes higher, the p1-03 fraction remained the fraction with a marginally higher mean r^2. Specifically fraction P_r = p1-03 had a mean r^2 of 0.0273 while fraction P_r = p1-05 had a mean r^2 of 0.027.

4.2.2 Comparison of PRS creation methods

While LDpred promised an improvement over traditional PRS creation methods, there were other competing methodologies available which also make use of LD structure in the creation of PRS. One such technology is PrediXcan. PrediXcan works by estimating the genetically regulated component of expression (defined by the Im lab as "GReX")[147], essentially using eQTL data and a reference dataset to "impute" expression-based polygenic weights for SNPs.

These weights can be used in place of SNP betas in the creation of PRS, providing a weighted estimate for each SNP. This is similar in practice to LDpred which also creates weighted betas for each SNP, based on the Gibbs sampling, which are then used in the standard PRS formula (**Equation 4.1**).

PrediXcan has been previously applied to every gene in the GTEx data set, compiled into a separate and publicly available data set called PredictDB, comprising of the weights for each gene (henceforth, discussion of the weights generated by PrediXcan will be referred to as PredictDB)[147]. These weights were then used to create PRS_{Gene} to compare the PredictDB-generated PRS with the LDpred-generated weights. As the method of PRS creation was the same following creation of the weights, this would be the important step in determining the best PRS-creation method. The best PRS was defined as those which could best predict the associated protein level in the ORCADES proteomics.

Of 1,068 unique protein measurements in the ORCADES dataset, using LDpred it was possible to create PRS_{Gene} for 280 of the associated genes from the GTEx whole blood data set. The PredictDB data for GTEx whole blood contained 264 of the genes within ORCADES. Between these two data sets there was an overlap of 198 genes.

A direct comparison of r^2 between the PredictDB-generated PRS_{Gene} and the LDpred-generated PRS_{Gene} showed that the PredictDB-based scores had a clear improvement in mean r^2 over those generated from LDpred (**Figure 4.3A**). Further, when the predictive ability of the two sets of PRS were compared only in the genes which overlapped between the two data sets then the PredictDB-generated PRS_{Gene} seemed to outperform those from LDpred (**Figure 4.3B**). However, this was only an extremely marginal improvement in both instances and also accounted for fewer genes from the GTEx data set.

Given the potential ambiguity from these data, with overall r^2 suggesting that PredictDB produced superior PRS while LDpred seemed to be capable of producing more PRS_{Gene} (with PredictDB potentially losing some of the lower-predictive expression scores), it is unclear which of the two methods was preferable for usage in creating PRS_{Gene}. However, as this was performed to test the utility of the associated technologies in creating PRS, one aspect of this was the computational time in the creation of each PRS_{Gene}. Here, LDpred provided a significant disadvantage, taking considerably more computational time than using PredictDB (as the time-consuming stage of calculating the weights had already been performed). For each gene, the process of calculating the LDpred-derived PRS from the GTEx summary statistics to the final scores in ORCADES could take upwards of 24 hours, while with PredictDB the PRS-creation step alone took less than an hour per gene. Given that there was no extremely obvious improvement in using either method, it was therefore decided for purposes of simplicity to use the PredictDB weights.

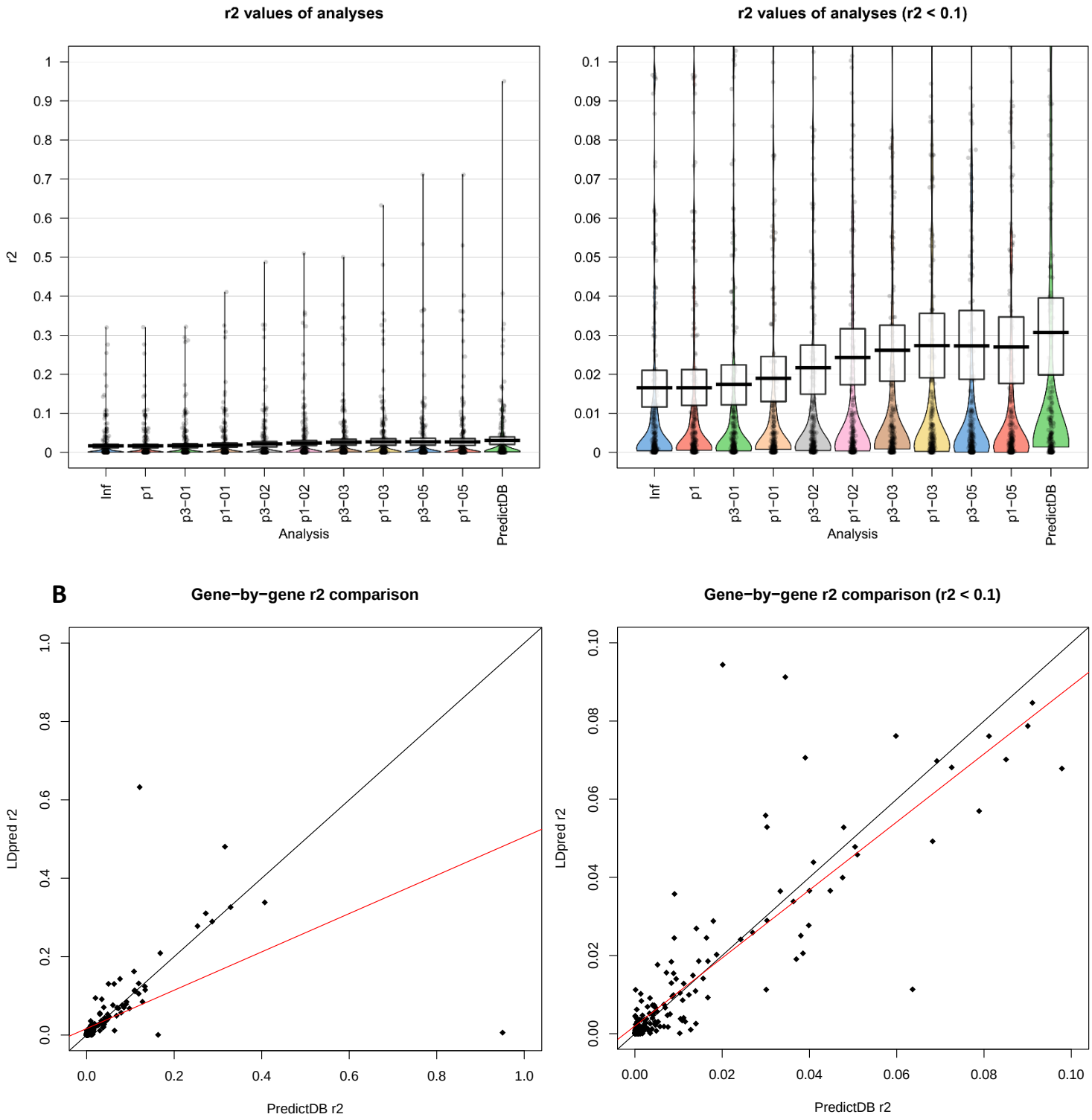


Figure 4.3. Comparison of LDpred-generated PRS and PredictDB-generated PRS. Panel A shows both the full and zoomed-in pirate plot of PRS created using differing LDpred fractions compared with the PRS created using the PredictDB weights. As can be seen, the mean predictive ability of the PredictDB scores is, again marginally, higher than that for the best LDpred fraction (p1-03). Specifically, the mean LDpred r^2 is 0.0273 and the mean PredictDb r^2 is 0.0307. Panel B shows a

gene-by-gene r^2 comparison of the PredictDB compared with the p1-03 LDpred fraction. The left panel shows all of the 198 genes which overlap between the data sets while the right panel shows the subset of those with an r^2 of less than 0.1 (accounting for 176 of the genes). In both plots there is a line of best fit in red, showing that when comparing the two methods directly in overlapping genes PredictDB produces PRS with superior predictive ability, but this superiority is drastically reduced when outliers are excluded.

4.2.3 Multi-tissue PRS_{Gene} in GTEx

The purpose in trying to create the best PRS_{Gene} is to be able to generate the most representative $PRS_{Pathway}$. One aspect of this with GTEx is the possibility of creating tissue-specific PRS_{Gene}, based on the idea that increased specificity would result in improved scores. In theory most pathways will have varying levels of activity depending on tissue (and even cell) type, so by increasing specificity of the source of the QTLs for the PRS_{Gene} it was theorised that this would improve the overall predictive ability of each score when combined into a $PRS_{Pathway}$.

However, I wished to examine the question of whether taking a broader view instead of a more specific view would improve prediction. As previously stated, one of the possible contributing factors in the discrepancy between transcriptomics and relevant protein level is protein transport and export, where a protein may be highly translated in one tissue (or cell) and exported elsewhere in the body. To account for this possibility, I intended to try and create individual tissue-specific PRS_{Gene} but then also compare these scores predictive ability with a multi-tissue PRS_{Gene}.

To create multi-tissue PRS_{Gene} I used K-fold cross validation (**Figure 4.4**) using 800 random individuals from the ORCADES proteomics dataset as a training set and then used the remaining 200 individuals as testing (from the proteomics sample of 1,048). The procedure was repeated five times (for k-5 cross-validation). We then compared the r^2 of the best individual tissue PRS_{Gene} with the multi-tissue PRS_{Gene}.

Of the 1,068 proteins in the ORCADES proteomics, 848 genes were present in PredictDB and of these a total of 710 genes could be used to create multi-tissue PRS_{Gene} for comparison. An initial examination of the comparison suggested that the multi-tissue PRS outperforms the predictive power of the tissue-specific PRS_{Gene}, with 54% of the genes having a higher r^2 from the multi-tissue score. However, of the 710 original genes, only 63 showed significantly better r^2 using the multi-tissue score compared to tissue specific (likelihood test, P-value $< 7.04 \times 10^{-5}$, **Figure 4.5**). A P-value threshold of 7.04×10^{-5} was used, correcting for multiple testing by dividing 0.05 by 710.

While this indicates that improvement in predictive power from the multi-tissue scores may at some point be possible, with increasing power and improving techniques, in practice these data showed that

less than 10% of the scores had demonstrable improvement. I.e., >90% of the proteins were either better predicted by individual tissue PRS_{Gene} or there was no difference. Because of this, for the purposes of the PathWAS methodology it was decided to focus only on individual tissue PRS_{Gene}.

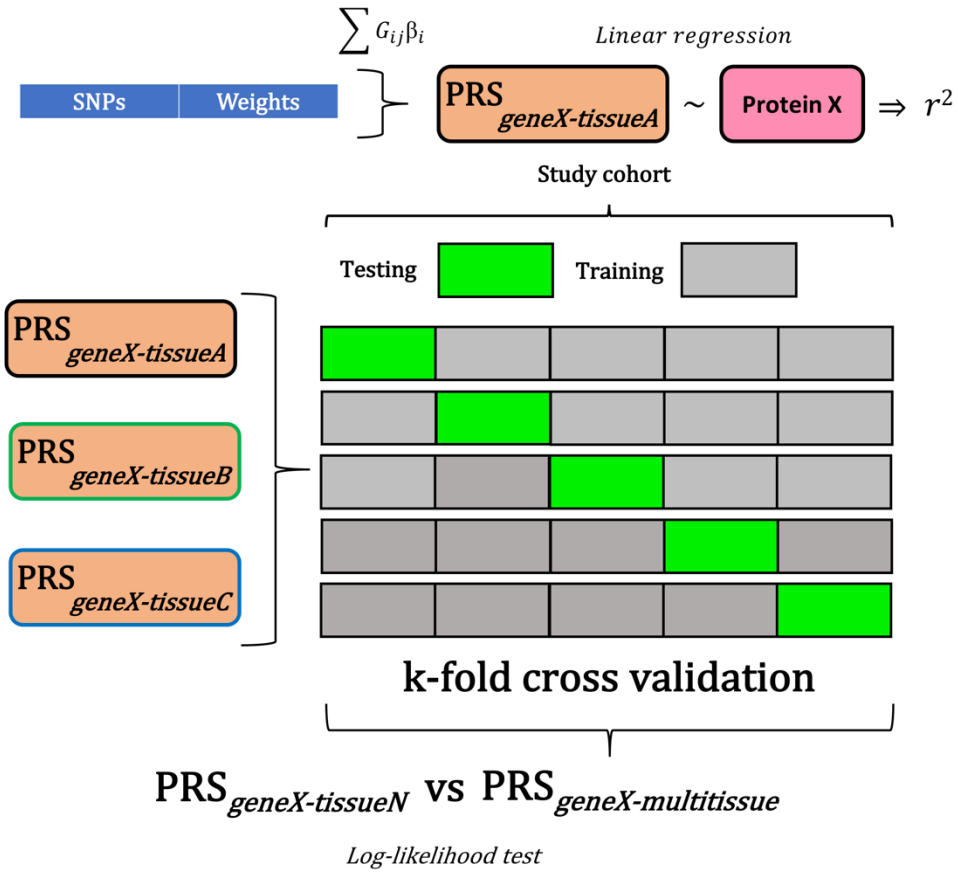


Figure 4.4. Methodology diagram of k-fold cross-validation. Depiction of the creation of the PRS for each gene (*geneX*) in different tissues, regressed against the measurement for the same protein (Protein X) to get the r^2. These PRS are then used in the cross-validation method within the ORCADES cohort where each individual is used as both a training and testing dataset. The PRS_{geneX} for specific tissues are then compared with those for the multi-tissue model.

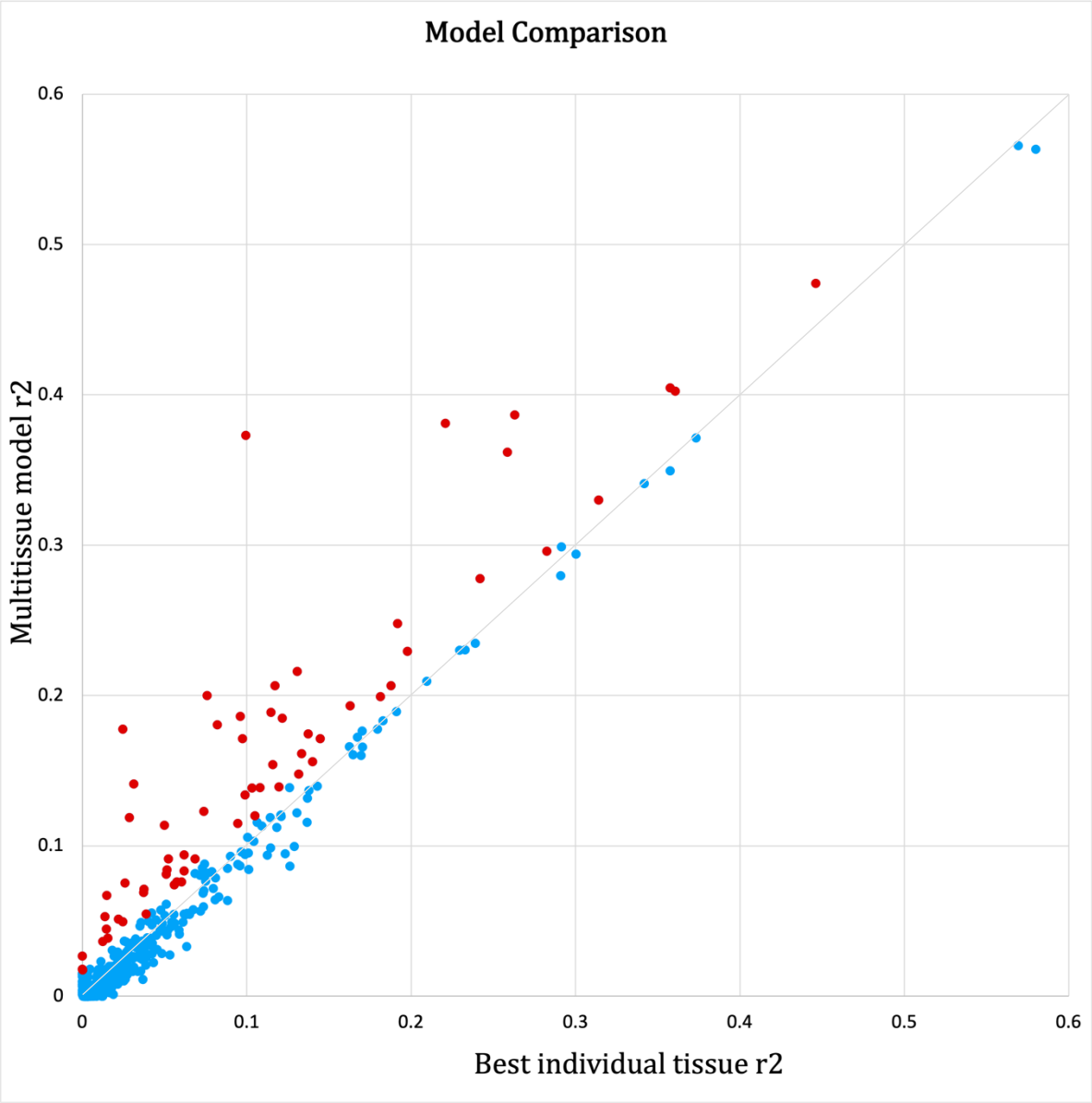


Figure 4.5. Comparison of Multi-tissue model with best tissue-specific PRS. A comparison of the best r^2 (predictive ability) from tissue-specific PRS_{Gene} created using PredictDB weights against the multi-tissue model PRS_{Gene}. The dots in red are those where the multi-tissue model was significantly better at predicting the equivalent protein, however a slight trend of improvement by the model can be seen as well.

4.2.3.1 Analysis of multi-tissue PRS_{Gene} improvement

While the multi-tissue models did not constitute an improvement in the creation of PRS_{Gene}, it was a source of curiosity why the scores for some genes did show some benefit. Several different pieces of online software were used to conduct gene-set enrichment analyses of the 63 genes which showed this significant improvement. Specifically, the FUMA-GWAS[175], ConsensusPathDB[184] and DAVID[185] websites were all utilised, as these allowed the use of a background list of genes for comparison.

These tools were selected because in many traditional pathway enrichment analysis software tools the "background" list will be every other gene in the human genome, however in this instance the genes were all selected based on the pre-existing ORCADES proteomics set. These proteins are pre-selected by the Olink company based on them being "interesting" biomarkers, and so enrichment conducted on a subset of this list would likely show enrichment for elements such as immune-response genes, as the Olink panels are already deliberately enriched with these.

Through use of the ORCADES proteomics as a background list, this still showed enrichment of immune response genes (ConsensusPathDB) and also signalling peptides (DAVID). This is intriguing as it suggests that the protein levels which are best predicted by the multi-tissue models may in fact be those which show greater levels of export from cells. This alone may be worthy of deeper future analysis.

Another interesting discovery was that in the case of the tissue-specific PRS_{Gene}, the best individual tissue at predicting the protein level was from the PredictDB whole blood weights 6% of the time, with the rest divided equally between the other tissues. This is in spite of the proteomics data being from blood serum.

4.2.4 ShinyApp creation

As part of the work conducted in the multi-tissue model analysis, I created an R ShinyApp to allow easy visualisation of the r^2 for each gene across each tissue where there were PredictDB weights available. An example output of this can be seen in **Figure 4.6**. This app allowed the rapid creation of bar plots for each gene showing the r^2 for each PRS_{Gene} compared between tissues and including the r^2 for the multi-tissue model.

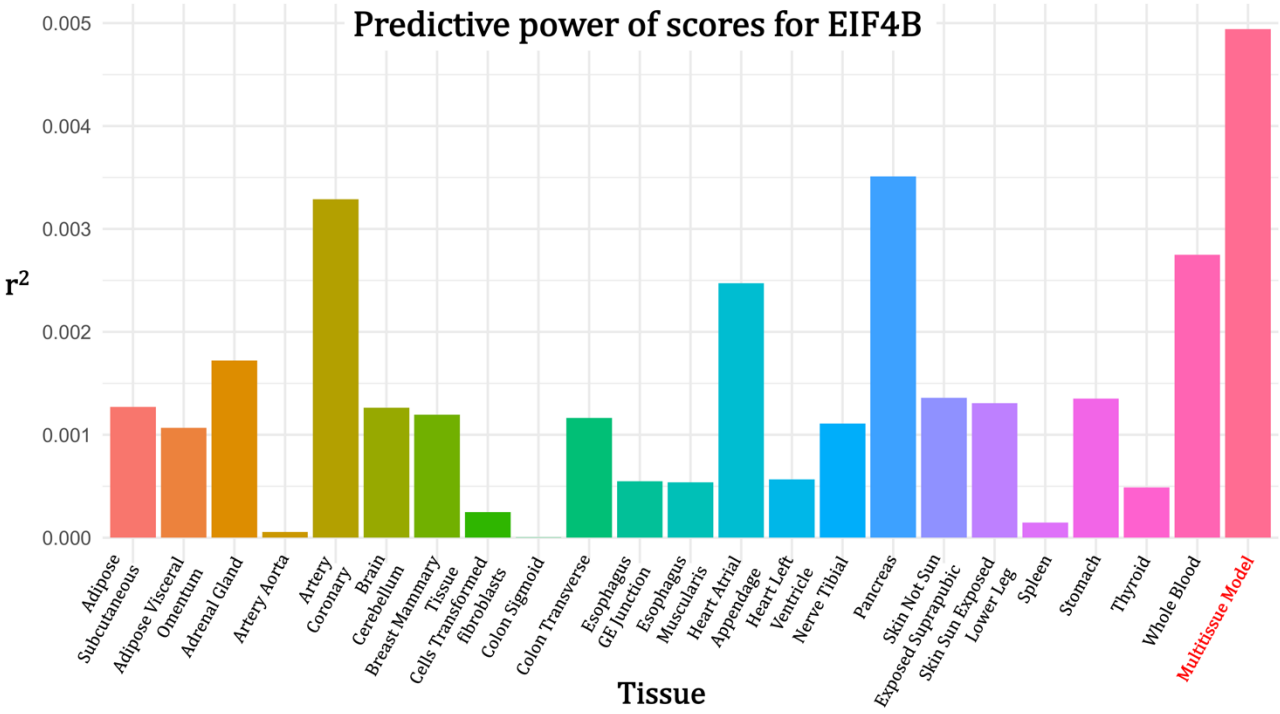


Figure 4.6. Example output from ShinyApp developed. This output was created for gene EIF4B with the r^2 compared between each tissue where there was a measurement for the gene in the PredictDB data set and then the multi-tissue model.

4.3 Conclusions

The initial aims of this early phase in the development of PathWAS was two-fold. First, following a literature-review of the available methods, it was intended to test a subset of the available PRS-creation methodologies in order to define the best, for later usage in creation of PRS_{Gene}. Secondly, it was attempted to compare the predictive abilities of PRS_{Gene} created from individual and multi-tissue eQTL sources, with the intention of examining whether multi-tissue scores would improve over tissue-specific scores. This was done in an attempt overcome some of the discrepancy between eQTL effect sizes and protein levels which poorly correlated in many instances.

This was performed with the broader goal of being able to produce the best PRS_{Gene} and therefore the best $PRS_{Pathway}$ with the PathWAS methodology, under the assumption that the expression PRS which would best predict protein level would also then be best at predicting phenotype when incorporated into a pathway score.

PRS_{Gene} were created with LDpred and through usage of PredictDB weighted SNPs, derived from the PrediXcan methodology, both using GTEx v7 eQTLs. In direct comparison the PredictDB-derived PRS appeared to slightly outperform the predictive ability of the LDpred-derived weights, both in

terms of mean predictive ability and overall. As such, given the computational concerns of LDpred, PredictDB was selected for the next stage of the analysis.

The field of PRS creation is a rapidly evolving one and as such there are many competing creation methods and it remains contentious about which methodology is best. LDpred is, however, regularly cited as one of the best tools in PRS creation[186,187] and as such, given that PredictDB appeared to match the predictive ability of it, it seemed like a valuable alternative.

I saw some limited improvement through usage of the multi-tissue model for PRS_{Gene} when compared with tissue-specific PRS_{Gene} (all derived from PredictDB). However, this improvement was only significant in a small subset of genes and so overall did not represent a worthwhile inclusion for the PathWAS pipeline, given that this meant there would not be any improvement in the majority of cases.

It is intriguing to note that where the multi-tissue model did however outperform the tissue-specific results, the PRS_{Gene} were enriched across several platforms for genes which are exported into the blood (immune response proteins and signalling peptides). This appears to suggest there is some validity to the usage of multi-tissue scores in specifically predicting exported proteins levels in blood. This in turn also may provide evidence for the export of proteins contributing to the poor correlation between eQTLs and protein levels in some capacity.

While these results are intriguing, exploring them further was beyond the scope of this project and so the primary conclusions from the early phase of the work was to focus on using PredictDB and tissue-specific PRS_{Gene} rather than LDpred or any form of multi-tissue scores.

Chapter 5

Results: PathWAS v.1.0

5.1 Introduction

After the development of the multi-tissue PRS for gene expression using genes from the Genotype-Tissue Expression (GTEx) dataset, several conclusions were drawn which were considered during the creation of the first version of PathWAS. While the GTEx dataset offered several advantages to use due to its tissue-specificity, and the PathWAS methodology would theoretically benefit from usage of tissue- or cell-type specific gene expression PRS, there were also some disadvantages with its inclusion.

A primary issue with GTEx is the relatively low sample size for each tissue, leading to a lack of power. With the advancement to GTEx v8, the largest individual sample size for a tissue in the dataset is skeletal muscle with a total of 803 samples, which is still relatively low. In addition to this, the majority of individual tissues have fewer than 500 samples. Moreover, due to the same lack of power in the GTEx dataset, it contains fewer associations with individual genes, an element which would be required for creation of pathway scores. Therefore, it was decided to employ a larger eQTL dataset to prioritise the method's available power over specificity.

It was also concluded that that the use of LDpred and PrediXCan showed little apparent difference in the development of polygenic weights. Consequently, since the PredictDB dataset was available for GTEx eQTLs, PrediXCan was utilized. However, given the shift in focus from GTEx to a different dataset, the applicability of the PredictDB data would no longer be relevant, and the methodology for creating the gene expression PRS would need to be reconsidered.

5.1.1 Using eQTLgen instead of GTEx for creating gene expression scores

In order to improve the power available to the PathWAS methodology it was decided to use an alternative set of eQTLs to the GTEx data. There are numerous options of eQTL data repositories available including ROSMAP[188], TwinsUK[189], GEUVADIS[190], GENCORD[191], BLUEPRINT[192] and more[193]. However, all of these, like GTEx, comprise of a similar and comparatively small sample size (with 411, 856, 462, 204 and 197 individuals used for RNA-sequencing respectively for those studies named). As such, we instead opted to use the eQTLgen dataset. It was selected primarily due to its very large sample size, as a meta-analysis consortium of eQTL studies it consists of 31,684 individuals[78] compared to the ~800 samples in GTEx v8.

The eQTLgen data contains the results of a meta-analysis of expression data from 37 cohorts which utilised whole blood and peripheral blood mononuclear cells (PBMC) . The individual cohorts provided expression data for eQTLgen using RNA-seq (20.3%) and expression arrays from Illumina (55%), Afymetrix U219 (8.7%) and Afymetrix Hu-Ex version 1.0 ST (16%)[78]. As such, while the eQTLgen consortium data lacks the tissue-specificity of the GTEx data, it had the dual benefit of an over ten-fold increase in sample size as well as having been drawn from a similar tissue to the proteomics used in the study. Specifically, the Olink proteomics data were derived from plasma samples while the transcriptomics were primarily from whole blood. Since the multi-tissue scores examined in **chapter 4** offered no improvement over tissue-specific expression PRS, maintaining similar tissue sources for the omics used was theorised to improve the capacity of the method to find associations for use in a proof-of-concept for the PathWAS methodology.

5.1.2 Testing additional QTL models

As was discussed in **chapter 4**, we attempted to determine whether multi-tissue PRS for expression levels showed any improvement over single-tissue PRS_{Gene}. This was done due to concerns over the potential limiting factor of eQTLs not corresponding well with their associated protein levels. As the theory behind the PathWAS methodology assumes that there is a cumulative effect of each gene in a pathway which combines into an overall pathway effect, this effect will be actioned directly by the proteins of the pathway, rather than indirectly from the mRNA. As such, PRS_{Gene} based on QTLs which best relate to protein level would be best for use in the PathWAS methodology.

Considering this, as well as creating PRS_{Gene} using eQTLgen, we also conducted an experiment to create PRS_{Gene} using pQTLs from the INTERVAL dataset. This is a cohort study with SomaLogic proteomics providing 3,283 protein measurements (described in **section 3.2.5**).

5.2 PathWAS approach version 1

The overarching aim of the project was to create polygenic scores to predict biological pathway function. Methods used initially to achieve this (and described in this chapter) were subsequently improved upon or replaced in the final iteration (described in **chapter 6**).

5.2.1 Proteomic data used

For the early versions of PathWAS we utilised two in-house sets of proteomic data from the ORCADES and Vis cohorts. These two cohorts comprised samples from ~1000 and ~800 people

(respectively), with the former used as a training dataset for the pathway models and the latter used to test (or predict) the models. Overlapping both cohorts, we had access to three panels of OLINK proteomics: CVD2, CVD3 (the Cardiovascular Disease 2 and 3 panels) and INF (the Inflammation panel). This provided a total of 261 proteins for use as end-point measurements.

5.2.2 Pathway genes and gene lists

Pathways were selected based upon availability of end-point proteins. As detailed in **section 2.4**, we required a proxy measurement for pathway function and for this were using proteomic data. As such, our pathways were limited to those for which we had an available end-point. With this in mind, we worked backwards from the 261 end-point protein measurements which overlapped between ORCADES and Vis so that we had a training and testing set.

The KEGG pathway database was used as the basis for all pathway data in the PathWAS methodology. It was possible to search the KEGG database using the KEGGREST R package for pathways in which the end-points were present[194]. These pathways could then be refined to only include those for which the protein was at the "end" of the pathway. Furthermore, the pathway was trimmed to only include genes which had an inward connection to the end-point (**Fig 5.1A**). This was performed using a mixture of the KEGGgraph[173] and igraph R packages. In practice this meant trimming down the broader pathway network to only include pathways which first contained one of the proteins, and then to only include genes in the extracted gene-list which were part of the sequence of gene-gene interactions which led to the selected final end-product. This was performed entirely automatically, to minimise manual curation and attempt to have a fully automatic pipeline which could be utilised by other researchers.

I extracted the KEGG pathways into simplified tables based on graph theory, with a table of nodes/vertices and directional edges. Here a vertex (or node) is any individual element within the graph or pathway, e.g. a gene, a molecule such as calcium, or a broader element of biology such as DNA or specific pathway interactions. An edge is then the connections between nodes and represent the relationships or interactions between them. The edges between two nodes can represent anything from physical connections such as protein-protein binding, to chemical reactions such as catalysis. It is also important that these edges included directionality, so it could be clear which genes were downstream and upstream from other in the chain of pathway reactions.

Within a given extracted pathway graph, the selected protein end-point had to have inward-facing edges, and no outward-facing edges. (I.e., at least one connection from any other element of the pathway and no connections from the end-point to another element). This would define it as an "end-point" for the pathway. As well as this, the end-point had to have at least 2 other genes connected

inwards, either in sequence or separately. This was done to ensure that the method would be creating cumulative pathway scores. With only one gene connected to the end-point this would no longer be a "pathway" and instead just an association test between an individual gene and a given protein. All of the genes kept for the overall pathway score had to eventually connect to the end-point (this was incorporated as many pathways in KEGG include multiple parallel and unconnected routes or pathway).

5.2.3 Combining GTEx with eQTLgen to retain tissue-specificity

Going forward, eQTLgen *cis*-eQTLs were used in place of eQTLs from GTEx for creation of the PRS_{Gene}. However, to attempt to retain an element of tissue-specificity, the GTEx transcript per million (TPM) files were incorporated into the methodology to create tissue-specific pathways. For this version of the method GTEx v.7 was used which includes expression data for 49 tissues.

To exemplify the process, a pathway would be selected with KEGGREST and converted into a directional graph of gene nodes connected to the final end-point using KEGGgraph. This directionality allowed the definition of genes at the end of a given pathway and also to differentiate between those genes which were connected upstream and those which existed within the same network but were not connected directly. Following this the TPM file was used to define whether a gene is "expressed" or not within each tissue. For this we used a value of 1 (transcript per million) as evidence of the presence of a gene being expressed. Then for each tissue, we attempted to create tissue-specific pathways based on the presence or absence of gene expression.

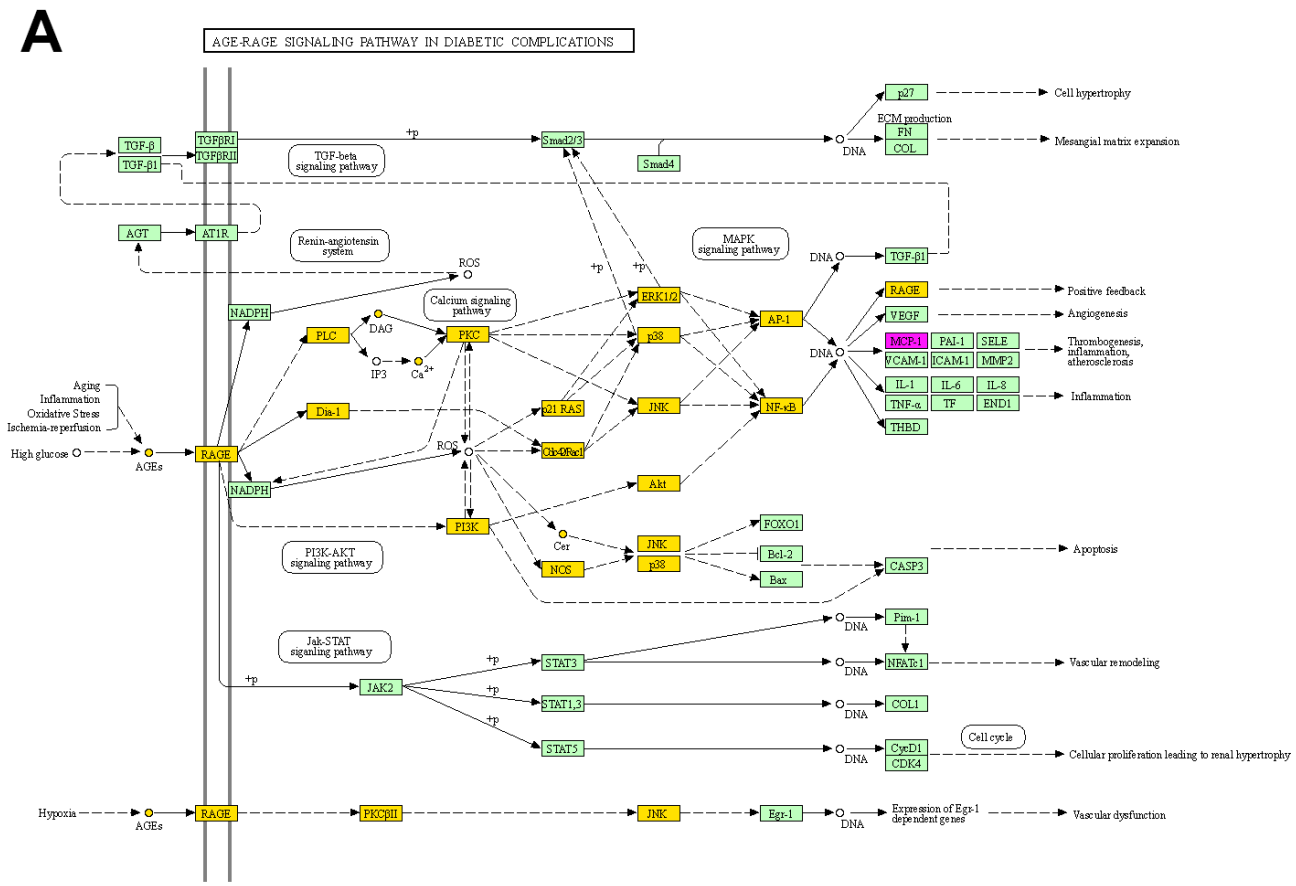
In order to do this, we created a matrix of simplified pathway "routes" throughout the overall pathway. Using the graph of nodes and edges, we first search for "starting" genes, I.e., nodes which only had outward edges and no inward edges (directly contrasting with the end-point genes). From these start genes we then searched to see if there was an eventual downstream connection with the end-point protein (to exclude start points which have no link to the end-point). From each start gene, linear routes were created within the pathway from each to the end-point, using the *all_smple_paths* function in the igraph R package. This allows the possibility of multiple chains from a given start point to end-point, as the start gene might connect to an end-point through multiple different linear routes.

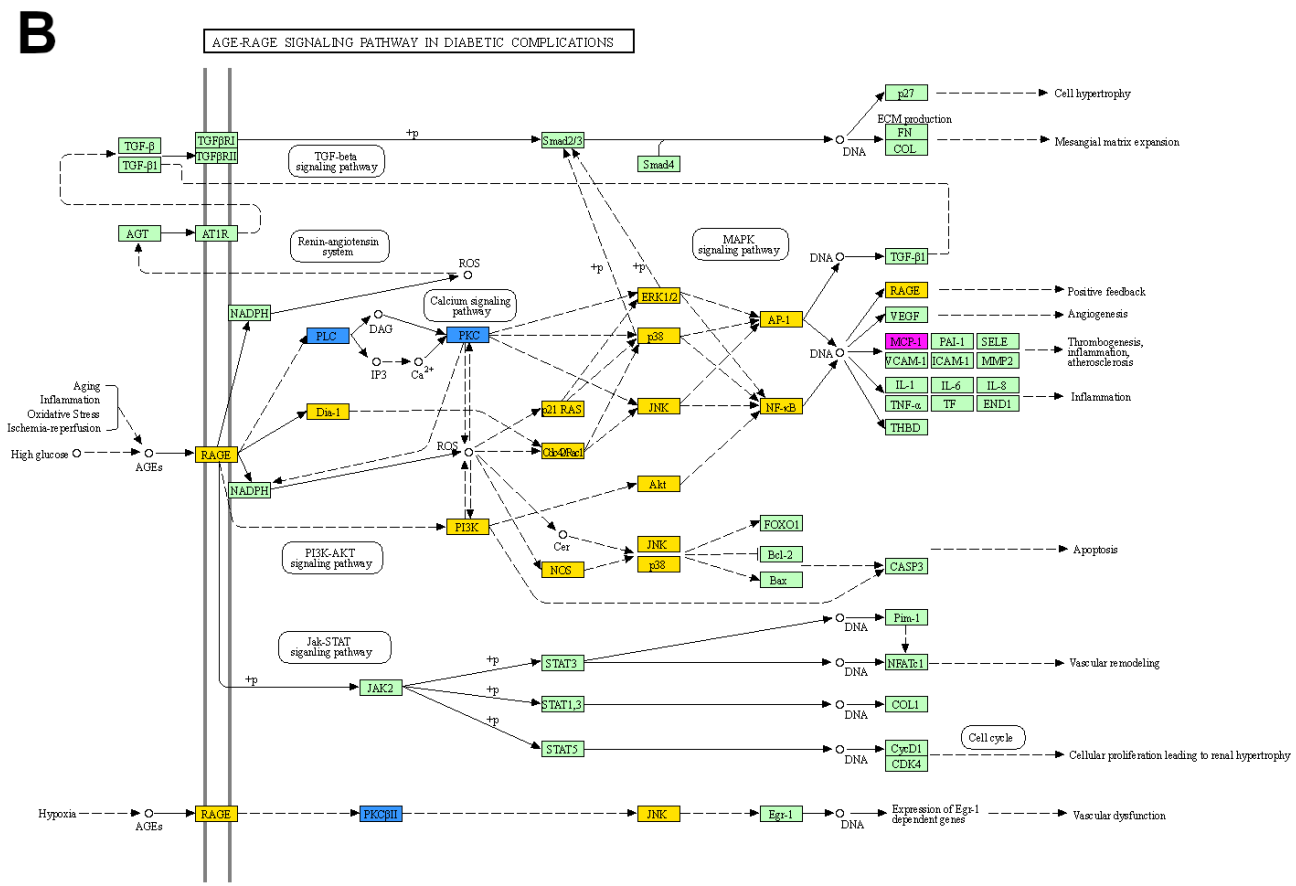
Each of these simplified directional routes were compiled into an encompassing matrix and then within each tissue the matrix was cross referenced with the TPM file and each gene within each simple pathway was classified as either being expressed or not, based upon the defined expression cut-off of 1. If any of the genes in the simple route was defined as unexpressed then this route was discarded, as we consider there to be no method of propagating the signal. The genes from the simple

pathways which were kept were then collated, providing a tissue-specific expression map for each pathway (**Fig 5.1B**). Due to the possibility of multiple routes from a given start gene to end-point, there may be redundancies within a pathway, and so even if one simple route is discarded, genes within it may still be included in the overall pathway for the tissue.

Figure 5.1: KEGG visualisation of AGE-RAGE signalling pathway with CCL2 as end-point. The figure depicts the AGE-RAGE signalling pathway derived from the KEGG database (hsa04933)., Genes are colour coded based on their relationship with a selected end-point from the ORCADES and Vis cohorts: Pink: CCL2 (end-point . labelled here as MCP-1[195]); Yellow: genes which have been extracted by the PathWAS methodology as being connected to the end-point; Green: genes which were removed by the method as having no inward connection to the end-point. Panel A shows the complete gene-list extracted by the PathWAS method for the CCL2 end-point, this also includes metabolites (such as Ca^{2+}) which are involved in the pathway chain. Panel B shows the tissue-specific pathway for Skeletal Muscle using the GTEx TPMs file, with genes in blue indicating those which did not pass the expression threshold for the method and would be excluded from the model. Here metabolites are also excluded as these are not included within the pathway PRS creation.

This figure was generated using the KEGG color site (https://www.genome.jp/kegg/mapper/color.html).





5.2.4 Using LDpred2

Given the switch from GTEx to eQTLgen we had to re-estimate the expression PRS. We had previously established LDpred to be one of the better methods for creating PRS and our need to re-create the PRS_{Gene} coincided with the development of LDpred2 by Privé *et al.* (2020)[140]. LDpred2 represented an improved version of LDpred, available as part of the R package *bigsnpr*[196]. In their report, Privé et al. demonstrated significant improvements through use of the LDpred2-auto function, both over LDpred1 and also PRS-CS, SBayesR and lassosum[140]. Thus, the newer LDpred2 potentially would not only outperform LDpred, but would also be substantially easier to apply due to its implementation in R. For the remainder of this chapter, all PRS_{Gene} are those created with LDpred2.

In order to create PRS_{Gene} with LDpred2, I created a pipeline for implementing LDpred2 on a given set of summary statistics. For the pipeline all of the SNPs for a given gene were extracted from the full eQTLgen dataset (usually between 5-12,000 SNPs) and the INTERVAL pQTLs. These SNPs were extracted from the relevant reference chromosomal bed file (using the custom UK Biobank reference panel, described in **section 3.4.3.1**) which was used to create a temporary genotype file. The betas for each SNP were aligned between the summary statistics and the reference allele in the genotype. This reference dataset was then used to create polygenic weights from the summary statistics using the *snp_ldpred2_auto* function as described in the LDpred2 tutorial[197]. Utilising this function also involved creating an estimation of SNP heritability for each gene by LDpred2 using LD score regression (LDSC). The weighted betas for each SNP were then used with the *big_prodVec* function to create PRS_{Gene} in UK Biobank, ORCADES and Vis from the full genotype files.

5.2.5 PathWAS models

With PRS_{Gene} created for every gene using LDpred2, next I used them to create PRS for each pathway (utilising the tissue-specific pathways created based on the GTEx TPM file).

To do this, we used GWAS of the proteomics end-points, as described in the rationale, as a proxy for pathway function. We then conducted a multivariable linear regression of the proteomics measurements against each of the combined PRS_{Gene} for the pathway. In short, for every individual the protein measurement was regressed, using the R *lm* function, against every PRS_{Gene} for the pathway. As well as providing a summed estimate of effect of all the PRS_{Gene} against the measured protein, the *lm* function also provides coefficients for each individual PRS_{Gene} within the pathway. The coefficients are used then as estimates of effect, or the weighted effect for the gene on the pathway.

The resulting model was trained utilising the ORCADES proteomics measurements, and then the coefficients of the model were used to predict their respective PRS_{Gene} in the Vis testing set. In other words, the effect of the PRS_{Gene} in ORCADES on the end-point protein was used to predict the

equivalent PRS_{Gene} in Vis, using the *predict* R function. This provided a total combined prediction value for each individual in Vis and this prediction value was regressed against the same end-point protein in Vis, thus providing an estimate of how well the pathway model in ORCADES could predict the measured protein in Vis and, as such, an estimate of the strength of the model.

As we wished to compare the difference between using linear and penalised regression, we also used elastic net from the Caret R package to create a second test and prediction set of values. While we hypothesised that incorporation of many PRS_{Gene} into the regression would improve the accuracy of the pathway model, it was possible that it would result in excessive noise from genes which had very little influence over the end-point protein. As such, through usage of elastic net penalised regression, the coefficients of the pathway would be pruned to include only those relevant to the end-point protein. This was done using the *train* function from the Caret package using the ORCADES PRS_{Gene} with the ORCADES end-point protein measurement, as well as trControl value of 10 (I.e., performing k-10 cross validation). The coefficients from this model were then used for the prediction with the Vis PRS_{Gene}, now using a "response" type of model and incorporating the λ for the elastic net model in ORCADES. This prediction value was then regressed against the Vis protein measurement as with the linear model to provide an overall estimate of effect and significance.

The entire process was conducted for every pathway in each tissue where a gene list was present. Those pathways which then could successfully predict the measured protein in Vis, either by linear regression or elastic net, were those which were carried forward for creation of the $PRS_{Pathway}$ which would be used in a PheWAS with UK Biobank.

5.2.6 PathWAS PheWAS

The significant pathway models are then used in a PheWAS against 44 traits in UK Biobank (**Table 3.3**), this included blood cell counts, anthropometry traits (such as height and BMI) illness and aging traits, lung function traits and bone density. These traits were selected for in order to test a spread of associations with the PathWAS methodology, including broad omnigenic traits such as height and BMI. We also included bone mineral density traits in order to test for known relationships between pathways and bone density. The blood cell count were selected due to their potentially close relationship with the blood proteins and transcriptomics. This set of 44 traits would be later expanded to 60 in **Chapter 6** to include more illness, death and metabolite measurements.

In order to do this the model coefficients for each gene were used in the creation of the $PRS_{Pathway}$. This involved weighting the PRS_{Gene} for each individual in UK Biobank by its respective linear regression coefficient, which is used as an estimate of the effect of the gene on the pathway. The weighted PRS_{Gene} were then summed into one overall $PRS_{Pathway}$ for each individual in UK Biobank.

Each $PRS_{Pathway}$ was then used in an association test as described in **section 3.6** against the 44 traits, a subset of the 60 traits from **Table 3.2**.

5.3 PathWAS v.1 test results

5.3.1 Pathway gene-lists

From the 261 proteins across the three overlapping proteomics panels between the ORCADES and Vis datasets it was possible to extract a total of 377 pathway-protein combinations, where there was a pathway in the KEGG dataset which had one of the proteins in the proteomics as an end-point. This resulted in a combined 107 unique KEGG pathways with 112 unique protein end-points, comprised of a total of 4,250 genes across each pathway for which a PRS_{Gene} had to be created using LDpred2 in ORCADES, Vis and UK Biobank.

Of the 4,250 genes across each pathway, 3,088 genes overlapped with eQTLs in the eQTLgen dataset, and so LDpred2 was applied to these. It was also not possible to create LDpred2 weights for 55 of the overlapping genes due to low h^2 estimates from the LDpred2 algorithm, and as such PRS_{Gene} could not be created in any dataset for these genes. As well as this, it was only possible to create PRS_{Gene} for 2,922 of the genes in Vis, potentially due to the smaller sample size of the cohort. As such, for the purposes of the prediction stage, the subsequent model creation phase of the process for each pathway was limited to only those PRS_{Gene} which overlapped between ORCADES and Vis.

5.3.1.1 INTERVAL pathway gene-lists and models

As well as creating PRS_{Gene} from the eQTLgen dataset, PRS_{Gene} were also created using the INTERVAL pQTLs using LDpred2. However, only 200 genes overlapped between all of the pathway genes and the ~3000 proteins measured by INTERVAL, and of these it was only possible to create LDpred2 weights for 189.

Due to this substantial lack of overlap between the pathways and the INTERVAL pQTLs, it ended up only being possible to predict models for 3 of the pathways in Vis. In a subsequent direct comparison between the INTERVAL models and eQTLgen models as well, for those three which overlapped, there was no consistent improvement in r^2, P-value or pathway model effect from using the pQTLs instead of the eQTLs. Given that the lack of pQTLs resulted in a corresponding difficulty in creation of $PRS_{Pathway}$, and that the few pQTL-derived models available showed no significant improvement, this aspect of the project was not taken further.

5.3.2 ORCADES and Vis PathWAS

From 377 unique pathway-protein combinations it was possible to create a PathWAS model for the pathway in ORCADES and predict this model in Vis in 221 instances in at least one tissue of 49 available. This was now refined to 4,940 models across 78 different pathways.

These results were then further pruned to only include pathways for which we were able to create both a linear regression model and an elastic net model, resulting in a final set of 1,536 models across 49 tissues and 187 pathway-endpoint combinations (and 74 unique pathways).

Of the 1,529 models created with both linear and elastic net regression, 5 elastic net models were significant and 5 linear models were significant. With significance defined by a Bonferroni corrected P-value of 0.05 divided by the number of unique pathways (74) giving a P-value threshold of 6.75×10^{-4}. Of these 5 significant results in each model, 3 overlapped between the two (**Figure 5.2**, **Table 5.1**). Given the equivalent number of significant results generated by each method of regression, it was not possible to come to a definitive conclusion about the superiority of either method.

Of the models which were significant, they were primarily involved in immune responses including response to infectious diseases like Shigellosis, Salmonella, and Legionellosis. Also NOD-like receptor signalling, and MAPK signalling pathways showed significance.

Of the 7 models that were significant, 5 have the end point protein Interleukin-18 (IL18). IL18 is a pro-inflammatory cytokine and is involved in responses to infection by salmonella, shigella and legionella and consequently is a key factor in these infection pathways. Two further IL18 models highlight the NOD-like receptor pathway, which plays an important role in innate immune responses. Another end-point protein, CCL3 (Chemokine C-C motif ligand 3) is a pro-inflammatory chemokine and here identified relation to the chagas disease pathway. The final end-point protein in the list, HSPB1, is a heat shock protein that is induced upon environmental stress and here is identified through the MAPK pathway.

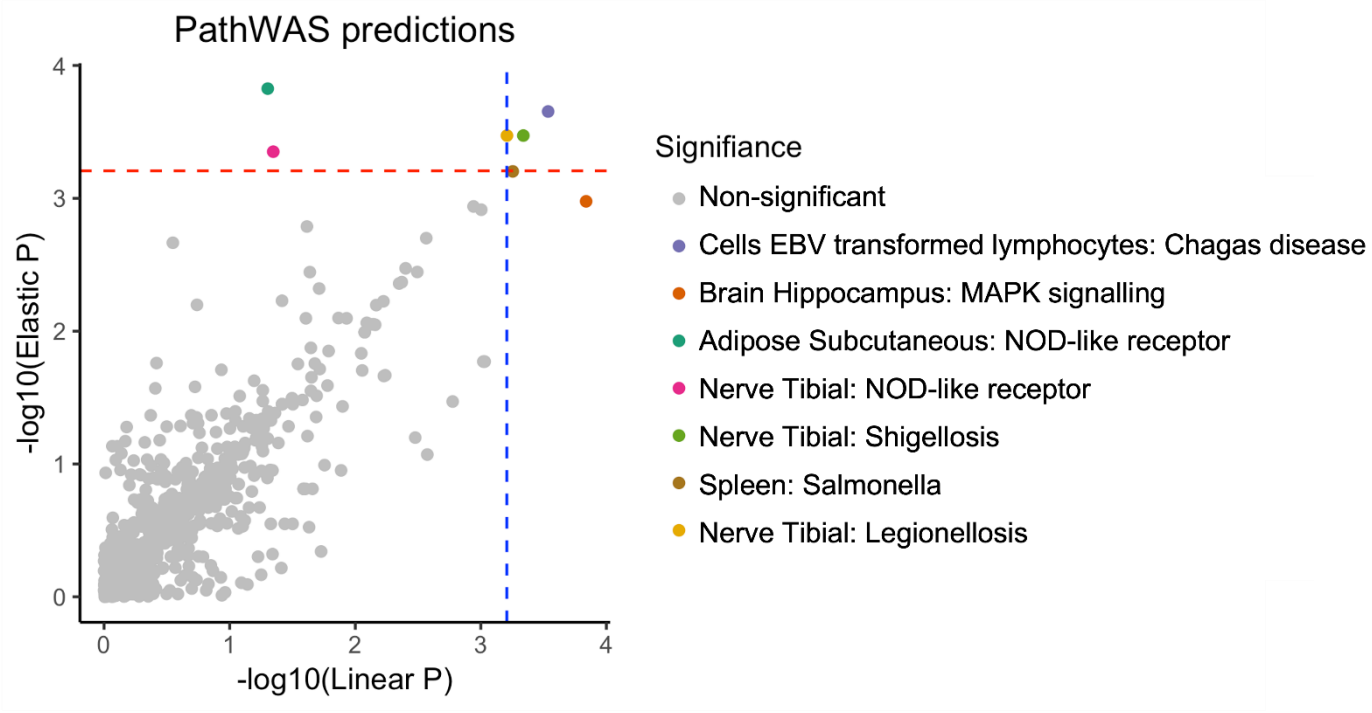


Figure 5.2. PathWAS model predictions in ORCADES and Vis. Shown are all 1,529 model P-values across both linear and elastic net regressions. The dots in grey are those models for which neither form of regression was significant, while those coloured show the pathways and tissues for which there was a significant model (as detailed in **table 5.1**).

End-point protein	KEGG Pathway ID	Pathway	Tissue	Linear P-value	Linear beta	Linear r^2	Elastic net P-value	Elastic net beta	Elastic net r^2
CCL3	hsa05142	Chagas disease	EBV-transformed lymphocytes	2.91×10^{-4}	1.03	1.45×10^{-2}	2.22×10^{-4}	1.44	1.50×10^{-2}
HSPB1	hsa04010	MAPK signalling pathway	Brain hippocampus	1.45×10^{-4}	2.23	1.59×10^{-2}	1.05×10^{-3}	3.35	1.19×10^{-2}
IL18	hsa04621	NOD-like receptor signalling pathway	Adipose subcutaneous	4.98×10^{-2}	0.24	4.27×10^{-3}	1.49×10^{-4}	0.77	1.59×10^{-2}
IL18	hsa04621	NOD-like receptor signalling pathway	Tibial nerve	4.50×10^{-2}	0.26	4.46×10^{-3}	4.46×10^{-4}	0.84	1.36×10^{-2}
IL18	hsa05131	Shigellosis	Tibial nerve	4.59×10^{-4}	0.62	1.36×10^{-2}	3.37×10^{-4}	0.73	1.42×10^{-2}
IL18	hsa05132	Salmonella	Spleen	5.57×10^{-4}	0.66	1.32×10^{-2}	6.27×10^{-4}	0.74	1.29×10^{-2}
IL18	hsa05134	Legionellosis	Tibial nerve	6.19×10^{-4}	0.61	1.29×10^{-2}	3.37×10^{-4}	0.72	1.42×10^{-2}

Table 5.1. Significant PathWAS v1 pathways predicted in Vis from ORCADES.

It is worth noting that pathways such as “Shigellosis” and “Legionellosis” in KEGG are pathways of immune response to infection and as such often consist of many overlapping elements and also include elements from other broader networks (including subsets of the MAPK and NOD-like receptor pathways). For example, in the KEGG shigellosis pathway, IL18 exists as an end-point as part of a subset of NOD-like receptor signalling which exists within the immune response to the shigella bacterium.

The tissues in which we discover these significant models is particularly curious. While the discovery of significant immune-response pathway models in the spleen makes sense given the spleen’s involvement with immunity[198,199], it is unexpected and intriguing to discover several strong immune response signals in the tibial nerve tissue.

It is plausible that this overrepresentation is to do with the source of both the proteomics and transcriptomics. As all the datasets involved are taken from blood, it is possible they would have notably higher concentrations of immune-response proteins and signalling peptides. However, the specific overrepresentation of IL18 amongst the end-points is intriguing particularly as a number of other interleukins (such as IL1A, IL6, IL2 and more) were present among the models as end-points for various pathways. The overrepresentation of immune-response proteins would also correlate with the results from the multi-tissue analysis, where those proteins most exported from different tissues like immune-response proteins and signalling peptides would potentially have the closest relationship to tissue-specific pathways.

5.3.3 ORCADES and Vis PathWAS PheWAS

The 7 pathway models (both linear and elastic net) which could predict the end-point protein in Vis were then used in the PheWAS analysis in UK Biobank as described. This finally resulted in 12 significant pathway-phenotype associations (**Figure 5.3**, **Table 5.2**), where significance was defined as 0.05 Bonferroni corrected by the 308 PheWAS tests made ($P < 1.62 \times 10^{-4}$). Five of the seven pathways were significantly associated with lymphocyte count, NOD-like receptor signalling with IL18 as end-point in both adipose and nerve tissue, chagas disease with CCL3 as an end-point and both shigellosis and salmonella with IL18 as the end-point. There were also additional associations between the chagas disease pathway and BMI, erythrocyte count, hip circumference, waist circumference and weight. Also, there were additional associations between the NOD-like signalling pathway in adipose tissue and both platelet count and height.

In this instance CCL3 exists as the end-point for a complex series of interactions within the chagas disease pathway (**Fig 5.4**) incorporating elements of MAPK-signalling, with expression of CCL3 itself driven by the JUN and FOS transcription factors. The axis of its expression is also downstream of toll-like receptors 2 and 6 (TLR2, TLR6) and is also downstream of inhibition by protein phosphatase 2A (PP2A) mediated by NFκB. As such, the expression of CCL3 exists at the end of a complex series of interactions involving numerous important genes. Therefore, it is perhaps unsurprising that the $PRS_{Pathway}$ derived from the model for this demonstrates some of the most significant associations seen in the experiment.

CCL3 itself has been extensively studied as a negative regulator of the proliferation of hematopoietic stem cells[200] as such if increased levels of the pathway leads to increased expression of CCL3 we would expect to see a negative association with blood cell counts, and we see a negative association with both lymphocyte and leukocyte counts. CCL3 has also been previous associated with obesity in both humans[201] and mice[202]. MAPK signalling has also been previously implicated in adipocyte differentiation and obesity[203], and so may be mediating part of its effect through CCL3. As such the positive relationship discovered between this pathway-axis and BMI and weight is also seemingly supported by existing literature.

The negative relationships between the other immune pathways with IL18 as the end-point and lymphocyte count is, however, unexpected. IL18 is a pro-inflammatory cytokine released by numerous cells, but often most predominantly macrophages and T-cells[204], and as such it is perhaps expected to have a positive association with white blood cell counts. It is possible that the pathway axis for the expression of IL18 may be involved in a form of negative feedback loop with lymphocyte expression, with higher levels of white blood cells leading to increased expression of the pathway and therefore increased levels of IL18, resulting in a subsequent downregulation of white blood cell proliferation.

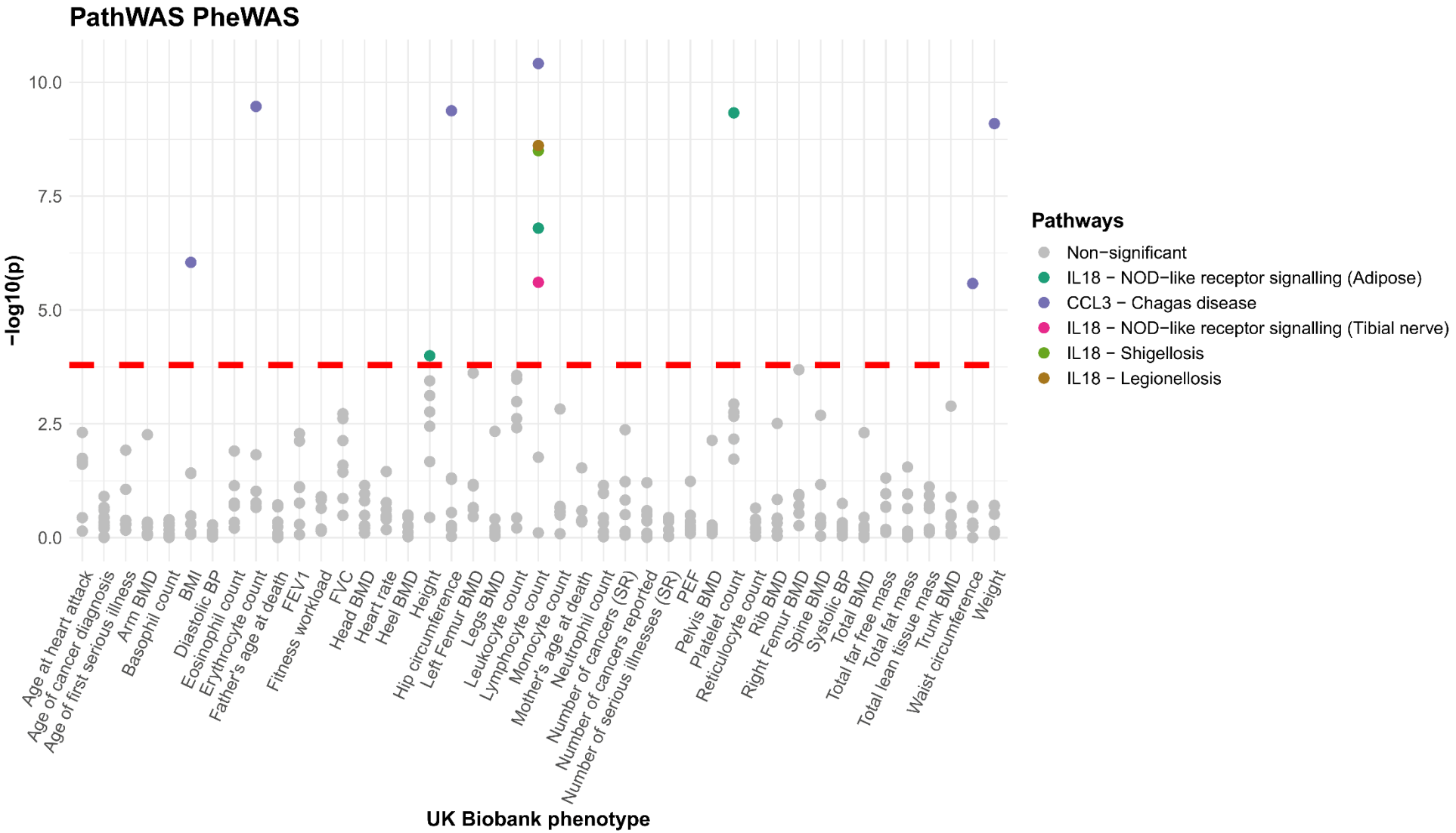


Figure 5.3. PathWAS v1 PheWAS. Shown are the significance (P-value) scores of each pathway from the initial version of PathWAS tested against 44 traits from UK Biobank across the X-axis.

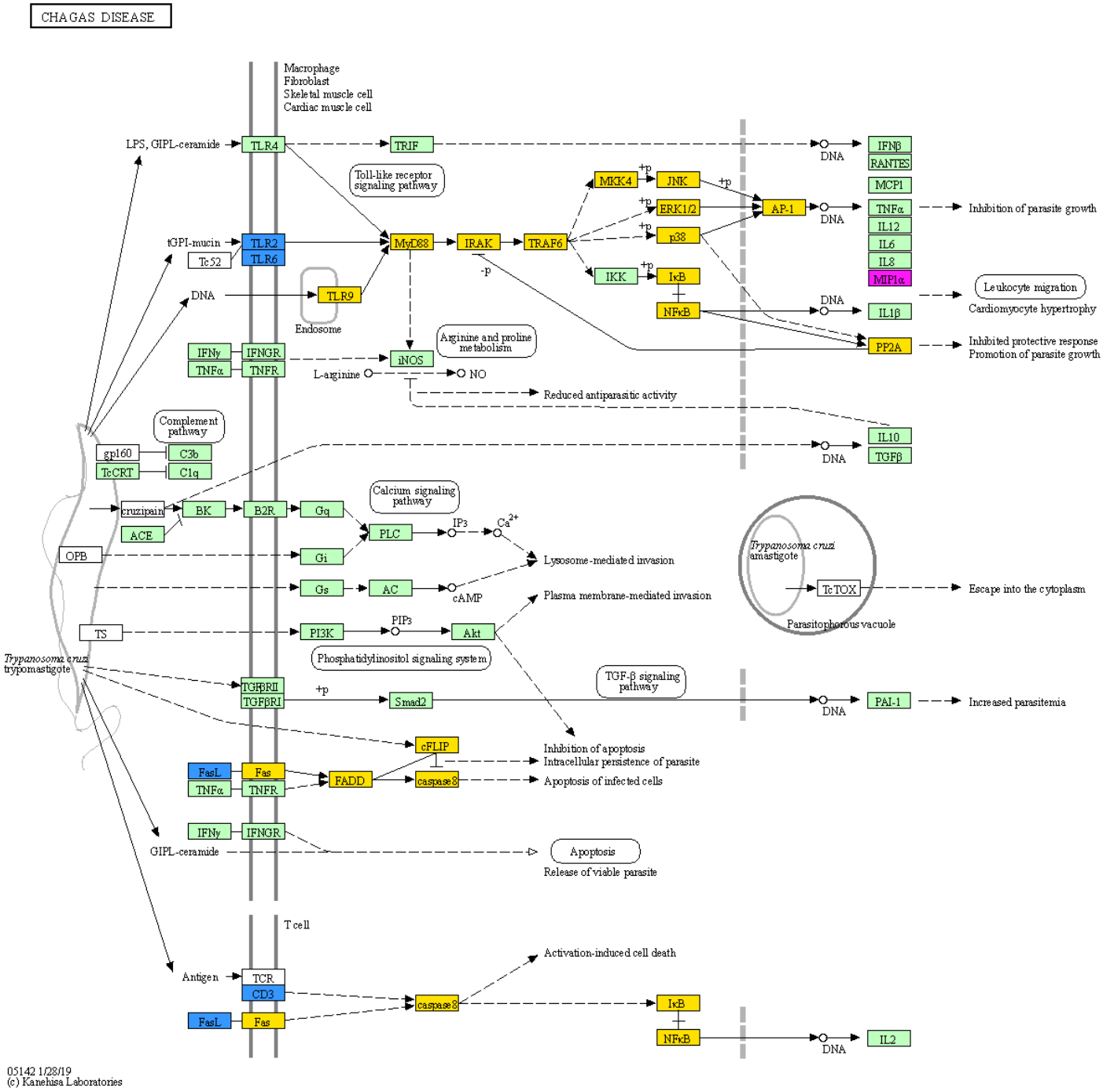


Figure 5.4. KEGG chagas disease pathway in EBV-transformed lymphocytes with CCL3 as end-point. Pink: CCL3 (end-point . labelled here as MIP1α); Yellow: genes which have been extracted by the PathWAS methodology as being connected to the end-point; Green: genes which were removed by the method as having no inward connection to the end-point; Blue: genes which are connected to the end-point but were deemed to be not expressed within the tissue as per the GTEx TPMs.

Pathway	Phenotype	Sample size	Beta	SE	Z-score	P-value	r^2
CCL3 – Chagas disease	Lymphocyte count	311493	-0.107	0.016	-6.610	3.86×10^{-11}	1.34×10^{-4}
CCL3 – Chagas disease	Waist circumference	321609	1.502	0.320	4.698	2.63×10^{-6}	1.42×10^{-5}
CCL3 – Chagas disease	Hip circumference	319772	1.451	0.232	6.246	4.21×10^{-10}	1.26×10^{-4}
CCL3 – Chagas disease	Weight	320302	2.265	0.369	6.144	8.04×10^{-10}	6.19×10^{-5}
CCL3 – Chagas disease	BMI	319557	0.597	0.122	4.912	9.01×10^{-7}	7.92×10^{-5}
CCL3 – Chagas disease	Erythrocyte count	313135	-0.060	0.010	-6.280	3.39×10^{-10}	2.33×10^{-5}
IL18 – NOD-like receptor signalling (Adipose)	Lymphocyte count	311493	-0.031	0.006	-5.240	1.60×10^{-7}	8.23×10^{-5}
IL18 – NOD-like receptor signalling (Adipose)	Height	323090	0.253	0.065	3.889	1.01×10^{-4}	-2.92×10^{-5}
IL18 – NOD-like receptor signalling (Adipose)	Platelet count	312001	-3.520	0.566	-6.230	4.68×10^{-10}	9.03×10^{-5}
IL18 – NOD-like receptor signalling (Tibial nerve)	Lymphocyte count	311493	-0.028	0.006	-4.711	2.47×10^{-6}	6.57×10^{-5}
IL18 – Legionellosis	Lymphocyte count	311493	-0.052	0.009	-5.966	2.44×10^{-9}	1.08×10^{-4}
IL18 – Shigellosis	Lymphocyte count	311493	-0.051	0.009	-5.923	3.17×10^{-9}	1.06×10^{-4}

Table 5.2. Significant PathWAS v1 PheWAS associations. The significant results of the association tests conducted between 44 UK Biobank phenotypes and the 7 significant pathway models created by PathWAS v1. Also shown are the betas and standard errors of the effect, as well as the Z-score. Here the r^2 predicted by the pathway alone was calculated by running a second model of the association including only the covariates and extracting the Δr^2 between both association tests.

5.3.4 ORCADES and Vis PathWAS TreeWAS

In order to further examine the effect of the pathways on phenotype, beyond the association test PheWAS, a hierarchical-structed form of PheWAS, TreeWAS, was conducted (by Dr. Xue Li) as described in **section 3.7**. This was conducted to further test the seven $PRS_{Pathway}$ against disease traits in UK Biobank in a structured and systematic manner. Of the seven $PRS_{Pathway}$ tested, only the NOD-like signalling pathway in adipose subcutaneous tissue (end-point: IL18) produced any significant associations (**Table 5.3**).

This analysis further implicated the NOD-like signalling pathway with circulatory disease, heart disease, cerebrovascular disease, headaches and diseases of the nervous system. The NOD-like signalling pathway is intricately involved in human innate immunity and inflammation[205,206]. However, there has been some previous evidence implicating the pathway in both heart disease[207] and neurodegenerative disorders[208], potentially through inflammatory mediators.

5.3.5 ORCADES and Vis PathWAS sensitivity analysis

In order to make sure that the pathway-phenotype associations were being driven by the pathway, as well as testing the $PRS_{Pathway}$ against the phenotype we also tested each individual PRS_{Gene} from within the pathway against the associated significant phenotype. If any of the individual genes was associated with the phenotype with a more significant P-value than the $PRS_{Pathway}$ then it was assumed that this gene would be driving the interaction between the pathway and the phenotype and would this would therefore fail the sensitivity analysis.

Of the 12 significant associations discovered, 11 failed this sensitivity analysis, with the exception of the relationship between the Chagas disease pathway, with the end-point CCL3 in EBV-transformed lymphocytes and the phenotype of lymphocyte count. I.e., in 11 instances of the sensitivity analysis, at least one PRS_{Gene} had a more significant P-value in association with the phenotype than the overall $PRS_{Pathway}$. In each instance of the sensitivity analysis which failed, there was only one gene which had a lower P-value than each of the other 11 phenotype-pathway associations (**Table 5.4**).

TreeWAS coding	Phenotype	Max beta	CI LHS	CI RHS	Probability
Chapter XVIII	Chapter XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	-0.08	-0.125	-0.065	0.998
Chapter IX	Chapter IX Diseases of the circulatory system	-0.08	-0.125	-0.065	0.996
Block R50-R69	R50-R69 General symptoms and signs	-0.08	-0.125	-0.065	0.994
Block I30-I52	I30-I52 Other forms of heart disease	-0.08	-0.125	-0.065	0.985
Chapter XIV	Chapter XIV Diseases of the genitourinary system	-0.08	-0.125	-0.065	0.984
Block N80-N98	N80-N98 Noninflammatory disorders of female genital tract	-0.08	-0.125	-0.065	0.977
Block I60-I69	I60-I69 Cerebrovascular diseases	-0.08	-0.125	-0.055	0.973
Chapter VI	Chapter VI Diseases of the nervous system	-0.08	-0.125	-0.055	0.972
R51	R51 Headache	-0.08	-0.125	-0.055	0.964
Block R00-R09	R00-R09 Symptoms and signs involving the circulatory and respiratory systems	-0.08	-0.125	-0.065	0.964
Block R25-R29	R25-R29 Symptoms and signs involving the nervous and musculoskeletal systems	-0.08	-0.125	-0.055	0.960
R69	R69 Unknown and unspecified causes of morbidity	-0.08	-0.125	-0.065	0.956

Table 5.3. Significant TreeWAS results for NOD-like receptor signalling in adipose tissue. Significance was defined as any result which had a probability >0.95. The confidence interval values on the left-hand side (LHS) and right-hand side (RHS) are also shown, as well as the beta.

In the sensitivity analyses which failed it was also frequently the same gene which appeared to be driving the various pathway-phenotype associations, with 4 different genes seemingly responsible for the 11 associations: NLRC4, MAPK3, MAPK9 and CASP8.

Phenotype	Pathway	Gene	PathWAS P-value	Gene P-value
Lymphocyte count	CCL3 – Chagas disease	TLR6	3.86×10^{-11}	2.81×10^{-10}
Lymphocyte count	IL18 - NOD-like receptor signalling (Adipose)	CASP8	1.60×10^{-7}	1.72×10^{-18}
Lymphocyte count	IL18 - NOD-like receptor signalling (Tibial nerve)	NLRC4	2.47×10^{-6}	2.20×10^{-9}
Lymphocyte count	IL18 - Legionellosis	NLRC4	3.17×10^{-9}	2.20×10^{-9}
Lymphocyte count	IL18 - Shigellosis	NLRC4	2.44×10^{-9}	2.20×10^{-9}
Waist circumference	CCL3 - Chagas disease	MAPK3	2.63×10^{-6}	4.17×10^{-18}
Hip circumference	CCL3 - Chagas disease	MAPK3	4.21×10^{-10}	3.31×10^{-17}
Height	IL18 - NOD-like receptor signalling (Adipose)	MAPK9	1.01×10^{-4}	5.27×10^{-16}
Weight	CCL3 - Chagas disease	MAPK3	8.04×10^{-10}	5.15×10^{-21}
BMI	CCL3 - Chagas disease	MAPK3	9.01×10^{-7}	6.05×10^{-18}
Erythrocyte count	CCL3 - Chagas disease	MAPK3	3.39×10^{-10}	1.32×10^{-40}
Platelet count	IL18 - NOD-like receptor signalling (Adipose)	CASP8	4.68×10^{-10}	2.37×10^{-15}

Table 5.4. Phenotype-pathway associations and associated P-value as well as individual gene which had the lowest P-value in the sensitivity analysis. As such only the first result passed the analysis.

Given the significance of these associations it was assumed that the significance of the pathway-phenotype relationship was primarily being driven by this one individual gene-phenotype association instead. As such, the only true pathway-phenotype relationship discovered was that between the chagas disease pathway and lymphocyte count.

5.3.6 ORCADES and Vis PathWAS PheWAS second iteration

An additional test was then conducted by incorporating an additional form of "tissue" into the PathWAS analysis. As well as examining tissue-specific pathways, we now also examined the "Complete" pathway. This included every gene connected to the end-point protein, but did not take into account the GTEx TPMs restrictions. However, this analysis revealed no new significant pathways predicted in Vis.

5.4 Discussion

The tests performed in this chapter acted as an proof-of-concept for PathWAS, indicating that it was possible to predict protein levels from a combined PRS for pathway function and subsequently it was possible to use these models to find traits significantly associated with the pathways in UK Biobank. There were, however, some obvious limitations with the approach, particularly given the small number of results and that even fewer of these survived the sensitivity analysis. Many of these issues would need to be solved in future versions in order to improve the discovered pathway-phenotype associations.

One important consideration as well was that while we used the prediction of protein levels in Vis to define the best $PRS_{Pathway}$, this was not the aim of the project or the model. While the protein levels were theoretically representative of pathway functionality, they were primarily used as a method of scaling or weighting the gene effects in creation of the PRS_{Gene}. More importantly, the overall level of contribution from pathway effect on the end-points was usually very small (based on small r^2s for each pathway model). As such, it is likely that those pathways which could predict the end-point protein in another independent dataset would rely on either finding pathways which contributed a large amount to the end-points abundance. Alternatively, those pathway models which best predicted the end-point also may have relied on particularly abundant or important proteins measured which maintain consistency between populations. As such the prediction step of the PathWAS pipeline was primarily used as a way of determining the strongest $PRS_{Pathway}$ but potentially should not exclude other models from being examined for phenotype associations.

As well as this there were some significant limitations in the methodology. One major issue was the lack of power conferred by the usage of ORCADES and Vis proteomics, with only 1000 and 800 individuals per cohort, this is a relatively underpowered dataset, as such we aimed from this point forward to include larger proteomics datasets.

The second primary limitation was in the use of straightforward linear and elastic-net regression of genes against the protein end-point. Given the use of eQTLs for the expression PRS, rather than individual level transcriptomics measurements, it was possibly too-far removed from the final objective of pathway function to be able to accurately predict it. Unfortunately, individual-level data is difficult to access and also raised the potential for new issues such as confounding. Therefore, in order to maintain the validity of using eQTL summary statistics, we would opt to switch from using either linear or elastic-net regression as the method for creating the gene effects on pathways. Instead, going forward we begin to utilise Mendelian randomisation (MR) of the QTL SNPs with the end-point proteomics, which theoretically ought to both overcome any possible confounding and also

provide a more direct link between the genetics of gene expression and the expression of the proxy for pathway functionality.

Another conclusion drawn during this phase of the project was that while the specificity and strength of the QTLs was an important factor in PathWAS, an equally important factor seemed to be the number of QTLs used in creation of the model. As the method is an attempt to estimate functionality of a pathway, I hypothesised that the more expression PRS used in creation of the model, the more representative of the overall pathway the model will be. This was because it would theoretically include more measurements of the varying directions of effect of each gene in the pathway. This would be in stark contrast to models with only a few gene measurements as in these cases the model would be closer to a relationship between one or two genes and a phenotype rather than an overall pathway. This was seemingly confirmed by the poor performance and viability of the use of the INTERVAL pQTLs in place of the SCALLOP eQTLs. One possible way to improve the PathWAS methodology in future would be the incorporation of multiple sources of QTL data, and primarily use eQTLs but use pQTLs for genes where they are available. However, this is an aspect of the project that was not revisited.

Another aspect which required re-examination was the sensitivity analysis utilised to determine which pathway-phenotype associations were legitimate and which were driven by individual genes. It was deemed that relying on P-value alone was not a valid method of determining whether a pathway-trait association was driven by any individual gene, as it was possible for there to be both gene-trait associations as well as the pathway-trait association, with a significant gene association perhaps only account for a small proportion of a pathway's effect on the trait. As such, for the future iterations of PathWAS a different sensitivity analysis method would need to be devised.

Another aspect of the PathWAS methodology which was not revisited, which however may be worth some consideration is the applicability of the tissue-specific pathway scores. While it is arguable how effective the methodology was at creating tissue-specific scores and predicting traits with them, one element which might be worth examining in more detail is the variation of pathway expression across different tissues. By comparing multiple different expression profiles it may be possible to discover the core gene set used in each pathway axis, finding those genes which are most relevant across all tissues in each pathway. While, this was beyond the scope of the nature of proving the usage of the PathWAS methodology, the functionality for creating tissue-specific networks was retained for this reason.

The TreeWAS conducted using the significant $PRS_{Pathway}$ yielded intriguing results further implicating the NOD-like receptor signalling pathway in a number of diagnostic phenotypes. However, due to the computationally demanding nature of the TreeWAS methodology and the subsequent expansion of the number of pathways analysed in **Chapter 6**, this form of analysis was not repeated. However, it

could be valuable to utilise the TreeWAS method in future iterations of the PathWAS method when specifically examining disease phenotypes.

5.5 Conclusions

In conclusion, PathWAS has shown promise as a method for analysing pathway-level associations between genetic variants and complex traits. While the proof-of-concept study has demonstrated its potential, it is clear that there are limitations and areas for improvement that need to be addressed. Further work would be needed to refine the methodology, optimise the statistical power, and improve the application of sensitivity analyses. By revisiting and improving the PathWAS approach, I theorised we could unlock its full potential to aid in understanding the genetic basis of complex diseases and ultimately advance precision medicine.

Chapter 6

PathWAS v.2 and R package results

6.1 PathWAS v.2 rationale and approach

The results of the first iteration of PathWAS, while promising, were disappointing in terms of the number of pathway-phenotype associations obtained from >300 different pathways across 49 tissues. Part of this was likely due to the incorporation of the prediction step of the method, in which the total number of pathway models used for PheWAS was reduced to only seven which could significantly predict the associated protein end-point in an additional cohort. Of note the aim of the PathWAS models is not to be able to predict the end-point omics measurement, but instead to use the omics to weight the individual expression PRS to create an estimate for pathway function. As such, the prediction phase could be argued to be overly stringent in removing models from use. However, in so doing it provides a possible estimation of the best models within the dataset, suggesting that the majority of the pathway models are not particularly strong.

The lack of results could also be attributed to a number of limitations within the methodology which were discussed in **section 5.4**. Particularly relevant are the following concerns: 1. a lack of statistical power from the proteomics datasets, 2. the potentially inappropriate usage of linear and elastic-net regression as the means of deriving the gene effects on the pathway and 3. the substandard sensitivity analysis used to discard pathway-phenotype associations based only on having an individual gene-phenotype association with a more significant P-value. As such, going forward we attempted to address as many of the concerns raised in subsequent versions of the PathWAS pipeline.

The next version of the PathWAS method was the result of several iterations of the methodology, which was repeatedly refined from the prior version before incorporation into the data presented in this chapter. As well as refining the methodology, much of the work conducted here has now been incorporated into an online R Package for PathWAS which would allow others to conduct their own PathWAS analyses.

Below the rationale behind some of the changes made is discussed.

6.1.1 Utilising whole pathways

A primary change to the methodology presented in this chapter is that we no longer used the pathway-specific PathWAS models. Instead, we relied on the "Complete" pathway scores, as detailed previously (**section 5.3.6**). This was done for two primary reasons: 1. In order to maximise the

number and strength of the pathway models available, it was important to incorporate as many genes as possible. A number of PathWAS runs failed due to having small pathway routes connected to the end-point and then having a key gene classified as "not expressed" within the majority of the tissues (based on the GTEx TPMs file). This would result in a pathway being classified as having fewer than the 2 genes necessary to constitute a "pathway" and so the methodology would fail. Furthermore, the best pathway models tended to be longer with genes than the weaker models, and so by incorporating all genes in a given pathway and ignoring tissue-specific effects was predicted to achieve stronger pathway effects. 2. I argue that it is valid to ignore the tissue-specific pathways due to the method by which I measure the end-point proteomics and the eQTLs used in the creation of the PRS_{Gene}. As neither data source is tissue-specific it would likely involve more proteins which were secreted from cells and tissues and so by using an overall view of the pathway, rather than a tissue-specific one, this may compensate for tissue-specific levels of secretion of the end-point protein.

As such, all of the $PRS_{Pathway}$ in this chapter are no longer tissue-specific, but instead are created for the whole or "complete" pathway associated with the end-point. The option for retaining tissue-specific pathways was, however, kept as an option within the PathWAS R package.

6.1.2 Conversion from regression to Mendelian randomisation

In the previous version of PathWAS we used linear and elastic-net regression of the individual PRS_{Gene} against the measured protein end-point to weight the scores by pathway function. Using regression models such as these however, is potentially vulnerable to confounding issues and as such might not be capturing the specific effects of the genes on the end-point as part of the pathway.

In order to overcome the potential limitations with usage of regression, we switched to using a multivariable Mendelian randomisation (MVMR) method, using the eQTL SNPs from which we created the PRS_{Gene} as the instrumental variables for the MVMR[209]. Now instead of using the measured protein levels as an end-point, we instead used the summary statistics from a GWAS of the protein levels as the outcome for the MVMR, with each individual gene as one of the multiple exposures.

By switching to MVMR instead of linear regression we theoretically would be more specifically weighting the effects of each individual gene against the end-point (the proxy for pathway function). Moreover, this would be free of confounding and also acknowledge the relationship between the different genes within the pathway, as each exposure would be weighted in accordance with each other. The MVMR method used relies on LASSO for variable selection to theoretically improve the accuracy of the prediction model created.

While the approach of using a GWAS of an end-point omics measurement as a proxy for pathway function is a novel one, the usage of eQTL SNPs as IVs in a Mendelian randomization analysis is not without precedent. A previously developed approach, transcriptome-wide Mendelian Randomization (TWMR), utilises eQTLs as IVs for gene expression to search for causal gene-trait associations, accounting for other gene expression profiles simultaneously[210]. This methodology differs from the PathWAS approach in that it directly searches for gene-trait relationships via MVMR, while the PathWAS methodology uses the MVMR effect sizes to weight PRS_{Gene} which are combined to search for pathway-trait relationships.

A major consideration of MR is ensuring that the instruments used in the analysis hold to the principles of MR, I.e. that they should be indepdent and. associated with only the exposure, not associated with the outcome and also should not be influenced by additional external confounders. Here the former caveat is accounted for by ensuring that the eQTL SNPs used as IVs are significant for their association with the given exposure of interest. It is difficult to account for confounding of individual SNP effects due to pleiotropy of influence with other genes (for example), however the effects of any pleiotropy may be reduced due to the MVMR approach, specifically if the SNP is an eQTL for multiple genes within the same pathway. However, the possibility of further pleiotropy or confounding by other genetic effects is not impossible and must be kept in account when analysing the results.

6.1.2.1 Selection of best MVMR exposures

In order to have a model which best represents the reality of the pathway, this would theoretically be one which includes as many gene exposures as possible. However, it is possible that this would introduce excessive noise into model creation and so could lead to false positives in the association testing between pathways and phenotypes. This was the same rationale behind testing elastic-net regression alongside linear regression. As such, we briefly experimented with two alternative methods of pruning exposure selection to include only the strongest associations within each pathway.

When creating the MVMR model, the method used would also provide a list of the exposures (the genes) which were significantly associated with the outcome (the end-point protein) from the list of genes provided. As such it was possible to prune the pathway scores to include only the weighted PRS_{Gene} which were significant exposures within the model.

An alternative method was, instead of using a LASSO MVMR method, to use one based on elastic net penalised regression. This involved incorporation of a function by Verena Zuber which would use the same input for the LASSO MVMR and instead conduct elastic net penalised regression to use the most relevant exposures on the end-point protein.

In the end both versions were kept as functions within the PathWAS package but were not used as part of the analysis for this project. This was because in the majority of cases, each pathway model would have at most one exposure which was significant or was retained by elastic net (often the same exposure in both methods), and thus the $PRS_{Pathway}$ would be created based on only one gene, thus invalidating the attempt of predicting overall pathway functionality.

6.1.3 Additional proteomics used

A further expansion of the PathWAS results was conducted by the incorporation of the SCALLOP and DeCODE proteomics datasets. The SCALLOP proteomics dataset uses the same OLINK proteomics technology which was used in **Chapter 5**, however now applied across a large multinational consortium (**section 3.2.3**). This meant a significant increase in sample size as well as changing which OLINK panels were used. Instead of using CVD2, CVD3 and INF (the three panels which overlapped between ORCADES and Vis) I switched to using the CVD2, CVD3, NEU (neurology) and NEX (neuro exploratory panels). Using SCALLOP with these panels represented upwards of a 10-fold increase in sample size as well as removal of 92 proteins from the dataset, but addition of another 184 instead. As described in the methods (**section 3.2.3**), the SCALLOP consortium involved multiple GWAS of these panels combined into a meta-analysis, with the specific exclusion of the ORCADES cohort which was retained as an independent testing dataset.

In addition to the SCALLOP proteomics, we also expanded the PathWAS methodology to the DeCODE proteomics, incorporating an additional 4,719 unique protein end-points, with some proteins overlapping with those from the SCALLOP data. This set of proteomics is based on the SomaLogic technology and as such was conducted as a separate analysis from the SCALLOP analysis. The DeCODE proteomics data included not only more proteins but also had a higher sample size than the largest of the SCALLOP datasets. Conversely however, there was no equivalent prediction data available for the DeCODE proteomics and therefore every model created was used in the subsequent analyses, as opposed to with the SCALLOP data where only those pathways with which the end-point could be predicted in the ORCADES dataset, were used for the subsequent PheWAS.

6.1.4 Revising PRS usage a third time

While PRS_{Gene} had been successfully created for all of the genes in the SCALLOP pathways using LDpred2, the incorporation of the DeCODE dataset also presented an opportunity to revisit the methodology for the creation of PRS. This was primarily done as the usage of LDpred2 continued to present computational concerns. While the PRS created by LDpred2 were believed to be accurate, the

creation of them was a slow process with each PRS_{Gene} taking up to 24 hours to create as well as extreme amounts of computing resources (this was in spite of using only the data from single chromosomes and often fewer than 10,000 SNPs per score). Additionally, LDpred2 relies on the creation of an internal estimate of heritability with LDSC and in a small number of cases this was not possible, meaning that a number of PRS had to be created with PRsice-2, providing an element of inconsistency to a new iteration of results.

In collaboration with Dr. Paul Timmers, the LDpred2 pipeline I had written was incorporated into a PRS pipeline which also provided the options for using PRsice-2 and PRS-CS. PRS-CS proved to be an attractive alternative to LDpred2 in the creation of PRS_{Gene}. Following the initial usage of LDpred2, further research in the field of PRS appeared to consistently suggest that PRS generated by PRS-CS would either be at least as accurate or even outperform those generated by LDpred2[138,211]. Furthermore, this PRS-CS was notably faster, taking less than an hour for most PRS_{Gene}. As such, for this final version of PathWAS all of the PRS_{Gene} created were made with PRS-CS.

6.1.5 Leave-one-out sensitivity analysis

To exclude the possibility that the observed significant effects were due to the effect of a single gene, we performed a Leave-One-Out Analysis (L1O) where the whole methodology of creation of $PRS_{Pathway}$ was performed again, omitting a single gene in the pathway each time. The gene was omitted starting from the SNP clumping step in order to make the new analysis completely independent of the gene. These L1O "$PRS_{Pathway\text{-}GeneN}$" were tested against the phenotypes for which significant associations were observed, essentially running a new PheWAS for each $PRS_{Pathway\text{-}GeneN}$ (for as many genes as there were in the original pathway).

In multiple instances, the PathWAS models were created based on 2 genes (with 2 genes being the minimum number used to initially designate a pathway) and thus removal of either would prevent the usage of the MVMR methodology. In these instances, more standard IVW Mendelian randomisation was conducted, with the individual gene representing the sole exposure in the analysis. For this we use the mr_ivw function from the Mendelian randomisation package.

As before, the phenotypes were standardised prior to analysis, and a new association test using speedglm function *glm* was conducted for each of the significant traits discovered.

To further verify that the observed effects were not mediated through the end-point protein, we created a PRS for each of the end-point proteins ($PRS_{Protein}$). We limited the creation of these $PRS_{Protein}$ for the analysis to *cis*-SNPs (within 1 Mb of the start or end of the open reading frame, defined by the start and end of the gene from the Ensembl build 19 data). For the creation of the $PRS_{Protein}$, we again

used PRS-CS with the same options as previously described. These $PRS_{Protein}$ were also then used in the same association analyses against significant traits as the corresponding $PRS_{Pathway}$.

Z-scores were then created (**Equation 6.1**) and analysed for each significant PheWAS result, comparing the initial significant result with each individual association test from the L1O. A Z-score (Z-statistic) is a statistical measure that quantifies how many standard deviations a data point is away from the mean of a distribution. Where the pchisq score for any separate $PRS_{Pathway\text{-}GeneN}$ was > 0.05, this relationship was deemed to have failed the L1O analysis and was therefore excluded from the final significant results. In this instance the Z-score method allows for the analysis of the distribution of the $PRS_{Pathway\text{-}GeneN}$ effect sizes following removal of any individual gene when compared with $PRS_{Pathway}$ with large divergences considered of the effect to have failed the test as we hypothesise that the large changes in effect are due to the gene. We also tested for nominal significance of the protein following the Z-score testing, where the $PRS_{Protein}$ passed the threshold of $P < 0.05$ it was deemed to be impossible to rule out mediation of the effect via the protein. A final Z-score test was then conducted between the significant protein associations and the $PRS_{Pathway}$ which was used to examine whether the protein effect was in the same or in opposing direction to that of the $PRS_{Pathway}$. This was calculated as follows:

$$Z = \frac{\beta_{Pathway} - \beta_{Test}}{\sqrt{SE_{Pathway}^2 + SE_{Test}^2}}$$

Equation 6.1.

Where the β (effect size) and standard error (SE) for "Test" is the β and SE for the equivalent $PRS_{Pathway\text{-}GeneN}$ or $PRS_{Protein}$ for the phenotype, while the β and SE for the Pathway are those for the $PRS_{Pathway}$.

6.1.6 Pathway gene selection refinement

In addition to the major refinements to the methodology, the method by which genes were extracted from the KEGG database for each pathway (**sections 5.2.2** and **5.2.3**) was altered. This was done partly to improve computational efficiency and speed, as in some instances the previous method would result in lists of many millions of simple pathway routes (which became impossible for the computer to use). It was also refined as the previous method incorrectly defined some proteins as end-points in pathways when in fact they were part of a protein complex.

Due to the way the KEGG XML files were encoded, most genes encoding proteins which exist in protein complexes would not have outward edges, with the exception of one gene within the complex. As such if a protein was within a complex and was not the gene selected in the KEGG files to represent the node with the outward edge, then it would be erroneously classified as an end-point. This was rectified utilising the KEGGlincs package to confirm whether genes existed in complex and then confirming whether the complex itself had outward edges.

The new gene lists were now generated from the bottom up, by iteratively searching for any node connected to the end-point by an inward-facing edge, and then searching for nodes attached to those and so-on until the number of iterations of the search exceeded the number of nodes tested (I.e., all inward connections had been discovered).

6.1.7 Approach summary

Incorporating the changes to the methodology, the final PathWAS v.2 method was conducted as follows (**Fig 6.1**).

Using the refined method of classifying end-points, 377 pathway-endpoint combinations were classified in SCALLOP across 368 unique proteins (notably this is the same number of combinations as in the PathWAS v.1 analysis, despite replacing one panel with two others, potentially suggesting that the NEX and NEU panels contain fewer proteins which play prominent roles in pathways). In DeCODE 2,832 pathway-endpoint combinations were detected, with many overlapping between both datasets.

2,721 PRS_{Gene} were created from eQTLgen *cis*-eQTLs using PRS-CS (from a total gene list of ~4,000, only 2,721 overlapped with eQTLgen, but PRS could be successfully made with all of them).

Pathway gene lists were created and the significant SNPs within eQTLgen (FDR < 0.05) for these same genes were extracted and then clumped. The clumped sets of SNPs for each pathway were then subjected to an MVMR analysis against a GWAS of the end-point protein, this was conducted in both datasets.

256 MVMR LASSO models were created from the SCALLOP pathways and for these, $PRS_{Pathway}$ were created in ORCADES and used to predict the same end-point protein. This resulted in 19 pathway models which could successfully predict the ORCADES protein.

As we lacked SomaLogic protein measurements for performing prediction, all of the DeCODE pathways were used for the next step (a total of 1,808 $PRS_{Pathway}$).

These 1,808 and 19 models were used to create $PRS_{Pathway}$ in UK Biobank which were then used in an association test with 60 phenotypes and tested for significance using FDR (instead of the previously

used Bonferroni corrected associations). This resulted in 9,105 significant associations (FDR < 0.05) from the DeCODE PheWAS and 123 associations from the SCALLOP PheWAS.

An improved sensitivity test was performed involving a Leave-1-Out analysis utilising testing of $PRS_{Pathway\text{-}GeneN}$ with each significant pathway-phenotype association. As well as the use of $PRS_{Pathway\text{-}GeneN}$, PRS for the end-point protein (also created with PRS-CS) were used to test for mediation of the phenotype by the protein. In DeCODE the L1O analysis could not be completed in an additional 124 pathways, leaving only 8,239 significant associations to be examined for sensitivity.

This resulted in a total of 2,108 significant pathway-phenotype associations from the DeCODE analysis and 28 from the SCALLOP analysis which passed the sensitivity analysis, with a further 610 and 7 which passed the L1O but also had a significant association between the phenotype and end-point protein.

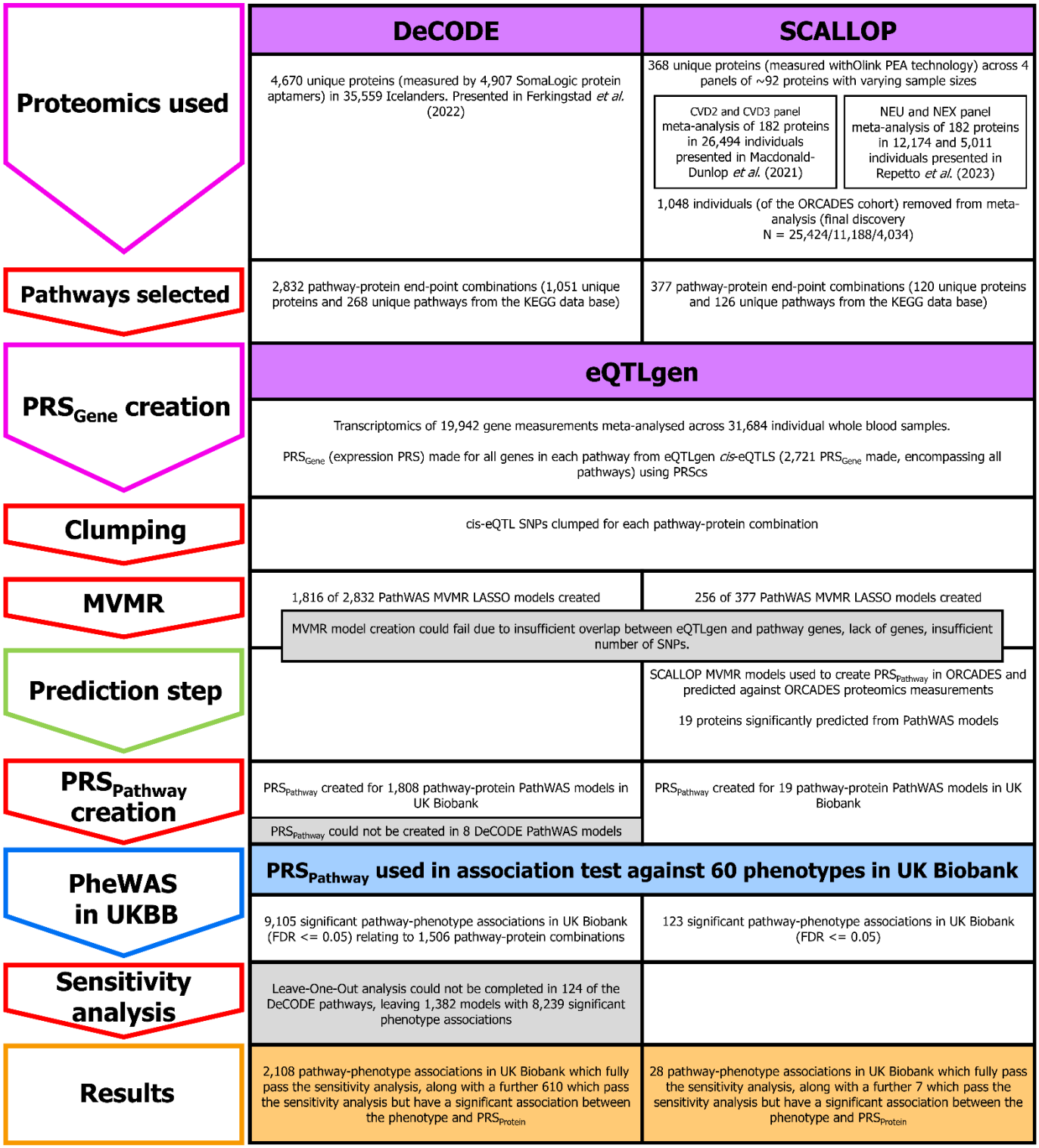


Figure 6.1. PathWAS summary chart. This shows the processes involved in the PathWAS v.2 methododology, as well as the various numbers of genes, pathways and traits involved in each step. Here purple indicates the usage of new summary statistics, red are the steps of the PathWAS method, green is the validation stage only used with the SCALLOP data, blue represents the association test in UK Biobank and orange is the final results of the entire process.

6.2 Outcomes

6.2.1 SCALLOP PathWAS

From the 368 SCALLOP proteins, it was possible to determine 377 unique pathway-protein-end-point combinations utilising the KEGG database, comprised of 120 different proteins across 126 KEGG pathways. Of these pathways, the PathWAS MVMR methodology was successfully applied to 256 combinations, creating a $PRS_{Pathway}$ model for each. There were a number of reasons why 121 combinations failed: if there were too few genes in the pathway (we limited the methodology to include only those with 2 or more genes as a "pathway"), too few SNPs associated with the extracted genes for the MVMR, insufficient overlap between the eQTLgen data and pathway genes (i.e. a pathway with 2 or more genes, but for which there were no eQTLs present), and a few for which it was computationally unfeasible to create the pathway gene lists (invariably these involved non-specific and very broad networks such as the KEGG "pathways in cancer", hsa05200).

Finally, to validate the $PRS_{Pathway}$, we tested their ability to predict the same protein used as end-point in an independent dataset of 1,048 people from the ORCADES study. Of the 256 MVMR models created, we were able to test prediction of the protein levels in 241 within the independent cohort to validate our findings. We used a significance threshold of $p < 0.1$ to determine the $PRS_{Pathway}$ models which could significantly predict the protein end-point. This significance test was a one-tailed test as we also limited the pathways to those which retained consistency in the directionality of the effect. This restricted the models used to those which had a positive beta for pathway effect (**Table 6.1**) on the protein. This was done to ensure that the analysis was limited to pathways where we were predicting a change in effect in the correct direction.

The final results produced 18 protein end-points which could be successfully predicted by a $PRS_{Pathway}$ pathway model (**Fig 6.2**), and it was these which were carried forwards for the PheWAS analysis as the strongest examples of pathways from PathWAS.

KEGG ID	Pathway	End-point protein	r^2	P-value	Beta	Number of genes	Significant gene model r^2	Significant gene model P-value	Significant gene model beta	Number of significant genes
hsa04621	NOD-like receptor signalling pathway	CXCL1	3.13×10^{-3}	7.26×10^{-2}	0.86	44	5.29×10^{-4}	4.61×10^{-1}	-0.63	5
hsa05167	Kaposi sarcoma-associated herpesvirus infection	CXCL1	4.82×10^{-3}	2.60×10^{-2}	0.00	72	2.84×10^{-4}	5.89×10^{-1}	-0.40	3
hsa04010	MAPK signalling pathway	HSPB1	2.81×10^{-3}	8.92×10^{-2}	0.33	52	3.94×10^{-4}	5.25×10^{-1}	0.26	5
hsa04621	NOD-like receptor signalling pathway	IL18	4.27×10^{-2}	2.24×10^{-11}	0.68	71	4.44×10^{-2}	8.49×10^{-12}	0.75	7
hsa05131	Shigellosis	IL18	4.42×10^{-2}	9.36×10^{-12}	0.75	5	4.44×10^{-2}	8.43×10^{-12}	0.75	4
hsa05132	Salmnoella infection	IL18	4.31×10^{-2}	1.73×10^{-11}	0.74	11	4.45×10^{-2}	8.04×10^{-12}	0.75	4
hsa05134	Legionellosis	IL18	4.54×10^{-2}	5.00×10^{-12}	0.76	3	4.54×10^{-2}	5.00×10^{-12}	0.76	3
hsa05135	Yersinia infection	IL18	3.40×10^{-2}	2.51×10^{-9}	0.65	30	3.56×10^{-2}	1.04×10^{-9}	0.67	7
hsa04620	Toll-like receptor signalling pathway	IL6	3.96×10^{-3}	4.35×10^{-2}	1.43	46	6.98×10^{-3}	7.34×10^{-3}	2.92	2
hsa05135	Yersinia infection	IL6	2.71×10^{-3}	9.52×10^{-2}	1.29	37	8.57×10^{-4}	3.48×10^{-1}	1.07	2
hsa04610	Complement and coagulation cascades	ITGB2	8.69×10^{-3}	2.48×10^{-3}	0.69	39	7.90×10^{-3}	3.92×10^{-3}	0.57	5
hsa03320	PPAR signalling pathway	LPL	3.67×10^{-3}	5.21×10^{-2}	2.72	5	-	-	-	-
hsa04115	p53 signalling pathway	SERPINE1	2.71×10^{-3}	9.13×10^{-2}	13.38	7	-	-	-	-
hsa04060	Cytokine-cytokine receptor interaction	TNFRSF13B	9.28×10^{-3}	1.98×10^{-3}	0.35	2	-	-	-	1
hsa04672	Intestinal immune network for IgA production	TNFRSF13B	9.28×10^{-3}	1.98×10^{-3}	0.35	2	-	-	-	1
hsa05170	HIV-1 infection	BST2	3.06×10^{-3}	7.83×10^{-2}	1.34	2	-	-	-	
hsa04612	Antigen processing and presentation	KIR2DL3	2.49×10^{-2}	4.40×10^{-7}	0.50	9	1.23×10^{-2}	4.10×10^{-4}	0.35	4

hsa05332	Graft-versus-host disease	KIR2DL3	2.20×10^{-2}	2.09×10^{-6}	0.50	6	1.78×10^{-2}	2.07×10^{-5}	0.42	4
hsa05225	Hepatocellular carcinoma	RPS6KB1	2.92×10^{-3}	8.57×10^{-2}	0.83	15	-	-	-	1

Table 6.1. SCALLOP PathWAS significantly predicted pathways. Shown are the 19 significantly predicted pathways and end-points from the SCALLOP PathWAS across 4 protein panels and 241 PathWAS models. Shown are the effect sizes (beta), predictive ability (r^2) and P-value of the pathway models ability to predict the associated end-point protein, using a P-value threshold of 0.1, restricting the analysis to pathways with at least 2 genes and to those with a positive beta. Also shown are the same values for the models created for the same pathways using only the significant MVMR exposures. The with end-point TNFRSF13B were also discovered to be identical, using the same genes in the creation of both. As such, for the PathWAS plot and discussion purposes, these were amalgamated into the "TNFRSF13 receptor signalling" pathway.

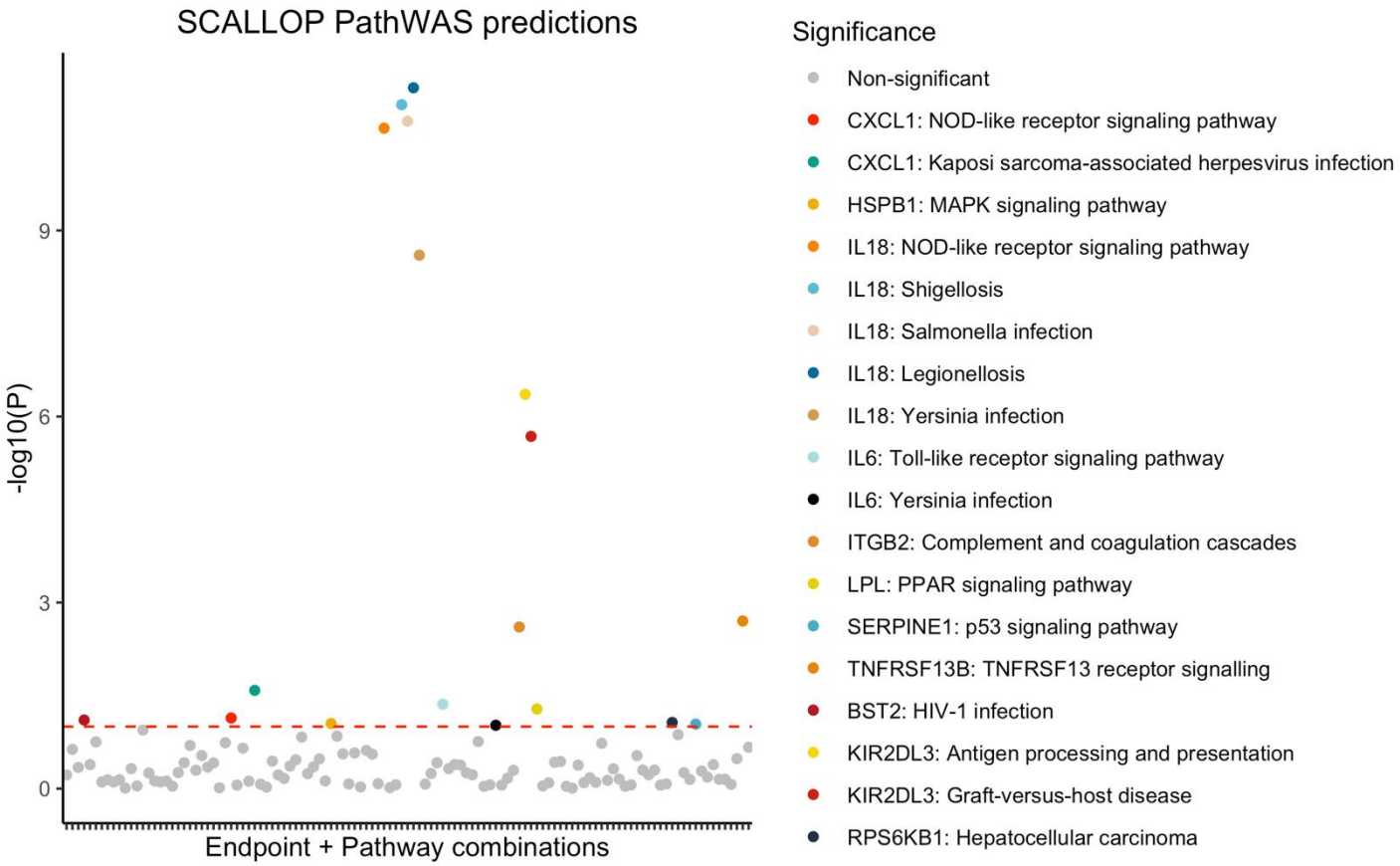


Figure 6.2: SCALLOP PathWAS predictions. From a total of 256 pathway models which could be created with the SCALLOP proteomics, 18 of the models could be used to successfully predict the same end-point protein in the independent ORCADES cohort.

From these models, the betas for each gene in the MVMR were used to create the $PRS_{Pathway}$, by multiplying the gene effect score with the individual PRS_{Gene} and then summing these for the pathway (as described above). Each $PRS_{Pathway}$ was also standardised to place them on the same scale.

When searching for pathways and conducting the model analysis, IL12A and IL12B were used interchangeably for the protein measurement for IL12 (a protein dimer of both IL12A and B) as it is these which could be found within KEGG. As well as this, the protein TNFRSF13B was an end-point for several KEGG pathways, however, this was always a subset of each pathway including the same genes in each instance. This has thus been renamed as the "TNFRSF13 receptor signalling" pathway, to prevent repetition. These changes are only accounted for in the final significant results and as such repetition may be present in the initial analyses.

6.2.2 SCALLOP PheWAS

Utilising the 18 significant pathway models, PRS were created for each pathway in UK Biobank. These PRS were then applied to an association analysis against 60 hand-selected phenotypes across all individuals in the cohort (Table S1). This revealed 123 significant (FDR < 0.05) associations between the pathways and different traits. Due to the high variation in $PRS_{Pathway}$, the betas of each trait in the pheWAS was also corrected by multiplication with the standard deviation of the $PRS_{Pathway}$.

6.2.3 SCALLOP sensitivity analyses

To verify that the observed phenotype associations with the pathways were driven by the pathway as a whole and not driven by either one individual gene from the pathway, or mediated through the levels of the protein used as the end-point, we conducted a sensitivity analysis. This analysis involved the testing of multiple $PRS_{Pathway\text{-}GeneN}$ and a $PRS_{Protein}$ for the end point protein against the significant phenotypes (**Fig 6.3** and **6.4**).

Out of the 123 significant connections found between pathways and phenotypes by SCALLOP, 35 were found to be valid after undergoing L1O analysis (**Fig 6.5**). Of these associations, 7 pathways also had a nominally significant association between the end-point $PRS_{Protein}$ and the trait ($P <= 0.05$). This suggests that 28 of the connections are likely driven by the pathway as a whole and not any individual gene, however it is not possible to rule out mediation or confounding of the pathway effect by the protein in the remaining 7.

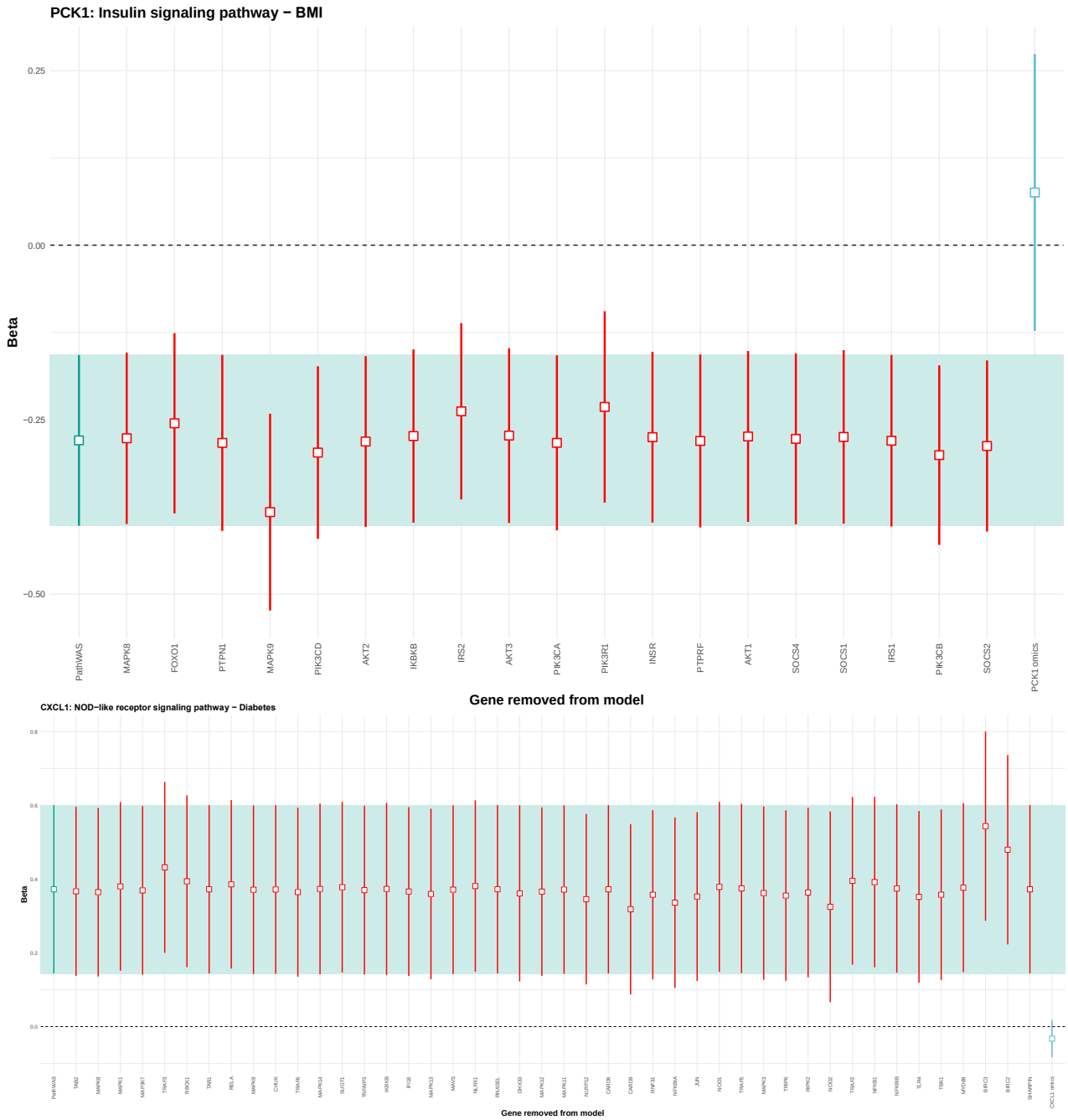


Figure 6.3. Example Passed Leave-One-Out Analyses. Results of two successful L1O analyses. The upper panel from the DeCODE dataset, the lower panel from SCALLOP. In the DeCODE results the Insulin signaling pathway with end-point PCK1 is examined in relation to its significant association with BMI. In SCALLOP the pathway analysed is the NOD-like receptor signaling pathway, with CXCL1 as the end-point and diabetes as the phenotype. The standard errors of the PathWAS association ($PRS_{Pathway}$, green) overlaps with each $PRS_{Pathway\text{-}GeneN}$ (red), which represent the same PRS with one gene removed. An association test between $PRS_{Protein}$ for the end-point protein and the phenotype (blue) also was non-significant (P = 0.45), indicating a complete pass for the sensitivity analysis.

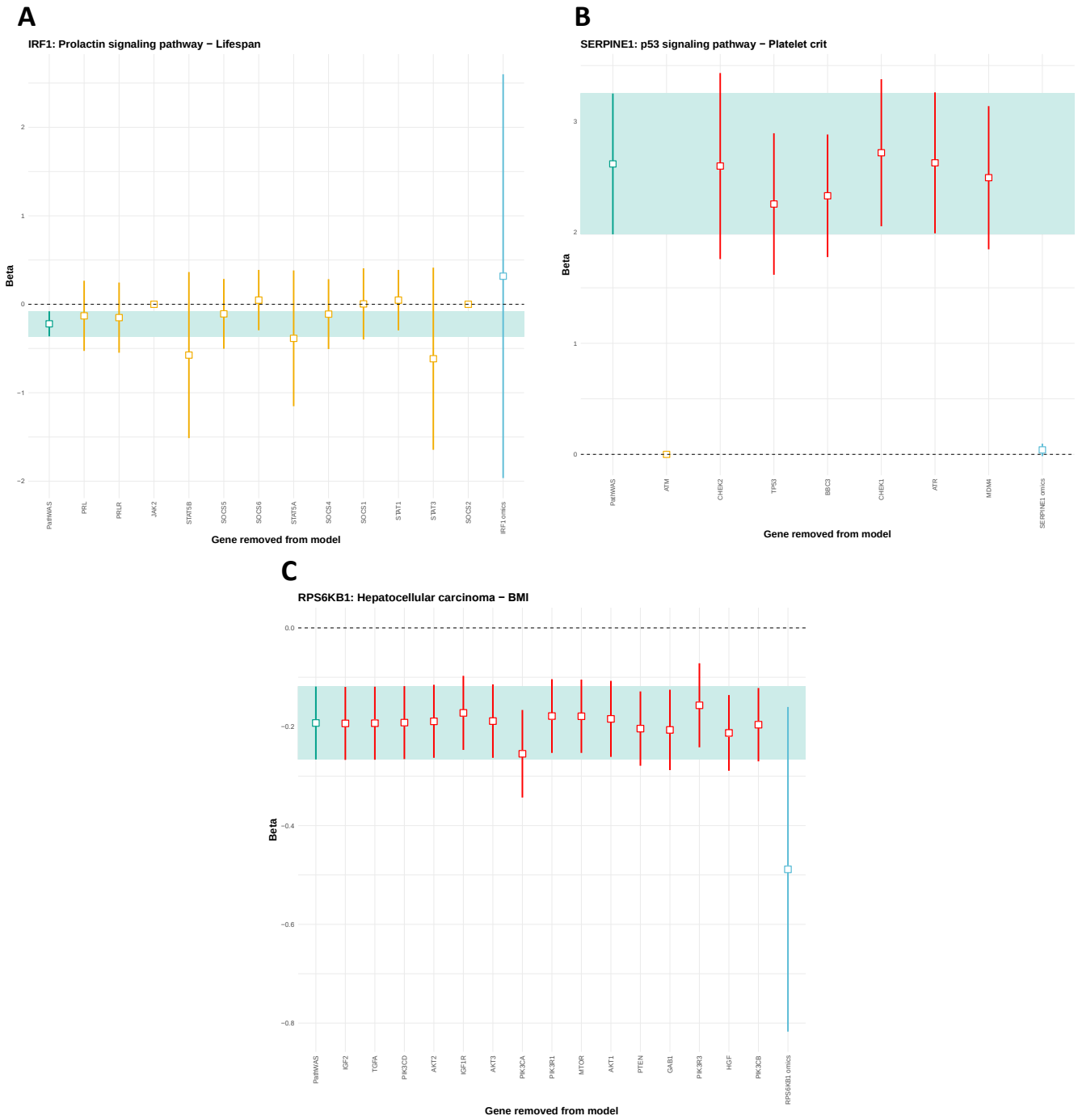


Figure 6.4. Example Failed Leave-One-Out Analyses. Three examples of significant PathWAS results which failed the leave-one-out analysis. Panel A is an example from the DeCODE data while panels B and C come from SCALLOP. Colours are the same as in **Fig 6.3**, with yellow lines now indicating when the association between trait and $PRS_{Pathway\text{-}GeneN}$ is now non-significant. In panel A several of the $PRS_{Pathway\text{-}GeneN}$ effects are totally different from the $PRS_{Pathway}$ effect, which is also clearly seen in panel B (I.e. where the ATM gene is driving the pathway effect). While panel C passed the Z-score testing, in this instance the $PRS_{Protein}$ for RPS6KB1 was significantly associated with the phenotype, and so is potentially driving the association (especially as it is in the same direction as the PathWAS result).

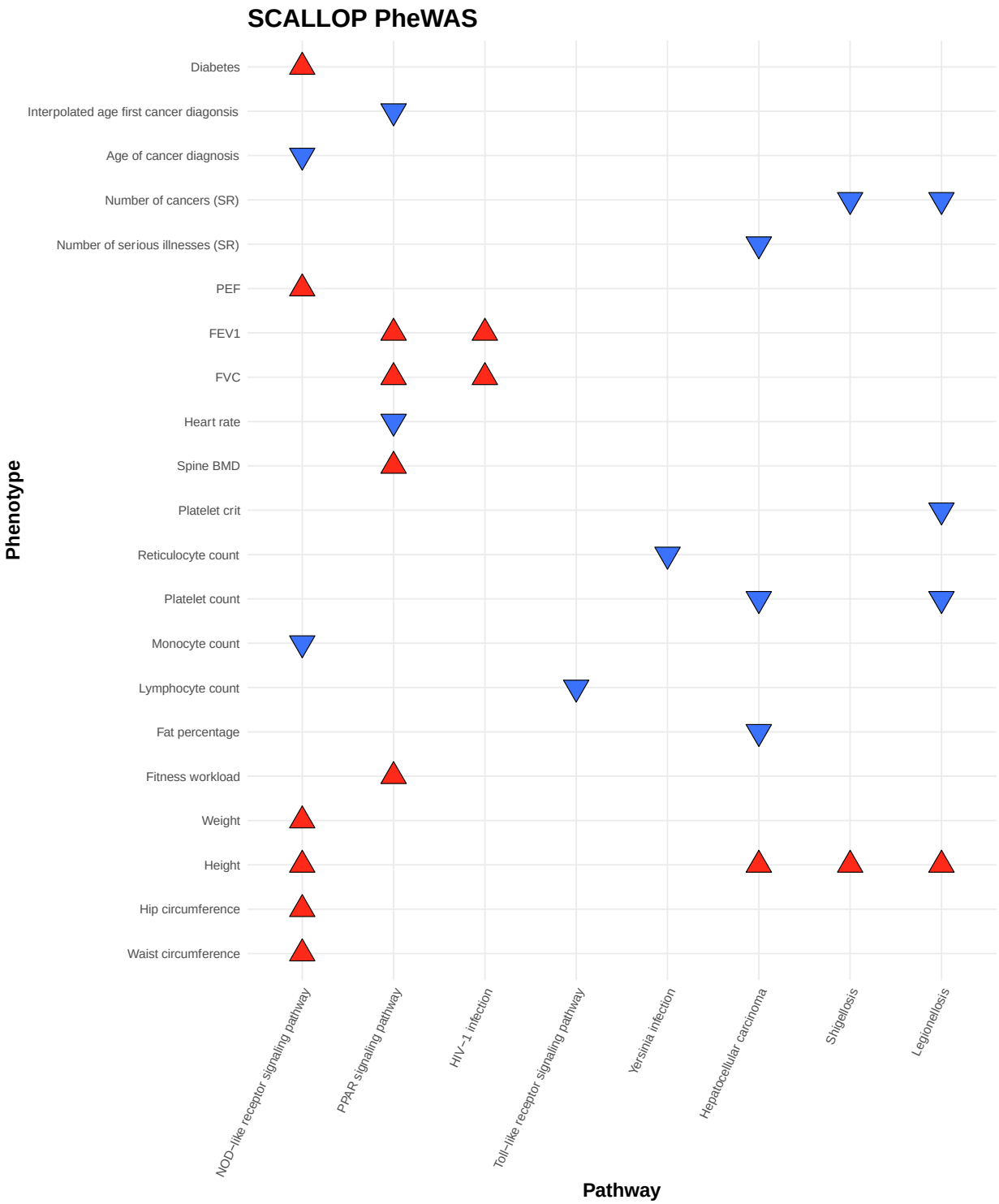


Figure 6.5. SCALLOP PathWAS PheWAS. From 18 significant pathways tested against 60 traits in the UK Biobank we discovered 123 significant pathway-phenotype associations. Of those, 28 fully passed the sensitivity analysis, indicating a likely causal relationship between the pathway and phenotype, not mediated by the end-point protein.

6.2.4 DeCODE PathWAS

A second and separate iteration of the PathWAS analysis was conducted using proteomics downloaded from the Icelandic DeCODE data set. From 4,907 proteins included in the analysis, 2,832 unique pathway-endpoint combinations were discovered in KEGG. From these the PathWAS MVMR methodology was successfully applied to 1,816 (with again, runs failing due to there being too few genes or eQTL SNPs in a pathway and too little overlap between genes and the eQTLgen-derived PRS_{Gene}).

Following this, PRS were created for the genes of each pathway to create the overall $PRS_{Pathway}$ for the PheWAS and the $PRS_{Pathway\text{-}GeneN}$ for the sensitivity analysis. As there was no additional validation proteomics dataset for prediction, all pathways were then carried forwards for association testing. This also meant no testing for direction of pathway effect was possible. From the pathway models, $PRS_{Pathway}$ could not be calculated in a further 8 pathways due to further computational issues relating to the MVMR weighting. This finally resulted in 1,808 pathway-endpoint combinations used in the PheWAS analysis.

These 1,808 pathways were tested against the same 60 phenotypes in UK Biobank as the SCALLOP data, including some disease traits, blood cell traits, anthropometry, physical and aging-related traits. This resulted in 9,105 significant (FDR < 0.05) traits, relating to 1,506 of the 1,808 pathways used in the test.

These 9,105 traits were subjected to sensitivity analysis testing with both the L1O analysis and testing for a relationship between $PRS_{Protein}$ and trait. The L1O analysis could not be correctly run in a further 124 instances (primarily because removal of individual genes from a given model resulted in too few SNPs for either MVMR or the IVW method of model creation). This then meant that the sensitivity analysis was applied across 1,382 pathways with 8,239 significant associations. Where the sensitivity analysis could be correctly employed, 2,718 significant pathway-trait associations successfully passed (**Fig 6.6-6.13**). This included 610 for which the L1O was passed, but where there was still a nominally significant association between the $PRS_{Protein}$ and the phenotype. The significant protein associations were also be divided up into those which have differing or complementing directions of effect from the pathway. Like with SCALLOP, the effect sizes of the associations were corrected by multiplication of the beta with the standard deviation of the $PRS_{Pathway}$.

We also found 1,485 pathways which failed the sensitivity analysis but had a nominally significant association with the end-point protein. Of those pathways which passed the sensitivity analysis and also had a significant association with the end-point protein, we further subcategorised them. 244 pathway-phenotype relationships which passed the sensitivity analysis but had a significant protein association had opposite directions of effect between the $PRS_{Pathway}$ and $PRS_{Protein}$. We categorised pathway-phenotype relationships where the significant protein effect was in the same direction as the

pathway, but the effect size was larger or equivalent to the pathway (292 and 74 associations respectively).

The DeCODE PathWAS results also revealed a number of instances (**Fig 6.6**, **6.7**, **6.9-6.13**) where the same pathway with multiple end-points would have significant associations for the same phenotype, passing the sensitivity analyses, but the direction of effect was different for each end-point. In these instances the arrow on top in the figure only corresponds to the alphabetical order of the protein end-point (proteins later in alphabetical order are layered on top). For example, in **Figure 6.6**, two overlapping arrows can be observed for the Maturity onset diabetes of the young (MODY) pathway (KEGG ID: hsa04950) and the heart rate phenotype. Here the arrows correspond to the end-points insulin (INS) and islet amyloid polypeptide (IAPP) with the effect sizes of -0.145 and 2×10^{-6} respectively.

Figure 6.6: DeCODE PathWAS PheWAS heart function. From a final total of 1,382 pathways, which could be subjected to each stage of the analysis, there was over 9,000 significant pathway-phenotype associations in UK Biobank. Of these associations 2,108 fully passed the sensitivity analysis with no mediation by the end-point protein. A subset of these relating to heart-function traits are displayed here. Blue arrows indicate that increased pathway activity leads to a decrease in the trait, red arrows implicate increased pathway activity with an increase in the trait. Where there are two arrows which overlap this indicates a pathway with multiple end-points, where each subset of the pathway for the specific end-point significantly predicts the trait in opposing directions.

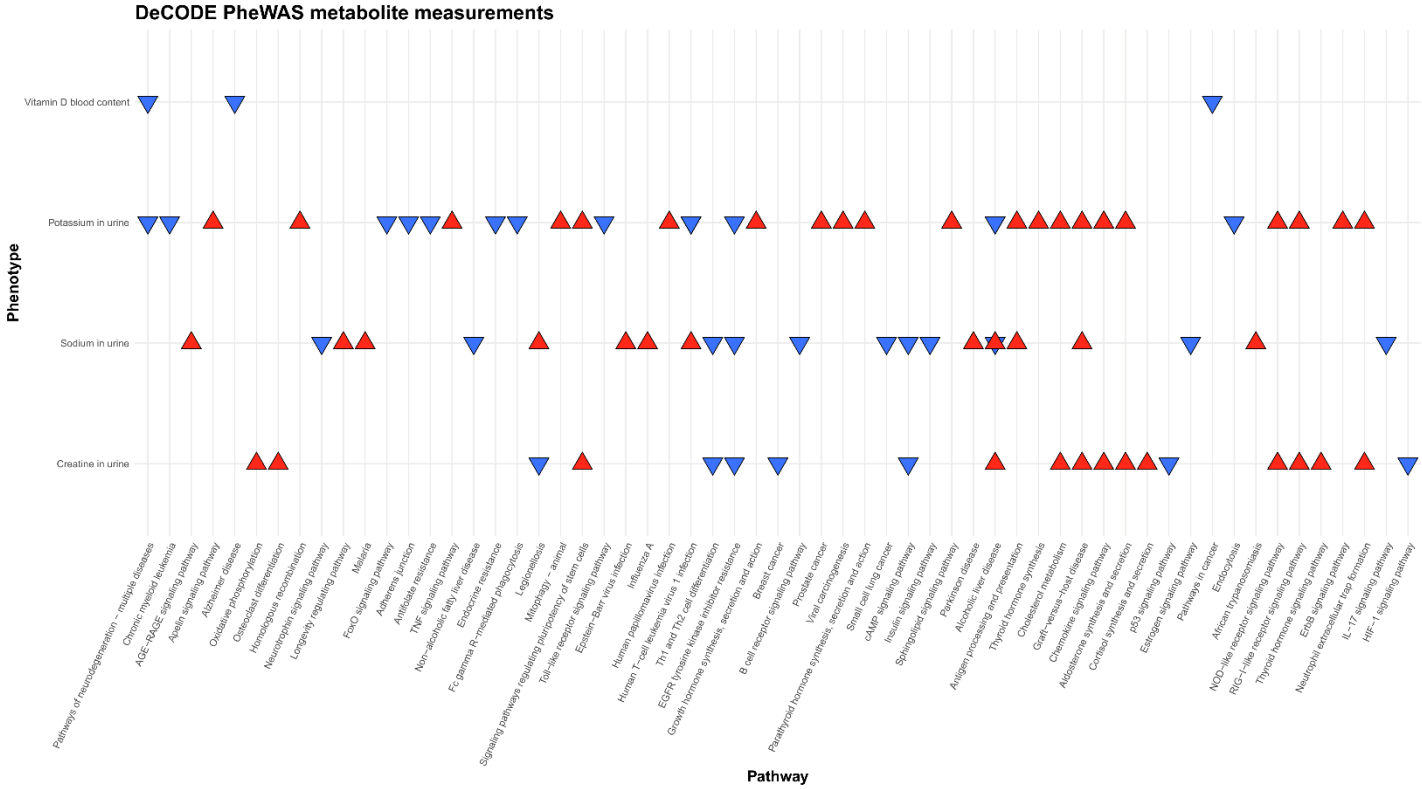


Figure 6.7: DeCODE PathWAS PheWAS metabolite measurements. Significant pathway-phenotype associations which passed the sensitivity analysis, relating to metabolite measurements. Blue arrows indicate that increased pathway activity leads to a decrease in the trait, red arrows implicate increased pathway activity with an increase in the trait.

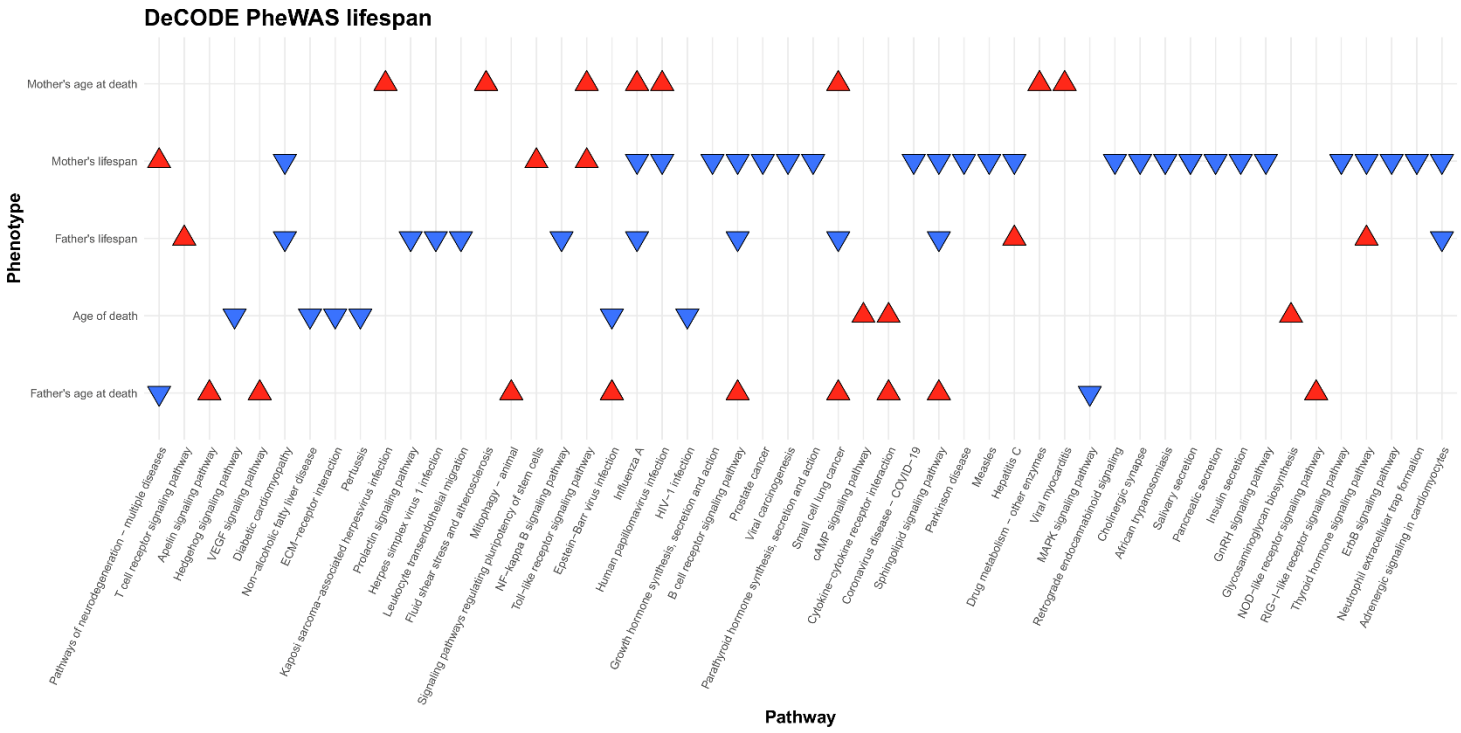


Figure 6.8: DeCODE PathWAS PheWAS lifespan. Significant pathway-phenotype associations which passed the sensitivity analysis, relating to lifespan traits. Blue arrows indicate that increased pathway activity leads to a decrease in the trait, red arrows implicate increased pathway activity with an increase in the trait.

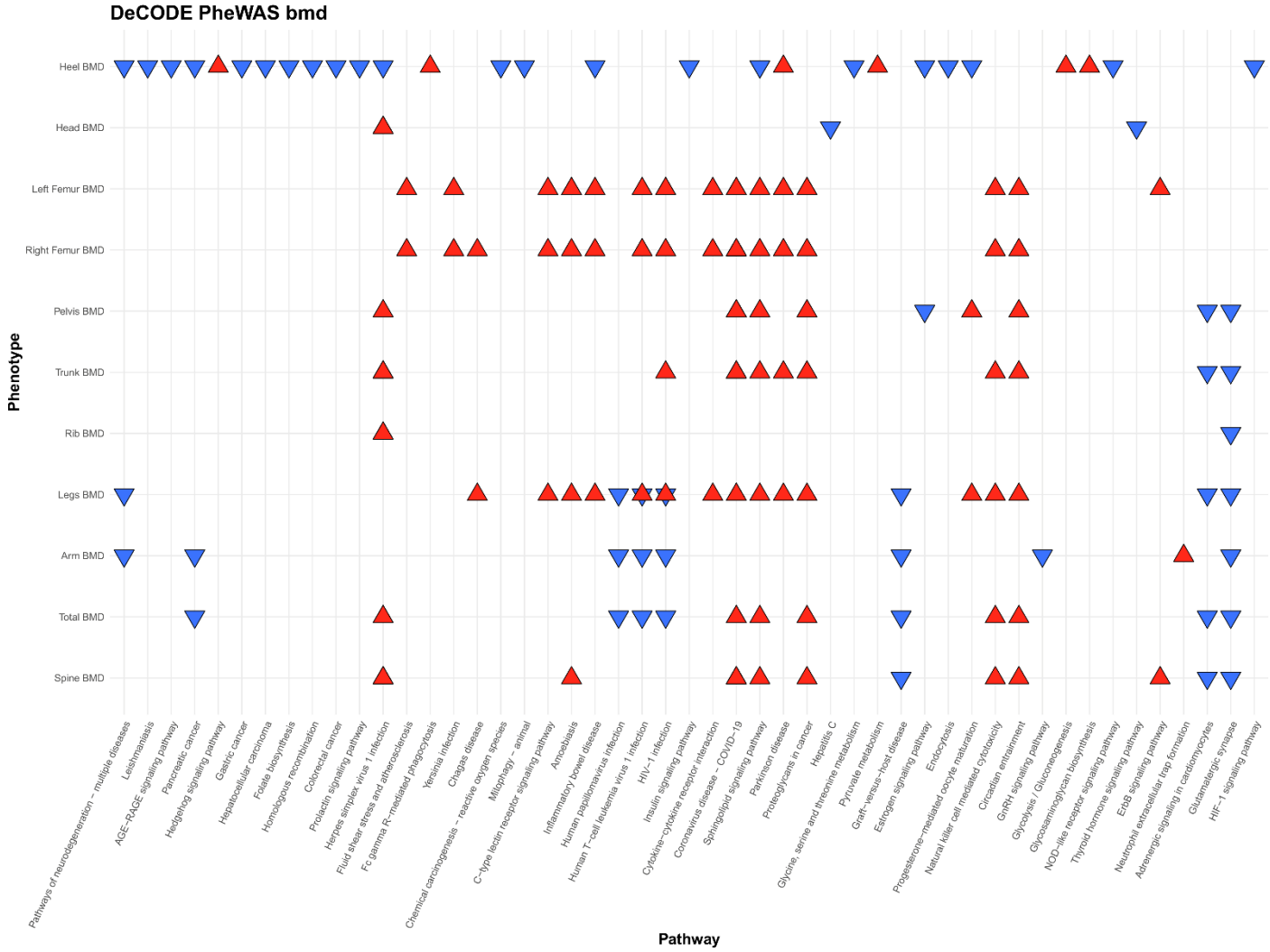


Figure 6.9: DeCODE PathWAS PheWAS bone mineral density (BMD). Significant pathway-phenotype associations which passed the sensitivity analysis, relating to bone mineral density traits. Blue arrows indicate that increased pathway activity leads to a decrease in the trait, red arrows implicate increased pathway activity with an increase in the trait.

DeCODE PheWAS serious illness
Phenotype
Age of cancer diagnosis
Number of cancers reported
Number of cancers (GR)
Interpolated age first illness
Diabetes
Number of serious illnesses (GR)
Age of first serious illness
Pathway
DeCODE PheWAS lung function traits
Phenotype
PEF
FVC
FEV1
Pathway

Figure 6.10: DeCODE PathWAS PheWAS anthropometry. Significant pathway-phenotype associations which passed the sensitivity analysis, relating to anthropometry traits. Blue arrows indicate that increased pathway activity leads to a decrease in the trait, red arrows implicate increased pathway activity with an increase in the trait

Figure 6.11: DeCODE PathWAS PheWAS blood cell counts. Significant pathway-phenotype associations which passed the sensitivity analysis, relating to blood cell count traits. Blue arrows indicate that increased pathway activity leads to a decrease in the trait, red arrows implicate increased pathway activity with an increase in the trait.

DeCODE PheWAS anthopometry
Phenotype
Pathway
DeCODE PheWAS blood cell count
Phenotype
Pathway

Figure 6.12: DeCODE PathWAS PheWAS serious illness. Significant pathway-phenotype associations which passed the sensitivity analysis, relating to serious illnesses. Blue arrows indicate that increased pathway activity leads to a decrease in the trait, red arrows implicate increased pathway activity with an increase in the trait

Figure 6.13: DeCODE PathWAS PheWAS lung function traits. Significant pathway-phenotype associations which passed the sensitivity analysis, relating to lung function traits. Blue arrows indicate that increased pathway activity leads to a decrease in the trait, red arrows implicate increased pathway activity with an increase in the trait.

6.3 Discussion

6.3.1 Pathway-phenotype discovery in SCALLOP

Using a combination of MVMR and association testing, we present a proof of concept for the PathWAS methodology. When applied using blood proteins and blood eQTLs, I show that it is possible to create pathway scores which can predict protein levels and which are reflective of the pathway functionality. I also demonstrate that it is possible to use these scores to link individual pathways to multiple and varying phenotypes in UK Biobank.

From 18 significant SCALLOP-derived pathways analysed in 60 different phenotypes in UK Biobank 123 significant associations between pathway and phenotype were discovered. Of these, 35 survived the sensitivity analysis of which a further 7 were potentially mediated through the end-point protein (either in an opposing or same direction) and so were discounted.

Of the 28 significant results which fully survived the sensitivity analyses, the associations relate to nine distinct pathway-end-point combinations: including the peroxisome proliferator-activated receptors (PPAR) signalling pathway (KEGG ID: hsa03320) with lipoprotein Lipase (LPL) as an end-point and hepatocellular carcinoma pathways (hsa05225) with ribosomal Protein S6 Kinase B1 (RPS6KB1) as the end-point.

Of note, the hepatocellular carcinoma pathway is one of many such KEGG pathways (denoted by the prefix hsa05-) which represent disease response or activation pathways. In these cases, the genes and reactions involved are often not specific human pathways but rather interactions which are modulated by the disease. As such these pathways will often incorporate elements from other major biological pathways. For example, the subset of the hepatocellular carcinoma pathways which has RPS6KB1 as an end-point is actually an element of the MAPK-signalling pathway.

In this study we found that this pathway has a positive association with height and a negative association with fat percentage (with a further relationship with BMI confounded by a significant

protein association with BMI). The relationship between MAPK-signalling and height may be well-founded as the MAPK-signalling pathway is heavily involved with cellular differentiation and proliferation[212]. Similarly, the protein RPS6KB1 is a serine/threonine kinase which phosphorylates the S6 subunit of the ribosome and promotes translation of specific proteins[213], making it an important component in the process protein synthesis. One could speculate that increased pathway activity may lead to increased cellular proliferation, protein synthesis and increased height.

Despite the relationship between this pathway and cell proliferation, previous studies have indicated that the relationship between MAPK-signalling and BMI or obesity is uncertain and potentially tenuous[214,215]. These findings suggest that this axis of the MAPK-signalling may be directly involved with growth but not in relation to obesity.

Another pathway in the significant SCALLOP results is the PPAR signalling pathway, which we find to have a positive association with spine bone mineral density, fitness workload and lung function traits along with a negative association with the interpolated age of first cancer diagnosis. The PPAR signalling pathway is involved in the transport and metabolism of fatty acids and lipids, agonists of the pathway have also been demonstrated to reduce the genesis of osteoclasts, cells which reabsorb and reduce bone density[216], providing evidence for the relationship discovered here. In contrast, it has been previously shown that increased exercise may upregulate PPAR signalling[217], while our results indicate that this relationship is likely to be more complex, with increased PPAR signalling allowing for increased fitness, potentially mediated by increased lung capacity. The relationship between PPAR signalling and cancer is more tenuous, particularly as this is a broad diagnosis of any cancer rather than specific. Recent studies have suggested shown the possibility of using both PPAR agonists and antagonists as chemotherapeutic agents, suggesting a contradictory role in increased PPAR signalling and cancer genesis[218,219].

6.3.2 Pathway-phenotype discovery in DeCODE

From the DeCODE proteomics results there were 2,108 significant associations which passed sensitivity testing, from over 9,000 initial significant results. While the results from the DeCODE analysis lacked the independent validation phase of the SCALLOP results, it is important to consider that the prediction of the protein levels is not an intended aim of the PathWAS methodology.

One of many significant signals in the DeCODE results is the relationship between the Legionellosis pathway (hsa05134), with the end-point interleukin 6 (IL6), and lymphocyte count ($P = 1.54 \times 10^{-118}$). As already discussed, the hsa05- pathways are those which are affected in response to a given disease, in this instance infection by *Legionella*. Here, the sub-component of the pathway which is specifically affected is part of the Toll-like receptor signalling pathway and is defined by KEGG as being involved in pro-inflammatory signalling and chemoattraction (the movement of cells in response to extracellular signals). The relationship between lymphocytes and inflammation is well-established[220].

Another interesting result is the significant association discovered between the lifespan survival trait, and the Toll-like receptor subset of the Toxoplasmosis pathway (hsa05145) and end-point MAPK11. Previous work has implicated both dysregulation and continuous activation of the Toll-like receptor pathway with increased morbidity and mortality associated with aging[221]. This result potentially further elucidates this relationship, with the studied pathway subset potentially being involved in continuous activation and reduced lifespan.

A third result is the positive association between a subset of the Parkinson's disease pathway (hsa05161) and end-point BH3 interacting domain death agonist (BID) with type-2 diabetes. BID is a regulator of apoptosis via the mitochondria (with the involvement of mitochondria in apoptosis contributing to the role it plays in Parkinson's disease), and there is evidence of mitochondrial dysfunction in type-2 diabetes[222,223], in the context of apoptosis. This relationship warrants further exploration, with increased pathway activity causing further expression and activation of downstream BID and thus potentially leading to increased incidence of diabetes through loss of beta cells.

Another result is observed with the alcoholic liver disease pathway (hsa0936) and three distinct end-point proteins. The end-points interferon alpha-8 (INFA8), IFNA10 and IFNA14, are each associated with the same branch of the pathway and are associated with fat percentage, weight, and both of these traits, respectively. This branch specifically is involved in the activation of Toll-like receptor signalling by lipopolysaccharides (LPS) through Toll-like receptor 4 (TLR4). This compliments existing literature which provides evidence for the relationship between increased LPS levels and increased obesity[224]. Furthermore, this pathway axis functions through downstream activation of interferon regulatory factor 3 (IRF3), which has also been implicated in driving obesity[225].

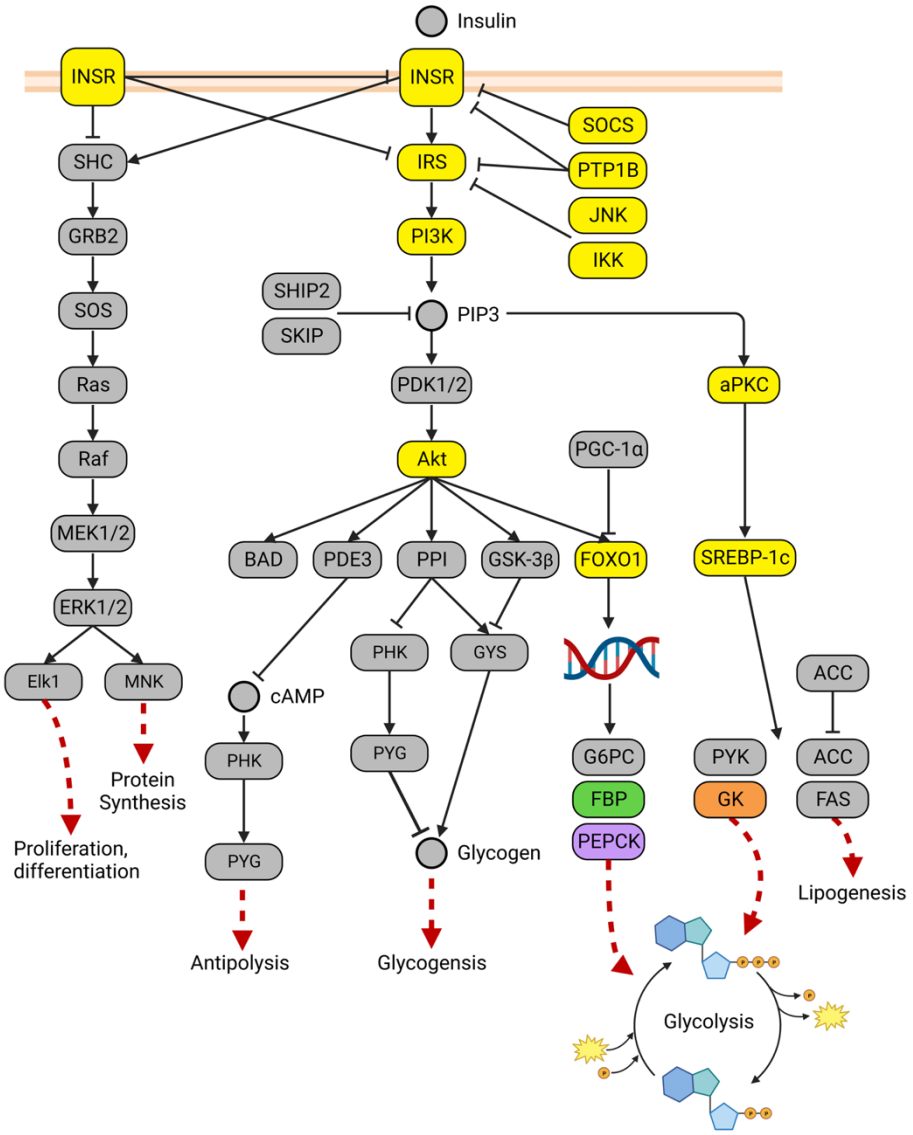


Figure 6.14. Excerpt of insulin secretion pathway from KEGG (hsa0910). An example pathway from the KEGG database with end-points in the DeCODE dataset, representing a subset of the genes involved in the insulin signalling pathway. Genes coloured in purple, green and orange represent three different end-point proteins used in measurement of the pathway. Genes in yellow are those which affect the end-points. Genes in grey are those not included in the pathway model. Due to the nature of KEGG nomenclature, here the FBP gene represents FBP2, PEPCK represents PCK1 and GK represents HK3. Also, vitally, the genes for PDK1/2 were not available in the eQTLgen dataset while those for SHIP2 and SKIP were unavoidably removed due to their interactions with the PIP3 molecule.

A final individual result is the insulin secretion pathway (hsa04910) with end-points fructose-bisphosphatase 2 (FBP2), phosphoenolpyruvate carboxykinase 1 (PCK1) and hexokinase 3 (HK3) (**Fig 6.14**, **Fig 6.15**). Both of which are end-points for the pathway involved in regulating glycolysis and also both show a negative association with BMI whilst also having no individual significant association between the proteins and BMI. This strongly suggests that increased glycolysis reduces

BMI and obesity, a result which has also been previously demonstrated[226]. This particular outcome, stemming from two separate chains of the same pathway seems particularly to support the validity of the results.

It is also curious to again see IL18 so highly represented in the pathway prediction stage as an end-point. This could suggest that IL18 levels are extremely tightly controlled such that it is far easier to predict the levels of the protein across individuals and populations. It also could indicate that the pathway models created for IL18 are predicting a far greater proportion of variance of the protein, and thus also are more easily predicted in others.

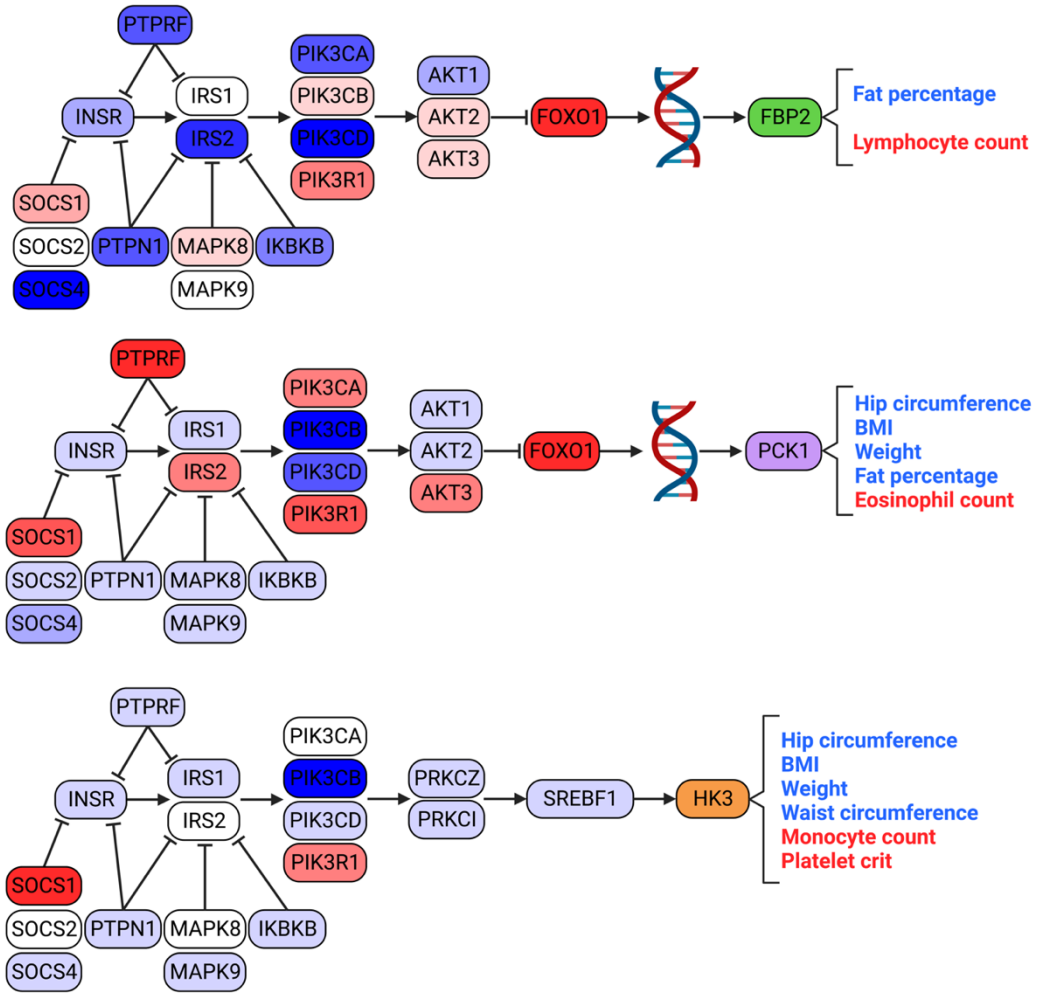


Figure 6.15: Refined pathway diagram. From the example pathway in **Fig 6.14** I show three refined pathway diagrams. In each case, I now only show the genes which have a downstream effect on the end-point protein. The genes are also coloured in accordance to their MR effect on the end-point (red is a positive effect and blue is a negative effect), with the intensity of the coloration stratified in accordance to the strength of the gene effect. We also show the significant traits which these pathway subsets are associated with (again, blue is negative, and red is positive). The end-points are coloured in relation to their place in **Fig 6.14**.

Another result of note is the previously mentioned result from **Section 6.2.4** of the two significant results for the MODY pathway associated with heart rate. Here the opposing direction of effect for the PathWAS results can likely be explained due to the end-points (IAPP and INS) used accounting for different sections of the overall KEGG pathway. In this instance the two axes may even be diametrically opposing each other as IAPP is a protein which inhibits insulin (INS) secretion[227]

and as such the pathway axis which relates to the function of IAPP would likely cause the opposite direction of effect to the pathway axis relating to INS function. It is also worth noting that the effect for INS is much strong than for IAPP, again potentially suggesting that the IAPP axis of the pathway has a lesser function due to its inhibition of the stronger INS effect. This highlights the importance of taking into account biology when analysing pathway results, as even two opposing directions of effect for the same pathway may actually make biological sense. It is worth noting however that the negative effect for the pathway with relation to INS and heart rate is opposed by literature which suggests that increased INS secretion should result in an according rise in heart rate[228]. This issue will be addressed in greater depth in **Section 6.3.3.2**.

6.3.3 Strengths and limitations of the PathWAS methodology

6.3.3.1 Lack of overlap between SCALLOP and DeCODE

An additional examination was conducted into the results of those pathways which overlapped between the SCALLOP and DeCODE results examining whether the same protein measurement with the same pathway produced equivalent results across both datasets. The results from this were inconclusive with varied levels of correlation between the PheWAS outputs. Likewise, there was no pattern of correlation seen in the $PRS_{Pathway}$ of these overlapping pathways. This is however possibly due to the differences between the SomaScan and Olink technologies used to measure the proteins as it has been previously discussed that not every protein measurement correlates well between these two[169,229].

The SomaScan platform utilises DNA oligonucleotides which bind to proteins, in a similar manner to antibodies, and can be used for the detection of very low quantities of proteins, and it is these bound "Somamers" which are quantified through DNA quantification. Meanwhile the Olink technology relies on antibodies binding to the proteins in question, with DNA nucleotides bound to the antibodies used for quantification. While both technologies have benefits and drawbacks, a lack of specificity has been reported with a set of the SomaScan proteins specifically[230]. Moreover the correlation between both technologies varies dramatically, consistently across cohorts, with a mean correlation of ~0.5[230] reported or <0.4 for the majority[231]. As such, this relative level of poor correlation may contribute to the poor correlation between the SCALLOP and DeCODE PathWAS results.

While only 16 protein-pathway models overlapped between the DeCODE and SCALLOP results (due to the prediction stage used with SCALLOP) it is still unexpected to see such an extreme divergence between both datasets. An element of poor correlation between the proteomics technology was to be expected, however the lack of correlation discovered was far more extreme than expected, and approaching a complete absence of any correlation in many cases (r^2 = ~0).

An additional experiment was conducted using the MAPK signalling pathway with endpoint HSPB1. The $PRS_{Pathway}$ for the SCALLOP and DeCODE omics had a correlation of -0.006 in UK Biobank. However, these PRS were re-created using a minor allele frequency (MAF) threshold of 0.01 and 0.05 respectively for the proteomics GWAS. This increased the correlation to 0.37 and 0.76 respectively. As such it is clear that much of the severe divergence in correlation between both the SCALLOP and DeCODE PathWAS analyses is potentially likely due to rare variants in both datasets causing population-specific effects.

6.3.3.2 Incorrect orientation of results

While many previously discovered significant associations between pathways and phenotypes were discovered, seemingly providing evidence for the validity of the PathWAS methodology, a large number of associations seemed to be oriented in the wrong direction. A primary example of this is the oestrogen signalling pathway and its associations with bone mineral density and white blood cell counts. With multiple end-points (KRT17, KRT18, EBAG9 and BCL2) this pathway demonstrated a significant negative relationship with both decreased bone mineral levels and decreased white blood cells (lymphocytes and neutrophils). This is in direct contrast to existing literature and known biological theory, which state that increased oestrogen signalling is protective of bone density and strength[232,233] and on their positive regulation of the immune system[234–236].

As such, this discovery stands in contrast to the validity of the PathWAS methodology. However, further examination suggests that a plausible cause for this discrepancy is the lack of alignment between the eQTLgen and DeCODE results. All the *cis*-SNPs from the DeCODE proteomics GWAS (within 500 kb of the start or end of the open reading frame) were filtered for SNPs with $P < 5 \times 10^{-8}$. These SNPs were aligned with those from the significant SNPs (FDR < 0.05) from the eQTLgen *cis*-SNPs dataset, with respect to their effect on specific genes. This provided an overlap of ~300,000 SNPs which were significant for the given gene in both datasets, an overlap of 1,172 genes. In ~40% of cases, the direction of effect of the SNP was opposite between the pQTL and eQTL (**Fig 6.16**).

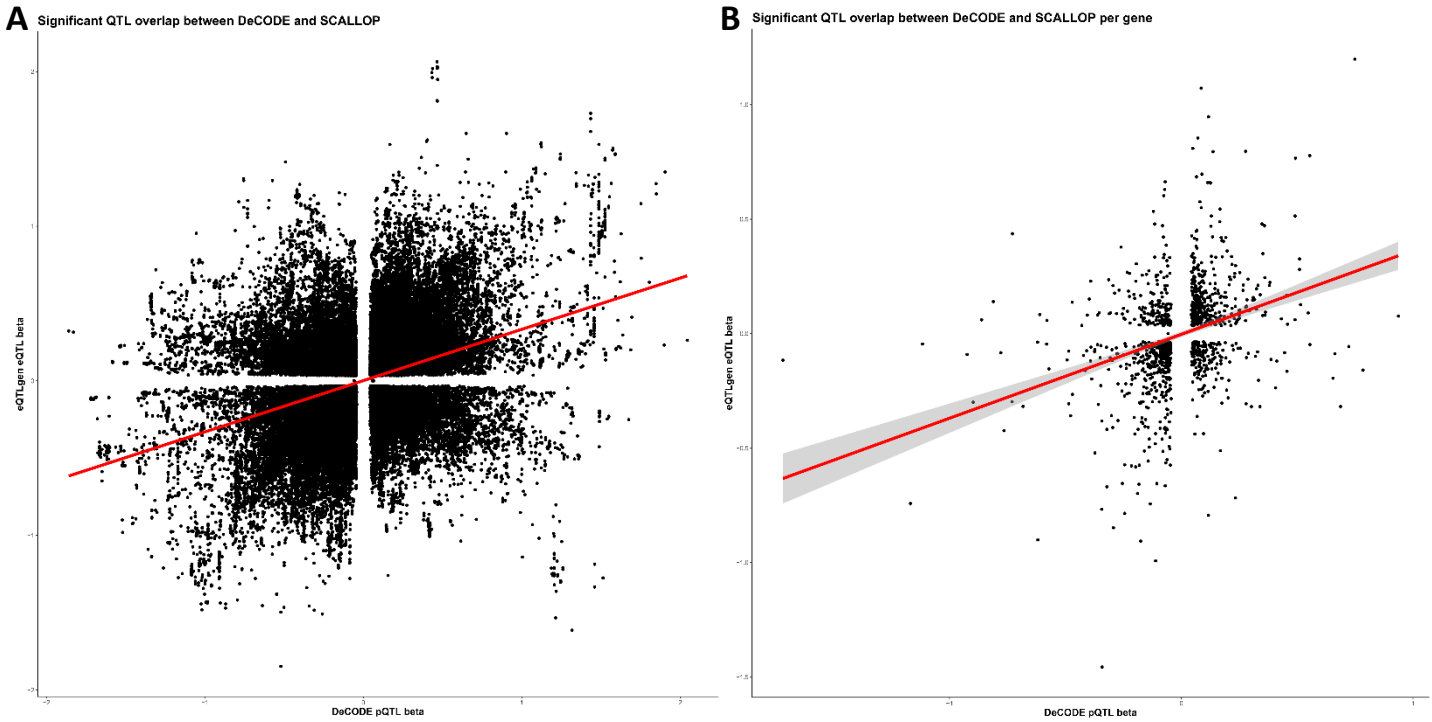


Figure 6.16. SNP overlap between DeCODE pQTLs and eQTLgen eQTLs. Panel A shows all SNPs which are significant in both DeCODE and eQTLgen datasets, an overlap of >1,000 genes and >300,000 SNPs. Panel B shows a randomly selected SNP from each individual gene from this subset of SNPs. The lines in red are lines of best fit, with standard error bars shown in both cases.

As such, this demonstrates the aforementioned poor correlation between eQTLs and protein measurements with a large number of significant SNPs for both datasets oriented in opposing directions. Taken at face value this would suggest an increase in transcription for a given gene would result in a decrease in translation for the protein product in 40% of cases (or vice versa). This is likely due to the tight layers of biological control in place, with elements such as negative feedback loops causing rapid degradation of proteins despite high levels of transcription. In practice however, if the transcriptomic and proteomic direction of effect are unaligned, then this would also result in an altered direction of effect for the PathWAS-derived pathway effect, potentially resulting in these incorrect directions of effect for associations.

It is possible that this issue of direction of effect could be responsible for the incorrect orientation of the effect of the $PRS_{Pathway}$ with phenotype in UK Biobank. Due to the significance of the effect, seen across multiple versions of the pathway with different end-points, this suggests that the inconsistent directionality of effect could be due to the discrepancy between eQTL and pQTL effects.

6.3.3.3 Missing results

One aspect of the results from PathWAS which was unexpected was the apparent absence of some results which were initially hypothesized to be discovered. I initially included the bone mineral density (BMD) phenotypes in the analysis in order to search for an association between these traits and MAPK-signalling (KEGG pathway: hsa04010) however, there were not even any results for this pathway (with multiple end-points in DeCODE and SCALLOP) even significantly associated with any of the BMD phenotypes. Likewise, the KEGG osteoclast differentiation pathway (hsa04380) was included in the analysis with several end-points and also produced no significant results associated with BMD phenotypes, a particularly confusing result given the importance of osteoclasts in bone degradation. A similarly missing association was between the p53 signalling pathway (hsa04115) and any form of cancer phenotype included in the analysis (number of cancers, age of cancer diagnosis or number of cancers reported), again unexpected due to the importance of p53 as a tumour suppressor.

In each instance there is potentially some unique biology at work which is preventing discovery of these associations, e.g. individual gene signals being more important in the BMD phenotype instead of the overall pathway, or the wrong axes of pathways being captured by the chosen end-points. In terms of the lack of results associated with cancer phenotypes, this may be due to the lack of specificity of the phenotypes as none of them refer to any specific cancer and instead to all cancers generically, which may be a misguided method of analysis. In each case this on serves to further highlight the importance of closely scrutinizing the specific biology of each pathway-phenotype relationship.

6.3.3.4 Proof of concept for methodology

The results seen here may in some cases warrant further investigation, but importantly also demonstrate the possibility for PathWAS to discover true associations between pathways and phenotypes. As such it is likely possible to be able to expand these results and discover novel associations with the methodology. By taking a broader and more holistic view of the links between biological pathways and complex traits we argue that this methodology may complement existing GWAS techniques, in combination with other post-GWAS analyses, to enhance our understanding of the consequences of dysregulation of biological pathways.

The method is also an attractive prospect in terms of precision medicine, due to the possibility of discovering novel and potentially druggable pathway relationships with complex disease. It has been shown previously that genetically-informed drug discovery has the potential to aid the process of developing new drugs, with drugs associated with GWAS targets more likely to pass phase III clinical trials[237,238]. By instead focusing on finding broader biological networks associated with complex

disease, it may be possible to further improve discovery of drug targets by instead allowing targeting of pathways rather than individual genes. Also, by combining signals into one overall pathway model it may be possible to discover commonalities between different complex diseases or indeed separate disease into unique sub-groups.

While PathWAS may offer a useful tool for analysis of complex traits, there are still several limitations to the methodology. A primary limit is the numerous assumptions made in the creation of the PathWAS models, including the treatment of biological pathways as discrete biological units rather than as metaphorical constructs and the use of specific "end-points" despite there not being any "end" for any biological network in reality.

The use of eQTLs to measure levels of the various acting genes in pathways is also not ideal. It has been previously well-established that eQTLs do not predict related protein-product levels well. So, given that we assume proteins and metabolites to be the major functional units within pathways this means that using eQTLs will likely not predict the pathway functionality as accurately as we would like. The argument however for using eQTLs instead of existing pQTL datasets is that at present there are more varied eQTL datasets in larger sample sizes, and more importantly, even the largest proteomics datasets only examine a smaller pre-selected group of proteins unlike transcriptomics which can examine almost all mRNA. Therefore, by using datasets like eQTLgen in favour of pQTLs we make a trade of accurately predicting individual genes for having measurements for more of the genes within different pathways.

Another of the primary limitations of PathWAS methodology is that the vast majority of biological pathways are intercellular and so transcriptomics and proteomics from whole blood will not provide an accurate picture of cell-type specific reactions. Ideally, to test for different phenotypes relevant to different tissues this would utilise single-cell transcriptomics and proteomics. Currently, some tissue-specific eQTL datasets do exist (like those of the GTEx consortium) but low sample sizes may then make power another issue.

Additional considerations for the PathWAS methodology include an acknowledgement that it relies heavily upon accurate and valid curation of pathway databases. While the method is capable of finding associations between pathways and phenotypes, it relies heavily on the KEGG database for creating pathway gene lists and thus will transfer incomplete knowledge from the database to the method. This highlights the issue that for more complex and thorough analysis of biological pathways, we need better tools and complete databases of pathways. While these limits must be acknowledged, it is also important to be aware that these are things which are being overcome. A greater focus on single-cell transcriptomic analyses will result in the existence of large single-cell QTLs in the near future, allowing a much more cell-specific form of PathWAS analysis, which we argue would ideally suit our methodology. Pathway databases also allow the constant refinement and

parsing of pathway knowledge and so allowing incorporation of the various pathways into the methodology will mean that the PathWAS method will only become more accurate over time. Thus, when combined with single-cell proteomics and transcriptomics, we hope it will become an attractive tool for the analysis of biological pathways and their relationship with complex traits.

Further considerations include more specific biological aspects such as rate-limiting steps and gene thresholds which could limit the validity of the method. This would mean that the assumption of a direct association between gene transcription levels and protein activity would be invalidated.

6.4 PathWAS conclusions

The results shown here in SCALLOP appear to demonstrate a clear proof-of-concept for the PathWAS methodology, affirming previously discovered pathway-phenotype relationships in the strongest pathway associations. A further analysis with all of the pathways examined from the DeCODE proteomics further expands upon this with many thousand significant relationships seemingly driven by pathways rather than individual gene-trait associations.

Whilst there is still much work which could be done on expanding and improving the PathWAS methodology, I believe this work presents a valid and useful method for usage in the field of precision medicine when combining multi-omics and genotype in the search of the relationships between genotype and complex traits.

As such, much of the work presented here has been incorporated into an R package available on GitHub for usage by other academics (https://github.com/Sabor117/PathWAS). This package is capable of estimating pathway functionality based on end-point omics in the same way as described through this chapter. I believe I have demonstrated the power of this package through creation of pathway PRS in multiple omics datasets which have then been used to find thousands of associations between biological pathways and complex traits in UK Biobank. It remains possible that this methodology may act as a compliment to GWAS and aid in discovery of novel pathway-phenotype relationships.

Chapter 7

Discussion and conclusions

7.1 Precision medicine

Precision medicine is a constantly evolving field with the broad aim of being able to tailor treatments and targeted interventions to individuals based on their unique physiology. In particular utilising genetic predictors of illness and complex disease may allow for the possibility of better targeted medicines, finding individuals who would best respond to different treatments or aiding in explaining why other don't respond as well.

In terms of determining why individuals with the same disease may respond differently to the same treatment, then one possible reason is the difference in activity of different biological pathways within those individuals. With subtly varying levels of different gene products and proteins, this would lead to different levels of individual activity which may be inconsequential on the scale of individual gene interactions but across multiple cumulative interactions and chemical reactions may lead to an overall larger biological effect.

In this project I have demonstrated the capacity of the PathWAS methodology of being able to predict known pathway-phenotype interactions as described in **Chapter 6** and so this raises the possibility of being able to discover novel relationships. This can be done without any prior knowledge of individual gene-to-phenotype relationships and in fact may even aid in the discovery of those for which there was not enough power in an original GWAS to discover all the related loci.

This method also has the capacity of extending known relationships between individual genes and complex traits by examining broader networks connected to the gene and searching for any differences in effect seen.

All of this has the potential to eventually be expanded upon and lead to improved therapies based on the individual activity of a given pathway in different patients and could also aid in targeting drug discovery in the running of clinical trials where efficacy of drugs may rely more on these varying pathway activity levels.

However, while the methodology may have promise for these goals, it likely still falls short of achieving them due to a number of important concerns.

7.2 Best practices

In terms of creating the scores which can best predict pathway function, one of the main areas I have attempted to address throughout this project is using the best sources of data and the best methodology for creation of the $PRS_{Pathway}$. However, while the project acts as a clear proof-of-concept for the validity of the PathWAS method, there are areas which could benefit from future improvements. I will now detail several of the areas where improvements were made to the methodology and where drawbacks still lie.

7.2.1 Importance of power

Creation of the best $PRS_{Pathway}$ relies on the best sources of data which can predict the function of the pathways in question. Two elements we focused on for the proof-of-concept were sample size and number of gene measurements. In terms of creation of the $PRS_{Pathway}$ from the constituent PRS_{Gene} in the SCALLOP data set, the pathway models which performed the best in the prediction data were primarily those which included multiple gene measurements. As well as this, when the same pathways were examined via elastic net penalised regression Mendelian randomisation, the elastic net models broadly performed either equally or worse than the LASSO MVMR models. The elastic net models excluded all but the best PRS_{Gene} from the models based on the strength of the associations, so in theory may have reduced noise from the models. However, it also removed many of the pathway models completely, suggesting that by removing gene signals from the pathway models this actually made them worse. Similarly, by switching from GTEx eQTLs to those from eQTLgen, thus increasing the number of available gene measurements, this also improved the $PRS_{Pathway}$ scores.

We theorise these differences are due to the same reason in that inclusion of more genes in a pathway model and a greater sample size in the measurement of those genes both equal an increase in power of detection for the model. This makes theoretical sense at least that a complete model for a pathway would include a measurement of activity for every gene within it. So, datasets with a greater number of measured genes will equal a more complete picture, while a larger sample size for these genes will result in a more accurate representation of expression for the given genes and thus a more accurate representation of their contribution to the overall pathway effect.

7.2.2 Importance of specificity

While we selected our data based on the requirement of power, focusing on sample size and number of measurements, this was at the sacrifice of other benefits offered, for example, by GTEx or the INTERVAL pQTLs. As discussed previously, it is likely there is merit to examining pathways on a

tissue or even cell-type specific basis and as such tissue-specific eQTLs in a large dataset would be an invaluable resource for PathWAS. Also, as we aim to create PRS_{Gene} which estimate a genes effect within a pathway based on gene expression levels, usage of proteomics instead of transcriptomics would certainly provide a better representation of abundance.

While power is important in detecting differences in effects, in some instances the specificity of a measurement could be far more important. For example, genetic effects on a pathway may be far more important in specific tissues or even cell-types for given traits. While traits such as height may be more omni-genic, certain phenotypes will unquestionably have larger effects in specific tissues. Clear examples include diseases such as cancers, e.g., colorectal cancer, much of which will be localised to intestinal tissues. In this case, there will clearly be tissue or cell-type specific effects. For example, the Wnt-signalling pathway is a vitally important pathway in canonical colorectal cancer, but effects of genetics on this pathway may not be associated with the cancer phenotype unless expression data from intestinal tissue or cells are used.

In the same way that it makes sense that increased power and increased numbers of measurements will improve the pathway models created, increased tissue and cell-type specificity also makes logical sense as an improvement for the methodology. This is especially true given that the vast majority of pathways studied are intracellular and so it makes inherent sense that gene expression measurements would be better from tissues and cells relevant to a trait of interest.

While not possible for this project specifically, in the coming years it would be extremely interesting to see work conducted on using the PathWAS pipeline in single-cell QTL data sets, of which a growing amount are becoming available in the near future.

Likewise, while single-cell proteomics or even tissue-specific proteomics are some years from having large enough data sets, the incorporation of these into the models would very likely improve any predictions made for more specific traits and diseases.

One aspect of this which would be intriguing is to see what, if any difference, would be made by mixing pQTL with eQTL data sets for use in the PRS_{Gene}. Specifically, where measurements for pQTLs are available for a given gene within a pathway, they could be used for the expression PRS in place of eQTLs and then combined with that from eQTLs, allowing for a mix of both specificity and increased gene number.

The use of pQTLs is another area which would theoretically improve the overall accuracy of PathWAS due to the concept that the pathway scores are caused by the cumulative effect of gene products, and then eQTLs often do not predict their associated protein[83]. Therefore, utilising pQTLs instead of eQTLs would theoretically improve the scores, but until proteomics is possible on multiple tissues or specific cell types, scRNA-seq might better capture the pathway effects.

Alternatively, incorporation of a method like the attempted creation of the multi-tissue PRS_{Gene}, which could more accurately reflect protein levels based on transcriptomics would be ideal for overcoming this known limitation in integrating transcriptomics with proteomics.

7.2.3 Pathway specificity

Another aspect of importance that was only slightly addressed by this project is the issue of pathways not truly being biological entities but instead a representation of sequences of chemical reactions. As such, it is important to use the pathways which would best reflect the reality of gene interactions within a cell.

We sought to address this by using the KEGG database[109], which is manually and carefully curated, and also by refining the pathways selected to those genes which are deemed to interact with the selected end-point proteins.

However, expanding upon the knowledge-base used by the methodology would likely improve the results discovered. On one hand it would be extremely interesting to see individually curated pathways used with the PathWAS method against specific GWAS of traits of interest. However, it has also been a consistent goal of the creation of the R package to extend the usage of the pathways beyond the KEGG database.

More specifically I have attempted to incorporate both the WikiPathways[151] and Reactome[150] databases into the PathWAS pipeline, however here I was restricted by the form of the data of both databases. What is necessary for PathWAS is a table of the nodes and edges in a pathway, and vitally it is necessary for this table to contain directionality (I.e. from GENE1 to GENE2). This proved to be significantly more challenging to extract from the XML documents of both WikiPathways and Reactome than the relative ease of using the KEGG database[173]. However, incorporation of these additional databases into the PathWAS methodology and package would certainly be extremely useful to if a viable option for disentangling them in an automatic way became available.

7.2.4 Protein-protein interactions and pathway scoring

Similarly to improving pathway access, another element which may have the capacity to improve the results of PathWAS would be to rely on protein-protein interaction (PPI) data instead of using pathway databases[239]. Alternatively, it could be interesting to see results of PathWAS informed by PPI in some way through databases such as StringDB[240] or IntAct[241], to see whether this produces improved results.

This would fundamentally alter the methodology of PathWAS where currently the pipeline relies on usage of pre-defined pathways, either manually curated or from pathway databases, to create a list of genes which would influence a given end-point. Instead, by using PPIs genes upstream of an end-point would be defined by discovered biological relationships, potentially providing a more holistic set of genes which influence a given end-point.

In a similar vein, improving the pathways used through use of an external pathway scoring tool, such as the pathway scoring algorithm (PASCAL)[242], may also in some way improve or further validate the results used by improving the pathway connections used in the models. This is a method of scoring pathways based on genes near significant loci discovered by GWAS, in a similar method to gene-set enrichment, however with greater possibility of discovering relevant biological pathways. It may therefore be possible to overlap this with the PathWAS methodology and define a set of pathways for examination in relation to a complex trait, assisting in confirmation of plausible relationships.

Incorporation of these additional methods of examination could lead to greater accuracy in the results of the PathWAS method.

7.2.5 The best tools for polygenic scoring

Throughout the project I have tried multiple methods for creating PRS for gene expression. It is well accepted in literature that these methods which take into account the LD structure of SNPs outperform more traditional methods of creating PRS which prune SNPs based on P-values and the clump[140]. Therefore, from the outset the aim of the project was to use PRS-creation tools which used LD-based algorithms.

The first sets of PRS were created based on the GTEx eQTLs using LDpred[139] and polygenic weights derived from Predi-X-Can[147] (in the form of the PredictDB weights). Initial testing and comparison of these scores by using them to predict the associated protein-product in ORCADES was, however, inconclusive. The Predi-X-Can-derived scores had a higher mean r^2 when predicting all the proteins which overlapped between the expression PRS and the proteomics, while the LDpred-derived scores had a slightly higher mean r^2 when comparing only those scores which overlapped between the two sets of PRS_{Gene}. As the predictive capability of both techniques seemed to be essentially equal, and due to the high computational requirements of LDpred, for the initial stages of the project using GTEx, I primarily focused on using the PredictDB data set.

However, when the work expanded to the use of the eQTLgen *cis*-eQTLs, as the PredictDB data was restricted to the GTEx eQTLs and due to the introduction of LDpred2 I switched to using LDpred2 in the creation of PRS for the SCALLOP dataset[140].

LDpred2 was used in two subsequent iterations of PathWAS, creating PRS for ~2000 genes from eQTLgen in order to create pathway scores for all of the SCALLOP proteomics pathways and 1,109 pathways from DeCODE (a subset of the complete DeCODE dataset). This also included the creation of a number of scores using PRsice-2[137] due to a low h^2 for a small subset of the gene summary statistics.

However, further iterations of the LDpred2 algorithm, made to improve the creation of PRS in a whole-genome context, made the package untenable for use in the creation of expression of PRS in subsequent iterations of PathWAS of the entirety of 2,922 pathways in the DeCODE data.

In order to maintain consistency across the creation of all PRS for gene expression, I finally switched to using PRS-CS, which utilises a Bayesian regression method for shrinking prior SNP effect sizes. The performance of PRS-CS compared with LDpred2 is uncertain, with Privé *et al.* (2020)[140] stating that LDpred2 outperforms PRS-CS while Zhang *et al.* (2022) state that PRS-CS outperforms LDpred2 at higher sample sizes[211]. More recently, however, in the context of expression PRS, Pain *et al.* (2022)[243] tested PRS creation in eQTLgen utilising multiple PRS creation methods, demonstrating a benefit to using PRS-CS even over LDpred2. This may be due to the small number of SNPs available for use and LDpred2 performing better in a whole-genome context. This remains contentious however, as others have also provided evidence for the superiority of LDpred2 (and other techniques)[244]. Due to this uncertainty I cannot make any conclusive statements about the best PRS-creation method, however it seems likely that both LDpred2 and PRS-CS are valid options.

Unfortunately, exhaustive testing of various PRS was beyond the scope of this project as in theory the PathWAS methodology and package should work independently of the methodology used in creation of the expression PRS. However, it has raised the question about which PRS creation tool is the best, which remains uncertain in the field. For the purposes of this project, however, it was assumed that given the somewhat smaller sample size of eQTLgen (relative to Biobank scale GWAS), LDpred2 would be the ideal tool for creating PRS_{Gene}. However, due to the inconsistent nature of the PRS created with LDpred2, we opted for the more consistent PRS-CS method.

7.2.6 Improvement in sensitivity analysis

The change in methodology of the sensitivity analysis is one aspect that improved considerably between early and later version of PathWAS. The initial attempt, utilising a comparison of P-values between individual PRS_{Gene} and $PRS_{Pathway}$ was almost certainly misguided and not a good representation of whether pathway-phenotype relationships were being driven by the pathway or by individual associations.

The conversion to usage of the Z-score testing within the bounds of the Leave-1-Out analysis not only improved the quantity of results discovered but likely also improved the quality of them as well. By examining each $PRS_{Pathway\text{-}GeneN}$, as well as testing for any pathway-phenotype associations driven by individual genes, this had the added benefit that this acted a form of test for horizontal pleiotropy across the pathway, as the effect would change drastically if SNPs directly involved with the outcome were removed.

The inclusion of testing for relationships between the $PRS_{Protein}$ and the phenotype also improved the robustness of the results, as each $PRS_{Protein}$ was weighted by the effect on a given protein and so it was possible that any discovered relationships would have been entirely to do with the interaction between the end-point and phenotype.

The switch to the L1O analysis thus represents an important improvement in the robustness and accuracy of the results discovered.

7.3 Improvement of results

Following the initial results of the PathWAS v.1 and then comparing these with the results of the PathWAS v.2 when using the Olink-derived proteomics provides the clearest example of the improvements in methodology. When first examined using eQTLs from eQTLgen, creating PRS_{Gene} with LDpred2, across >300 pathways with scores in 49 different tissues, only 7 total significant models were discovered (utilising both linear and elastic net regression models). This was then followed by several significant improvements in methodology and the data sources used.

Instead of using just proteomics studied in ORCADES and Vis, we now had access to the SCALLOP proteomics consortium data. PRS-CS was used instead of LDpred2 for creation of the PRS_{Gene}, and MVMR was used instead of regression. This lead to the discovery of 18 significantly predicted pathways in the ORCADES dataset. This increase in discovered significant pathways is also despite only using one "tissue" source for the pathways (using the "Complete" pathway instead of any of those pruned by the GTEx TPMs). As such I propose that had the tissue-specific pathways also been included this would have included a similar increase in significant models.

An early version of the PathWAS v.2 also continued using the LDpred2 PRS_{Gene} and here we discovered 12-14 significant pathway associations with MVMR, some of which were also discovered with PRS-CS. This seems to suggest that the changes made to the PRS creation may not have been the driving force in the discovery of new significant pathway models. Instead, the more vital change was the switch from regression models to Mendelian randomisation and the vast increase in power lent by the SCALLOP data.

7.4 Pathway-phenotype discovery

As discussed in **Chapter 6** we found over 2,000 significant pathway-phenotype associations across both proteomics datasets which passed the sensitivity analysis. Many of these were validated by existing literature or were part of known biological theory. The results from the independently validated SCALLOP proteomics appear to act as a proof-of-concept of the method, followed by expansion to the DeCODE dataset, allowing discovery of many more relationships. Particularly in combination with the sensitivity analysis, these results seem to link genuine causal relationships between many pathways and complex traits.

The volume of significant results was in fact unexpected, even given a starting point of over 2,000 pathway-end-point combinations to examine. A considerable challenge was even parsing the results of the DeCODE data, and much more work could be performed here on examining individual relationships of interest.

Those relationships which I extracted and discussed make up only a tiny number of those seen, and there is in fact a number of other relationships which also are worth consideration. It may be important to further-examine those pathway-phenotype associations which passed the L1O analysis but also had a significant protein association with phenotype. While this would suggest that the pathway itself is affecting the phenotype in some way, it cannot be ruled out that the relationship is due to the scaling with the protein. However, individual and expert examination of the individual pathways here could further elucidate this.

One aspect which would be extremely useful for the PathWAS methodology would be to directly compare the results generated by the method with other more traditional methods of pathway-enrichment analysis (E.g. FUMA[175]) or other forms of pathway analysis like pathway scoring with PASCAL[242] or prediction of gene functions using DEPICT[245]. As the analyses conducted here were primarily performed as proof-of-concept I have not highlighted novel discoveries in favour of selecting associations which have been previously defined. However in future, discovering novel pathway-phenotype associations with specific GWAS would be entirely plausible with the PathWAS methodology and then comparison with methods which work backwards from GWAS to pathways would be extremely interesting to analyse.

7.5 Remaining limitations

One of the primary limitations of the PathWAS methodology discussed in **section 6.3.3.2** is the orientation of some pathway-phenotype results being in opposition to known biological theory. I hypothesise that this inconsistency could be primarily due to the poor correlation between the eQTL and pQTL SNP effects. Due to the usage of the eQTL effect sizes and pQTL effect sizes in the MR

framework, this would mean that in many cases where there are increased levels of transcription and thus an assumed increase in gene expression, the associated end-point protein may have effects oriented in the opposite direction, resulting in a flipped phenotype-pathway association from the $PRS_{Pathway}$.

While an element of poor correlation between eQTL and pQTLs is to be expected, having ~40% of signals oriented in the opposite direction was unexpected. However, this proportion of lack of overlap was replicated in an additional experiment using similar datasets by Zhijian Yang and Dr. Xia Shen. This discrepancy appears to highlight the previously reported poor relationship between mRNA and associated protein abundance but at an unexpected scale.

This represents a final challenge in the PathWAS method and solving it goes beyond the scope of this thesis. In particular, unpicking the relationship between mRNA and protein levels is an exceedingly complex topic and may require incorporation of more single-cell sources of data alongside newer statistical techniques to solve. This is assuming that it is possible to generate scores from eQTLs which predict protein level that do not require case-by-case analysis due to varying and specific biology.

While currently lacking the necessary methods to truly overcome this discrepancy, one technique that may potentially provide a solution in the interim, would be the use of other omics technologies either for the gene estimates in the MVMR or for the outcome as the proxy of functionality. With a large-enough proteomics dataset it may be possible to use pQTLs as the IVs in the MVMR. Alternatively, and more realistically, it already would be possible to use either metabolomics or transcriptomics as the end-points for the pathways. While we use a measured protein due to protein due to protein level being more representative of phenotype, in practice eQTLs from an independent dataset would function just as well, with the pathway effect being scored by how much it effects transcription of a given end-point gene.

7.6 Further improvements of the methodology

There remains three primary ways in which the PathWAS methodology could be easily improved in the near future, beyond the inclusion of more powerful and specific data sources. These are aspects which I will continue to work on as I will maintain the PathWAS R package beyond the duration of my PhD.

The first step will be the refinement of the data used with MAF filters in order to improve the reproducibility of the $PRS_{Pathway}$ between the omics sources used as the end-point proteins. The current overlapping results between the Olink and SomaScan technologies displayed very little correlation,

likely majorly driven by population-specific rare variants in both datasets. For publication purposes, this will be corrected and the results re-examined.

Another major element will be the possible incorporation of the Reactome or Wikipathways databases into the search aspects of PathWAS, to expand the pathways which can be examined and compare pathway results between varying sources. It is possible that combined usage of the R packages *jsonlite* and *httr* could allow reading the Reactome pathways in an automatic and reproducible manner.

Finally, the phenotypes used in the PheWAS stage of the PathWAS method were pre-selected for a variety of reasons. Blood cell traits were selected due to the omics sources both being from blood, anthropometry traits were selected due to their broad interest in genetics, bone mineral density traits selected in order to search for known existing relationships. However, many of the other traits were selected primarily to have a varied spread of different traits and due to personal interest. In other words, the traits were not selected through any systematic manner. As such a possible and useful expansion of the PathWAS method could be the incorporation of techniques such as PhenoScanner which cross-references genetic variants with known phenotypes to find possible associations[246], which could potentially provide plausible targets for each pathway model in a systematic manner.

Beyond this, the next stage in testing the PathWAS method would be in usage along with a specific pathway or searching for pathway relationships with a specific GWAS of a given trait in order to search for novel associations.

7.7 Conclusions

The PathWAS methodology provides a novel and holistic method for the analysis of biological pathways and their relationship with complex traits. By analysing pathways and their relationships with complex from the bottom-up rather than in the method of traditional gene-set enrichment analyses, I believe we have the potential to overcome many of the shortcomings of this method. Specifically, the potential lack of power to discover all variation associated with a trait or the possibility of defining incorrect gene targets for significant loci would both be bypassed by selecting all genes and variants *a priori* with PathWAS.

In doing so the method has the capacity to discover novel relationships or to clarify and elucidate known ones. In this work I have shown that PathWAS has the ability to create scores predictive of pathway function through the use of a proxy of functionality and that these scores can be used to predict the associated proxy in independent datasets. I also show that the $PRS_{Pathway}$ generated with this technique are capable of finding many significant associations in UK Biobank, providing many intriguing relationships to examine.

This method has the benefit of utilising multiple sources of genetic and omics data which are all publicly available and as such can be reproduced by other academics in search of pathway-trait relationships based on a trait or pathway of their own interest. It thus also has the added benefit of further leveraging existing data, as it is possible that in many such cases there is still a great deal of knowledge to be extracted from them.

Development of the L1O methodology and application of MVMR also means that PathWAS-generated scores are potentially genuinely-causal pathway-phenotype relationships, with no evidence of horizontal pleiotropy and reduced confounding influences. This further improves the robustness and viability of the pathway-phenotype relationships discovered.

While there remain limitations of the method, in particular the surprising lack of correlation between $PRS_{Pathway}$ for the same pathway between two different omics technologies, and overcoming the inconsistent direction of effect between eQTLs and pQTLs, the method still seems to show clear promise in discovery of pathway-phenotype associations.

This same potential could mean that PathWAS could have a multitude of uses within the field of precision medicine. Finding pathways associated with complex diseases could allow better understanding of the mechanism of disease aetiology and progression, potentially even allowing differentiation between disease severity and subclasses. With this possibility it would also potentially be a valuable tool in the search of drug targets for specific diseases, instead of examining genes which are associated with GWAS hits, PathWAS could allow for examination of pathways instead. By examining pathways instead of on a gene-by-gene basis, it may be easier to extract viable drug targets as it may be more apparent what the driving causes are of the trait.

The method I have developed here is publicly available via GitHub and ready for application with additional data and already there are many avenues for possible further exploration. With the work here acting as a proof-of-concept, I believe I have demonstrated the validity of the methodology.

Bibliography

1. Jelenkovic, A. *et al.* Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. *Sci. Rep.* **6**, 28496 (2016).
2. Sahu, M. & Prasuna, J. G. Twin Studies: A Unique Epidemiological Tool. *Indian J. Community Med. Off. Publ. Indian Assoc. Prev. Soc. Med.* **41**, 177–182 (2016).
3. Friedman, N. P., Banich, M. T. & Keller, M. C. Twin studies to GWAS: There and back again. *Trends Cogn. Sci.* **25**, 855–869 (2021).
4. Schwabe, I., Janss, L. & van den Berg, S. M. Can We Validate the Results of Twin Studies? A Census-Based Study on the Heritability of Educational Achievement. *Front. Genet.* **8**, (2017).
5. Yengo, L. *et al.* A Saturated Map of Common Genetic Variants Associated with Human Height from 5.4 Million Individuals of Diverse Ancestries. *bioRxiv* 2022.01.07.475305 (2022) doi:10.1101/2022.01.07.475305.
6. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
7. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
8. Young, A. I. Solving the missing heritability problem. *PLOS Genet.* **15**, e1008222 (2019).
9. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).
10. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
11. Garbade, S. F. *et al.* Allelic phenotype values: a model for genotype-based phenotype prediction in phenylketonuria. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **21**, 580–590 (2019).
12. Roos, R. A. Huntington's disease: a clinical review. *Orphanet J. Rare Dis.* **5**, 40 (2010).
13. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).

14. Holland, D. *et al.* Estimating Effect Sizes and Expected Replication Probabilities from GWAS Summary Statistics. *Front. Genet.* **7**, (2016).

15. Stringer, S., Wray, N. R., Kahn, R. S. & Derks, E. M. Underestimated Effect Sizes in GWAS: Fundamental Limitations of Single SNP Analysis for Dichotomous Phenotypes. *PLOS ONE* **6**, e27964 (2011).

16. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).

17. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).

18. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).

19. Wijmenga, C. & Zhernakova, A. The importance of cohort studies in the post-GWAS era. *Nat. Genet.* **50**, 322–328 (2018).

20. Montaner, J. *et al.* Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. *Nat. Rev. Neurol.* **16**, 247–264 (2020).

21. Paananen, J. & Fortino, V. An omics perspective on drug target discovery platforms. *Brief. Bioinform.* **21**, 1937–1953 (2019).

22. Savino, R., Paduano, S., Preianò, M. & Terracciano, R. The Proteomics Big Challenge for Biomarkers and New Drug-Targets Discovery. *Int. J. Mol. Sci.* **13**, 13926–13948 (2012).

23. Li, B. & Ritchie, M. D. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Front. Genet.* **12**, (2021).

24. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).

25. Thompson, D. J. *et al.* UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. 2022.06.16.22276246 Preprint at https://doi.org/10.1101/2022.06.16.22276246 (2022).

26. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).

27. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).

28. Armstrong, B. NHS launches new polygenic scores trial for heart disease. *Genomics Education Programme* https://www.genomicseducation.hee.nhs.uk/blog/nhs-launches-new-polygenic-scores-trial-for-heart-disease/ (2021).

29. Sun, L. *et al.* Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLOS Med.* **18**, e1003498 (2021).

30. O'Sullivan, J. W. *et al.* Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation* **146**, e93–e118 (2022).

31. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32**, 650–654 (2002).

32. Haga, H., Yamada, R., Ohnishi, Y., Nakamura, Y. & Tanaka, T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. *J. Hum. Genet.* **47**, 605–610 (2002).

33. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

34. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).

35. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).

36. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).

37. Jang, S.-K. *et al.* Rare genetic variants explain missing heritability in smoking. *Nat. Hum. Behav.* **6**, 1577–1586 (2022).

38. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).

39. Auwerx, C. *et al.* The individual and global impact of copy-number variants on complex human traits. *Am. J. Hum. Genet.* **109**, 647–668 (2022).

40. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

41. Mathieson, I. The omnigenic model and polygenic prediction of complex traits. *Am. J. Hum. Genet.* **108**, 1558–1563 (2021).

42. Han, X., Aslanian, A. & Yates, J. R. Mass Spectrometry for Proteomics. *Curr. Opin. Chem. Biol.* **12**, 483–490 (2008).

43. Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **7**, 681–685 (2010).

44. Nakayasu, E. S. *et al.* Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. *Nat. Protoc.* **16**, 3737–3760 (2021).

45. Mirauta, B. A. *et al.* Population-scale proteome variation in human induced pluripotent stem cells. *eLife* **9**, e57390 (2020).

46. Solomon, T. *et al.* Identification of Common and Rare Genetic Variation Associated With Plasma Protein Levels Using Whole-Exome Sequencing and Mass Spectrometry. *Circ. Genomic Precis. Med.* **11**, e002170 (2018).

47. Ahsan, M. *et al.* The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. *PLOS Genet.* **13**, e1007005 (2017).

48. Benson, M. D. *et al.* Genetic Architecture of the Cardiovascular Risk Proteome. *Circulation* **137**, 1158–1172 (2018).

49. de Vries, P. S. *et al.* Whole-genome sequencing study of serum peptide levels: the Atherosclerosis Risk in Communities study. *Hum. Mol. Genet.* **26**, 3442–3450 (2017).

50. Assarsson, E. *et al.* Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. *PLOS ONE* **9**, e95192 (2014).

51. Hathout, Y. *et al.* Large-scale serum protein biomarker discovery in Duchenne muscular dystrophy. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7153–7158 (2015).

52. Ngo, D. *et al.* Aptamer-Based Proteomic Profiling Reveals Novel Candidate Biomarkers and Pathways in Cardiovascular Disease. *Circulation* **134**, 270–285 (2016).

53. Reed, B. D. *et al.* Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device. *Science* **378**, 186–192 (2022).

54. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, (2020).

55. Connally, N. J. *et al.* The missing link between genetic association and regulatory function. *eLife* **11**, e74970 (2022).

56. Venkata Subbiah, H., Ramesh Babu, P. & Subbiah, U. Determination of deleterious single-nucleotide polymorphisms of human LYZ C gene: an in silico study. *J. Genet. Eng. Biotechnol.* **20**, 92 (2022).

57. Savas, S., Tuzmen, S. & Ozcelik, H. Human SNPs resulting in premature stop codons and protein truncation. *Hum. Genomics* **2**, 274–286 (2006).

58. Qi, T. *et al.* Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat. Genet.* **54**, 1355–1363 (2022).

59. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).

60. Perzel Mandell, K. A. *et al.* Genome-wide sequencing-based identification of methylation quantitative trait loci and their role in schizophrenia risk. *Nat. Commun.* **12**, 5251 (2021).

61. Li, N. *et al.* Genetic Predisposition to Multiple Myeloma at 5q15 Is Mediated by an ELL2 Enhancer Polymorphism. *Cell Rep.* **20**, 2556–2564 (2017).

62. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).

63. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

64. Wang, Q. S. & Huang, H. Methods for statistical fine-mapping and their applications to auto-immune diseases. *Semin. Immunopathol.* **44**, 101–113 (2022).

65. Jaroszewicz, A. & Ernst, J. An integrative approach for fine-mapping chromatin interactions. *Bioinformatics* **36**, 1704–1711 (2019).

66. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).

67. Çalışkan, M. *et al.* Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. *Am. J. Hum. Genet.* **105**, 89–107 (2019).

68. Broekema, R. V., Bakker, O. B. & Jonkers, I. H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* **10**, 190221 (2020).

69. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genet.* **10**, e1004383 (2014).

70. Huber, C. D., Kim, B. Y. & Lohmueller, K. E. Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PLOS Genet.* **16**, e1008827 (2020).

71. Eijsbouts, C. Q., Burren, O. S., Newcombe, P. J. & Wallace, C. Fine mapping chromatin contacts in capture Hi-C data. *BMC Genomics* **20**, 77 (2019).

72. Choi, J. *et al.* Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat. Commun.* **11**, 2718 (2020).

73. Long, E. *et al.* Massively parallel reporter assays and variant scoring identified functional variants and target genes for melanoma loci and highlighted cell-type specificity. *Am. J. Hum. Genet.* **109**, 2210–2229 (2022).

74. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

75. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120362 (2013).

76. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

77. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation Biobanking* **13**, 311–319 (2015).

78. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).

79. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145–1149 (2016).

80. Liu, Y. *et al.* Methylomics of gene expression in human monocytes. *Hum. Mol. Genet.* **22**, 5065–5074 (2013).

81. Richardson, T. G., Hemani, G., Gaunt, T. R., Relton, C. L. & Davey Smith, G. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nat. Commun.* **11**, 185 (2020).

82. Zhu, H. & Zhou, X. Transcriptome-wide association studies: a view from Mendelian randomization. *Quant. Biol. Beijing China* **9**, 107–121 (2021).

83. Battle, A. *et al.* Impact of Regulatory Variation from RNA to Protein. *Science* **347**, 664–667 (2015).

84. He, B., Shi, J., Wang, X., Jiang, H. & Zhu, H.-J. Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. *BMC Biol.* **18**, 97 (2020).

85. Robins, C. *et al.* Genetic control of the human brain proteome. *Am. J. Hum. Genet.* **108**, 400–410 (2021).

86. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).

87. Nie, L., Wu, G. & Zhang, W. Correlation of mRNA Expression and Protein Abundance Affected by Multiple Sequence Features Related to Translational Efficiency in Desulfovibrio vulgaris: A Quantitative Analysis. *Genetics* **174**, 2229–2243 (2006).

88. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).

89. Stoney, R., Robertson, D. L., Nenadic, G. & Schwartz, J.-M. Mapping biological process relationships and disease perturbations within a pathway network. *Npj Syst. Biol. Appl.* **4**, 1–11 (2018).

90. Alberts, B. *et al.* Signaling Pathways That Depend on Regulated Proteolysis. *Mol. Biol. Cell 4th Ed.* (2002).

91. Karin, M. & Smeal, T. Control of transcription factors by signal transduction pathways: the beginning of the end. *Trends Biochem. Sci.* **17**, 418–422 (1992).

92. Gitig, D. M. & Koff, A. Cdk pathway: cyclin-dependent kinases and cyclin-dependent kinase inhibitors. *Mol. Biotechnol.* **19**, 179–188 (2001).

93. Ding, L. *et al.* The Roles of Cyclin-Dependent Kinases in Cell-Cycle Progression and Therapeutic Strategies in Human Breast Cancer. *Int. J. Mol. Sci.* **21**, 1960 (2020).

94. Ozaki, T. & Nakagawara, A. Role of p53 in Cell Death and Human Cancers. *Cancers* **3**, 994–1013 (2011).

95. Chen, J. The Cell-Cycle Arrest and Apoptotic Functions of p53 in Tumor Initiation and Progression. *Cold Spring Harb. Perspect. Med.* **6**, a026104 (2016).

96. Morris, E. J. *et al.* Cyclin-Dependent Kinases and P53 Pathways Are Activated Independently and Mediate Bax Activation in Neurons after DNA Damage. *J. Neurosci.* **21**, 5017–5026 (2001).

97. Komiya, Y. & Habas, R. Wnt signal transduction pathways. *Organogenesis* **4**, 68–75 (2008).

98. Kozmikova, I. & Kozmik, Z. Wnt/β-catenin signaling is an evolutionarily conserved determinant of chordate dorsal organizer. *eLife* **9**, e56817 (2020).

99. Lee, M. S., D'Amour, K. A. & Papkoff, J. A yeast model system for functional analysis of β-catenin signaling. *J. Cell Biol.* **158**, 1067–1078 (2002).

100. Schatoff, E. M., Leach, B. I. & Dow, L. E. Wnt Signaling and Colorectal Cancer. *Curr. Colorectal Cancer Rep.* **13**, 101–110 (2017).

101. Klaus, A. & Birchmeier, W. Wnt signalling and its impact on development and cancer. *Nat. Rev. Cancer* **8**, 387–398 (2008).

102. Patel, S., Alam, A., Pant, R. & Chattopadhyay, S. Wnt Signaling and Its Significance Within the Tumor Microenvironment: Novel Therapeutic Insights. *Front. Immunol.* **10**, (2019).

103. Wei, X. *et al.* The evolutionarily conserved MAPK/Erk signaling promotes ancestral T-cell immunity in fish via c-Myc–mediated glycolysis. *J. Biol. Chem.* **295**, 3000–3016 (2020).

104. Braicu, C. *et al.* A Comprehensive Review on MAPK: A Promising Therapeutic Target in Cancer. *Cancers* **11**, 1618 (2019).

105. Dhillon, A. S., Hagan, S., Rath, O. & Kolch, W. MAP kinase signalling pathways in cancer. *Oncogene* **26**, 3279–3290 (2007).

106. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14**, 482–517 (2019).

107. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* **9**, 4361 (2018).

108. White, M. J. *et al.* Strategies for Pathway Analysis using GWAS and WGS Data. *Curr. Protoc. Hum. Genet.* **100**, e79 (2019).

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

110. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).

111. Pico, A. R. *et al.* WikiPathways: Pathway Editing for the People. *PLOS Biol.* **6**, e184 (2008).

112. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

113. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

114. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).

115. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

116. Wang, H., Huang, B. & Wang, J. Predict long-range enhancer regulation based on protein–protein interactions between transcription factors. *Nucleic Acids Res.* **49**, 10347–10368 (2021).

117. Pott, S. & Lieb, J. D. What are super-enhancers? *Nat. Genet.* **47**, 8–12 (2015).

118. Mubeen, S. *et al.* The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Front. Genet.* **10**, 1203 (2019).

119. Simillion, C., Liechti, R., Lischer, H. E. L., Ioannidis, V. & Bruggmann, R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics* **18**, 151 (2017).

120. Li, D. & Chan, D. W. Proteomic cancer biomarkers from discovery to approval: it's worth the effort. *Expert Rev. Proteomics* **11**, 135–136 (2014).

121. Amiri-Dashatan, N., Koushki, M., Abbaszadeh, H.-A., Rostami-Nejad, M. & Rezaei-Tavirani, M. Proteomics Applications in Health: Biomarker and Drug Discovery and Food Industry. *Iran. J. Pharm. Res. IJPR* **17**, 1523–1536 (2018).

122. Enroth, S. *et al.* High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. *Commun. Biol.* **2**, 1–12 (2019).

123. Dhingra, R. & Vasan, R. S. Biomarkers in Cardiovascular Disease. *Trends Cardiovasc. Med.* **27**, 123–133 (2017).

124. Chan, M. *et al.* Development of a Next-Generation Sequencing Method for BRCA Mutation Screening: A Comparison between a High-Throughput and a Benchtop Platform. *J. Mol. Diagn.* **14**, 602–612 (2012).

125. Ross, C. A. & Tabrizi, S. J. Huntington's disease: from molecular pathogenesis to clinical treatment. *Lancet Neurol.* **10**, 83–98 (2011).

126. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

127. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

128. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. A guide to performing Polygenic Risk Score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

129. Rietveld, C. A. *et al.* GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science* **340**, 1467–1471 (2013).

130. Allegrini, A. G. *et al.* Genomic prediction of cognitive traits in childhood and adolescence. *Mol. Psychiatry* **24**, 819–827 (2019).

131. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

132. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).

133. Peyrot, W. J. *et al.* Effect of polygenic risk scores on depression in childhood trauma. *Br. J. Psychiatry* **205**, 113–119 (2014).

134. Martin, A. R., Daly, M. J., Robinson, E. B., Hyman, S. E. & Neale, B. M. Predicting Polygenic Risk of Psychiatric Disorders. *Biol. Psychiatry* **86**, 97–109 (2019).

135. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. B. Making the Most of Clumping and Thresholding for Polygenic Scores. *Am. J. Hum. Genet.* **105**, 1213–1221 (2019).

136. Liu, W., Zhuang, Z., Wang, W., Huang, T. & Liu, Z. An Improved Genome-Wide Polygenic Score Model for Predicting the Risk of Type 2 Diabetes. *Front. Genet.* **12**, (2021).

137. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, giz082 (2019).

138. Pain, O. *et al.* Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet.* **17**, e1009021 (2021).

139. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

140. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinforma. Oxf. Engl.* btaa1029 (2020) doi:10.1093/bioinformatics/btaa1029.

141. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

142. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).

143. Robinson, M. R. *et al.* Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* **1**, 1–13 (2017).

144. Yang, S. & Zhou, X. Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. *Am. J. Hum. Genet.* **106**, 679–693 (2020).

145. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).

146. Pain, O. *et al.* Imputed gene expression risk scores: a functionally informed component of polygenic risk. *Hum. Mol. Genet.* **30**, 727–738 (2021).

147. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

148. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. 2022.05.07.491045 Preprint at https://doi.org/10.1101/2022.05.07.491045 (2022).

149. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci. Publ. Protein Soc.* **28**, 1947–1951 (2019).

150. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).

151. Martens, M. *et al.* WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–D621 (2021).

152. Mi, H. *et al.* PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394–D403 (2021).

153. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles | PNAS. https://www.pnas.org/doi/10.1073/pnas.0506580102.

154. Harvey, S. L. *et al.* A phosphatase threshold sets the level of Cdk1 activity in early mitosis in budding yeast. *Mol. Biol. Cell* **22**, 3595–3608 (2011).

155. Schwarz, C. *et al.* A Precise Cdk Activity Threshold Determines Passage through the Restriction Point. *Mol. Cell* **69**, 253-264.e5 (2018).

156. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).

157. Burgess, S. & Thompson, S. G. Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).

158. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* (2015) doi:10.1038/ng.3367.

159. Bretherick, A. D. *et al.* Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits. *PLoS Genet.* **16**, e1008785 (2020).

160. Rudan, I. *et al.* "10 001 Dalmatians:" Croatia Launches Its National Biobank. *Croat. Med. J.* **50**, 4–6 (2009).

161. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

162. Wain, L. V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir. Med.* **3**, 769–781 (2015).

163. Eastwood, S. V. *et al.* Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PLOS ONE* **11**, e0162388 (2016).

164. Timmers, P. R. H. J. & Wilson, J. F. Limited Effect of Y Chromosome Variation on Coronary Artery Disease and Mortality in UK Biobank—Brief Report. *Arterioscler. Thromb. Vasc. Biol.* **42**, 1198–1206 (2022).

165. Macdonald-Dunlop, E. *et al.* Mapping genetic determinants of 184 circulating proteins in 26,494 individuals to connect proteins and diseases. 2021.08.03.21261494 Preprint at https://doi.org/10.1101/2021.08.03.21261494 (2021).

166. Repetto, L. *et al.* Genetic mechanisms of 184 neuro-related proteins in human plasma. 2023.02.10.23285650 Preprint at https://doi.org/10.1101/2023.02.10.23285650 (2023).

167. Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).

168. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).

169. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

170. Xu, Y. *et al.* An atlas of genetic scores to predict multi-omic traits. 2022.04.17.488593 Preprint at https://doi.org/10.1101/2022.04.17.488593 (2022).

171. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* (2015) doi:10.1016/j.ajhg.2015.09.001.

172. Cortes, A. *et al.* Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nat. Genet.* **49**, 1311–1318 (2017).

173. Zhang, J. D. & Wiemann, S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **25**, 1470–1471 (2009).

174. White, S. *KEGGlincs Design and Application: An R Package for Exploring Relationships in Biological Pathways*. (2017). doi:10.13140/RG.2.2.26158.41289.

175. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

176. Pontén, F. *et al.* A global view of protein expression in human cells, tissues, and organs. *Mol. Syst. Biol.* **5**, 337 (2009).

177. Alberts, B. *et al.* An Overview of Gene Control. *Mol. Biol. Cell 4th Ed.* (2002).

178. Harper, J. W. & Bennett, E. J. Proteome complexity and the forces that drive proteome imbalance. *Nature* **537**, 328–338 (2016).

179. Hennrich, M. L. *et al.* Cell-specific proteome analyses of human bone marrow reveal molecular features of age-dependent functional decline. *Nat. Commun.* **9**, 4004 (2018).

180. Pham, N., Hu, F., Evelo, C. T. & Kutmon, M. Tissue-specific pathway activities: A retrospective analysis in COVID-19 patients. *Front. Immunol.* **13**, (2022).

181. Zhang, Y. *et al.* Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Comput. Struct. Biotechnol. J.* **18**, 2953–2961 (2020).

182. Candia, J., Daya, G. N., Tanaka, T., Ferrucci, L. & Walker, K. A. Assessment of variability in the plasma 7k SomaScan proteomics assay. *Sci. Rep.* **12**, 17147 (2022).

183. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).

184. Kamburov, A. *et al.* ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.* **39**, D712-717 (2011).

185. Dennis, G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, R60 (2003).

186. Ma, Y. & Zhou, X. Genetic prediction of complex traits with polygenic scores: A statistical review. *Trends Genet. TIG* **37**, 995–1011 (2021).

187. Ni, G. *et al.* A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol. Psychiatry* **90**, 611–620 (2021).

188. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).

189. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).

190. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

191. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**, e00523 (2013).

192. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).

193. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).

194. KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG) version 1.30.1 from Bioconductor. https://rdrr.io/bioc/KEGGREST/.

195. Deshmane, S. L., Kremlev, S., Amini, S. & Sawaya, B. E. Monocyte Chemoattractant Protein-1 (MCP-1): An Overview. *J. Interferon Cytokine Res.* **29**, 313–326 (2009).

196. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).

197. Polygenic scores and inference using LDpred2 • bigsnpr. https://privefl.github.io/bigsnpr/articles/LDpred2.html.

198. Lewis, S. M., Williams, A. & Eisenbarth, S. C. Structure-function of the immune system in the spleen. *Sci. Immunol.* **4**, eaau6085 (2019).

199. Mebius, R. E. & Kraal, G. Structure and function of the spleen. *Nat. Rev. Immunol.* **5**, 606–616 (2005).

200. Baba, T. & Mukaida, N. Role of macrophage inflammatory protein (MIP)-1α/CCL3 in leukemogenesis. *Mol. Cell. Oncol.* **1**, e29899 (2014).

201. Iodice, S. *et al.* The Independent Role of Body Mass Index (BMI) and Severity of Depressive Symptoms on Biological Changes of Women Affected by Overweight/Obesity. *Int. J. Environ. Res. Public. Health* **18**, 2923 (2021).

202. Surmi, B. K., Webb, C. D., Ristau, A. C. & Hasty, A. H. Absence of macrophage inflammatory protein-1{alpha} does not impact macrophage accumulation in adipose tissue of diet-induced obese mice. *Am. J. Physiol. Endocrinol. Metab.* **299**, E437-445 (2010).

203. Bost, F., Aouadi, M., Caron, L. & Binétruy, B. The role of MAPKs in adipocyte differentiation and obesity. *Biochimie* **87**, 51–56 (2005).

204. Zhang, J.-M. & An, J. Cytokines, Inflammation and Pain. *Int. Anesthesiol. Clin.* **45**, 27–37 (2007).

205. Saxena, M. & Yeretssian, G. NOD-Like Receptors: Master Regulators of Inflammation and Cancer. *Front. Immunol.* **5**, (2014).

206. Zhong, Y., Kinio, A. & Saleh, M. Functions of NOD-Like Receptors in Human Diseases. *Front. Immunol.* **4**, 333 (2013).

207. Jaén, R. I. *et al.* Innate Immune Receptors, Key Actors in Cardiovascular Diseases. *JACC Basic Transl. Sci.* **5**, 735–749 (2020).

208. Kong, X., Yuan, Z. & Cheng, J. The function of NOD-like receptors in central nervous system diseases. *J. Neurosci. Res.* **95**, 1565–1573 (2017).

209. Burgess, S. *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30**, 543–552 (2015).

210. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 3300 (2019).

211. Zhang, C., Ye, Y. & Zhao, H. Comparison of Methods Utilizing Sex-Specific PRSs Derived From GWAS Summary Statistics. *Front. Genet.* **13**, (2022).

212. Zhang, W. & Liu, H. T. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.* **12**, 9–18 (2002).

213. Bahrami-B, F., Ataie-Kachoie, P., Pourgholami, M. H. & Morris, D. L. p70 Ribosomal protein S6 kinase (Rps6kb1): an update. *J. Clin. Pathol.* **67**, 1019–1025 (2014).

214. Donohoe, F., Wilkinson, M., Baxter, E. & Brennan, D. J. Mitogen-Activated Protein Kinase (MAPK) and Obesity-Related Cancer. *Int. J. Mol. Sci.* **21**, 1241 (2020).

215. Lim, A. K. H. *et al.* Role of MKK3–p38 MAPK signalling in the development of type 2 diabetes and renal injury in obese db/db mice. *Diabetologia* **52**, 347–358 (2009).

216. Kasonga, A., Kruger, M. C. & Coetzee, M. Activation of PPARs Modulates Signalling Pathways and Expression of Regulatory Genes in Osteoclasts Derived from Human CD14+ Monocytes. *Int. J. Mol. Sci.* **20**, 1798 (2019).

217. Thomas, A. W. *et al.* Exercise-associated generation of PPARγ ligands activates PPARγ signaling events and upregulates genes related to lipid metabolism. *J. Appl. Physiol. Bethesda Md 1985* **112**, 806–815 (2012).

218. Tachibana, K., Yamasaki, D., Ishimoto, K. & Doi, T. The Role of PPARs in Cancer. *PPAR Res.* **2008**, 102737 (2008).

219. Tan, Y. *et al.* PPAR-α Modulators as Current and Potential Cancer Treatments. *Front. Oncol.* **11**, (2021).

220. Koyasu, S. & Moro, K. Role of Innate Lymphocytes in Infection and Inflammation. *Front. Immunol.* **3**, (2012).

221. Shaw, A. C. *et al.* Dysregulation of Human Toll-like Receptor Function in Aging. *Ageing Res. Rev.* **10**, 346–353 (2011).

222. Kwak, S. H., Park, K. S., Lee, K. & Lee, H. K. Mitochondrial metabolism and diabetes. *J. Diabetes Investig.* **1**, 161–169 (2010).

223. Rovira-Llopis, S. *et al.* Mitochondrial dynamics in type 2 diabetes: Pathophysiological implications. *Redox Biol.* **11**, 637–645 (2017).

224. Hersoug, L.-G., Møller, P. & Loft, S. Role of microbiota-derived lipopolysaccharide in adipose tissue inflammation, adipocyte size and pyroptosis during obesity. *Nutr. Res. Rev.* **31**, 153–163 (2018).

225. Kumari, M. *et al.* IRF3 promotes adipose inflammation and insulin resistance and represses browning. *J. Clin. Invest.* **126**, 2839–2854 (2016).

226. Wu, C. *et al.* Enhancing hepatic glycolysis reduces obesity: differential effects on lipogenesis depend on site of glycolytic modulation. *Cell Metab.* **2**, 131–140 (2005).

227. Westermark, P., Andersson, A. & Westermark, G. T. Islet amyloid polypeptide, islet amyloid, and diabetes mellitus. *Physiol. Rev.* **91**, 795–826 (2011).

228. Mogensen, C. E., Christensen, N. J. & Gundersen, H. J. The acute effect of insulin on heart rate, blood pressure, plasma noradrenaline and urinary albumin excretion. The role of changes in blood glucose. *Diabetologia* **18**, 453–457 (1980).

229. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, (2021).

230. Raffield, L. M. *et al.* Comparison of Proteomic Assessment Methods in Multiple Cohort Studies. *Proteomics* **20**, e1900278 (2020).

231. Haslam, D. E. *et al.* Stability and reproducibility of proteomic profiles in epidemiological studies: comparing the Olink and SOMAscan platforms. *PROTEOMICS* **22**, 2100170 (2022).

232. Khalid, A. B. & Krum, S. A. Estrogen Receptors Alpha and Beta in Bone. *Bone* **87**, 130–135 (2016).

233. Khosla, S., Oursler, M. J. & Monroe, D. G. Estrogen and the Skeleton. *Trends Endocrinol. Metab. TEM* **23**, 576–581 (2012).

234. Dai, R. *et al.* Neutrophils and neutrophil serine proteases are increased in the spleens of estrogen-treated C57BL/6 mice and several strains of spontaneous lupus-prone mice. *PLoS ONE* **12**, e0172105 (2017).

235. Khan, D. & Ansar Ahmed, S. The Immune System Is a Natural Target for Estrogen Action: Opposing Effects of Estrogen in Two Prototypical Autoimmune Diseases. *Front. Immunol.* **6**, (2016).

236. Kovats, S. Estrogen receptors regulate innate immune cells and signaling pathways. *Cell. Immunol.* **294**, 63–69 (2015).

237. Cao, C. & Moult, J. GWAS and drug targets. *BMC Genomics* **15**, (2014).

238. King, E. A., Wade Davis, J. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, (2019).

239. Farber, C. R. & Mesner, L. D. Chapter 3 - A Systems-Level Understanding of Cardiovascular Disease through Networks. in *Translational Cardiometabolic Genomic Medicine* (ed. Rodriguez-Oquendo, A.) 59–81 (Academic Press, 2016). doi:10.1016/B978-0-12-799961-6.00003-2.

240. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

241. Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).

242. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLOS Comput. Biol.* **12**, e1004714 (2016).

243. Pain, O. *et al.* Polygenic Prediction of Molecular Traits using Large-Scale Meta-analysis Summary Statistics. 2022.11.23.517213 Preprint at https://doi.org/10.1101/2022.11.23.517213 (2022).

244. Ni, G. *et al.* A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biol. Psychiatry* **90**, 611–620 (2021).

245. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).

246. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinforma. Oxf. Engl.* **35**, 4851–4853 (2019).

Appendices

A.1 Significantly improved multi-tissue models

Gene name	Entrez ID	Multi-tissue model r^2	Best individual tissue r^2	GTEx best tissue	Likelihood P-value
ALDH3A1	218	0.139	0.103	Adrenal_Gland	4.02E-10
AMN	81693	0.069	0.037	Liver	1.11E-08
ARSA	410	0.178	0.025	Brain_Frontal_Cortex_BA9	1.66E-38
ASAH2	56624	0.053	0.014	Stomach	3.12E-10
BAMBI	25805	0.027	0.000	Testis	3.25E-07
BCL2L11	10018	0.018	0.000	Brain_Cerebellar_Hemisphere	3.75E-05
BST1	683	0.216	0.131	Nerve_Tibial	5.24E-24
CAPG	822	0.248	0.192	Pituitary	4.67E-17
CD274	29126	0.055	0.039	Brain_Cerebellum	6.33E-05
CD300C	10871	0.174	0.137	Adipose_Subcutaneous	5.73E-11
CD33	945	0.186	0.096	Adipose_Visceral_Omentum	6.11E-24
CD6	923	0.207	0.188	Brain_Cortex	1.97E-06
CLEC7A	64581	0.296	0.283	Nerve_Tibial	1.93E-05
CLUL1	27098	0.161	0.133	Brain_Cerebellum	1.44E-08
CPXM1	56265	0.134	0.099	Skin_Sun_Exposed_Lower_leg	5.68E-10
CRISP2	7180	0.074	0.056	Whole_Blood	1.28E-05
CTSH	1512	0.387	0.263	Esophagus_Muscularis	4.27E-41
DAG1	1605	0.018	0.000	Heart_Left_Ventricle	2.27E-05
DCBLD2	131566	0.045	0.015	Esophagus_Muscularis	3.75E-08
DPEP1	1800	0.094	0.062	Adrenal_Gland	6.34E-09
ENTPD6	955	0.067	0.015	Cells_Transformed_fibroblasts	3.50E-13
FAM3B	54097	0.123	0.074	Stomach	3.66E-13
FCGR2B	2213	0.091	0.052	Artery_Tibial	1.98E-10
FCN2	2220	0.115	0.095	Adrenal_Gland	2.27E-06
FCRLB	127943	0.141	0.031	Artery_Tibial	3.05E-27
FGF2	2247	0.156	0.140	Testis	1.93E-05
FOLR3	2352	0.330	0.314	Thyroid	1.75E-06
FRZB	2487	0.039	0.016	Lung	1.43E-06
GPC5	2262	0.181	0.082	Testis	5.63E-26
HNMT	3176	0.091	0.069	Artery_Aorta	9.14E-07
IDUA	3425	0.139	0.108	Adipose_Visceral_Omentum	6.55E-09
IGF2R	3482	0.076	0.058	Heart_Atrial_Appendage	1.02E-05
IL17D	53342	0.050	0.025	Testis	6.88E-07
IL17RB	55540	0.278	0.242	Nerve_Tibial	6.83E-12
IL18R1	8809	0.189	0.115	Skin_Sun_Exposed_Lower_leg	2.63E-20

IL1RL2	8808	0.076	0.060	Brain_Nucleus_accumbens_basal_ganglia	5.69E-05
KLB	152831	0.154	0.116	Liver	6.12E-11
KLK10	5655	0.193	0.163	Skin_Sun_Exposed_Lower_leg	1.62E-09
KLK12	43849	0.474	0.446	Small_Intestine_Terminal_Ileum	1.23E-12
LGALS8	3964	0.120	0.105	Brain_Nucleus_accumbens_basal_ganglia	4.83E-05
LRP11	84918	0.084	0.052	Skin_Sun_Exposed_Lower_leg	4.84E-09
MDGA1	266727	0.405	0.358	Prostate	5.60E-18
MGMT	4255	0.148	0.132	Skin_Sun_Exposed_Lower_leg	1.76E-05
NQO2	4835	0.381	0.221	Brain_Caudate_basal_ganglia	3.27E-51
NT5E	4907	0.171	0.098	Whole_Blood	7.59E-20
PDCD1LG2	80380	0.171	0.145	Esophagus_Gastroesophageal_Junction	3.33E-08
PILRA	29992	0.200	0.076	Heart_Atrial_Appendage	7.43E-33
PILRB	29990	0.373	0.100	Adipose_Subcutaneous	1.69E-79
PON2	5445	0.362	0.259	Heart_Left_Ventricle	7.83E-34
PRTG	283659	0.037	0.013	Cells_Transformed_fibroblasts	9.69E-07
PVR	5817	0.199	0.181	Cells_Transformed_fibroblasts	3.20E-06
SERPINB8	5271	0.083	0.062	Artery_Aorta	1.94E-06
SFTPD	6441	0.114	0.050	Pancreas	1.43E-16
SIRPB1	10326	0.119	0.029	Spleen	9.56E-23
SLAMF8	56833	0.075	0.026	Heart_Atrial_Appendage	1.19E-12
SNCG	6623	0.403	0.361	Cells_Transformed_fibroblasts	4.63E-16
SUMF2	25870	0.051	0.022	Adipose_Visceral_Omentum	5.71E-08
TACSTD2	4070	0.207	0.117	Muscle_Skeletal	1.96E-24
TCN2	6948	0.229	0.198	Skin_Sun_Exposed_Lower_leg	3.55E-10
TLR3	7098	0.071	0.038	Cells_Transformed_fibroblasts	3.61E-09
TMPRSS5	80975	0.185	0.122	Colon_Transverse	1.07E-17
TREM1	54210	0.081	0.051	Nerve_Tibial	2.17E-08
XCL1	6375	0.139	0.120	Whole_Blood	3.17E-06