Omics Measures of Ageing and Disease Susceptibility

Erin Macdonald-Dunlop

The University of Edinburgh

Doctor of Philosophy with Integrated Study

2021

Declaration

I declare that this thesis has been composed by myself and the work has not been submitted for any other degree or qualification. I confirm that the work is my own, except where work which has formed part of jointly-authored manuscripts has been included, in which case my contribution and those of others to the work has been explicitly indicated below and in chapters 3-5. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

……………………………………………………………………………………………..

Erin Macdonald-Dunlop, 8th August 2021

Abstract

While genomics has been a major field of study for decades due to relatively inexpensive genotyping arrays, the recent advancement of technology has also allowed the measure and study of various "omics". There are now numerous methods and platforms available that allow high throughput and high dimensional quantification of many types of biological molecules. Traditional genomics and transcriptomics are now joined by proteomics, metabolomics, glycomics, lipidomics and epigenomics.

I was lucky to have access to a unique resource in the Orkney Complex Disease Study (ORCADES), a cohort of individuals from the Orkney Islands that are extremely deeply annotated. Approximately 1000 individuals in ORCADES have genomics, proteomics, lipidomics, glycomics, metabolomics, epigenomics, clinical risk factors and disease phenotypes, as well as body composition measurements from whole body scans. In addition to these cross-sectional omics and health related measures, these individuals also have linked electronic health records (EHR) available, allowing the assessment of the effect of these omics measures on incident disease over a ~10-year follow up period. In this thesis I use this phenotype rich resource to investigate the relationship between multiple types of omics measures and both ageing and health outcomes.

First, I used the ORCADES data to construct measures of biological age (BA). The idea that there is an underlying rate at which the body deteriorates with age that varies between individuals of the same chronological age, this biological age, would be more indicative of health status, functional capacity and risk of age-related diseases than chronological age. Previous models estimating BA (ageing clocks) have predominantly been built using a single type of omics assay and comparison between different omics ageing clocks has been limited. I performed the most exhaustive comparison of different omics ageing clocks yet, with eleven clocks spanning nine different omics assays. I show that different omics clocks overlap in the information they provide about age, that some omics clocks track more generalised ageing while others track specific disease risk factors and that omics ageing clocks are prognostic of incident disease over and above chronological age.

Second, I assessed whether individually or in multivariable models, omics measures are associated with health-related risk factors or prognostic of incident disease over 10 years post-assessment. I show that 2,686 single omics biomarkers are associated with 10 risk factors and 44 subsequent incident diseases. I also show that models built using multiple biomarkers from whole body scans, metabolomics, proteomics and clinical risk factors are prognostic of subsequent diabetes mellitus and that clinical risk factors are prognostic of incident hypertensive disorders, obesity, ischaemic heart disease and Framingham risk score.

Third, I investigated the genetic architecture of a subset of the proteomics measures available in ORCADES, specifically 184 cardiovascular-related proteins. Combining genome-wide association (GWAS) summary statistics from ORCADES and 17 other cohorts from the SCALLOP Consortium, giving a maximum sample size of 26,494 individuals, I performed 184 genome-wide association meta-analyses (GWAMAs) on the levels of these proteins circulating in plasma. I discovered 592 independent significant loci associated with the levels of at least one protein. I found that between 8-37% of these significant loci colocalise with known expression quantitative trait loci (eQTL). I also find evidence of causal associations between 11 plasma protein levels and disease susceptibility using Mendelian randomisation, highlighting potential candidate drug targets.

Lay Summary

While the human DNA sequence has been relatively easy and cost effective to measure and is therefore extensively studied, recent technological advancements have allowed many other different types of biological molecules that circulate in the blood to be measured. As well as DNA, it is now possible to measure large numbers of proteins, metabolites, fat molecules and sugar molecules for example. Together these are known as “omics”, due to proteomics, metabolomics, lipidomics and glycomics etc. All of this information about an individual, these different types of omics, can be measured in a single blood sample.

I was lucky to have access to a unique resource in the Orkney Complex Disease Study (ORCADES), a study population of individuals from the Orkney Islands in Scotland, that we have a lot of biological information about. There are nine different types of omics measured as well as information about their health, lifestyle, and any disease diagnoses. In addition to all of this information that was collected at time of recruitment, there is also information about any diseases these individuals were hospitalised with over a 10-year follow up period. In this thesis I use this biological information from the ORCADES study to investigate the relationship between multiple types of omics measures and both ageing and health outcomes.

First, I used the different types of omics in ORCADES to measure how healthy individual’s bodies are compared to other individuals of the same age, this measure is called biological age. I compared the biological age measures based on different omics and found: that they massively overlap in the information they provide about age, that some omics biological age measures track more generalised ageing while others track specific disease risk factors and that omics biological age measures are able to predict future hospitalisation due to disease, better than just knowing how old someone is.

Second, I tested whether either on their own or together, omics measures are able to predict commonly measured disease risk factors such as cholesterol and blood pressure. Further, I looked to see if omics measures can predict if an individual is likely to be hospitalised for a particular disease in the future.

Third, I combined data from 18 different study populations including ORCADES to investigate how variation in the DNA sequence between individuals leads to the differing levels of 184 cardiovascular related proteins that are measured circulating in the blood. I found 592 regions of the DNA that are associated with changes in at least one of the 184 protein levels. I also show that these regions of the DNA overlap with regions known to be associated with many different diseases and that there is evidence that changes in some of these protein levels in the blood cause a change in disease risk.

Acknowledgements

I would like to offer my thanks to:

The MRC Precision Medicine Doctoral Training Programme and the University of Edinburgh for funding this work.

The creators and maintainers of the computing tools I used throughout this work and those researchers who contribute to open science and make their data publicly available, particularly eQTLGen and GTEx.

Everyone who contributed to the funding, creation, maintenance and handling of all of the twenty-one cohorts used throughout this thesis. With a particular thanks to the participants who volunteered their time and their data, without their contribution scientific research would not be possible.

Collaborators in Estonia, Lausanne and Russia for hosting me. Thanks to Krista Fischer and Nele Taba for their guidance and assistance on the work on ageing clocks. Thanks to Yurii Aulchenko and his group at Novosibirsk State University for inviting us to visit on multiple occasions, giving us the opportunity to teach and especially for looking after us given our negligible Russian.

All of the members of the SCALLOP Consortium who contributed their data, time and expertise to this project. With particular thanks to Jim Wilson and Anders Mälarstig for giving me the opportunity to present the work done on this project on an international stage.

All of the co-authors whose contributions are highlighted throughout this thesis for their assistance and their comments on the included manuscripts.

All of the members of the Wilson Group for providing a welcoming, supportive and sometimes insane environment, it has been a pleasure to work with such kind, funny, knowledgeable and motivated people for the last 4 years. Especially Paul

Timmers for his endless patience when being bombarded with coding questions and Kat Kentistou for keeping me sane and picking me up no matter the country.

My supervisors Peter Joshi, Jim Wilson and Chris Haley for providing me with access to their data, advice, support, encouragement, opportunities to collaborate and feedback throughout my PhD. But I would also like to thank both Peter and Jim for pushing me, during this PhD I have done things I never thought I would do, and indeed would not have done, without the push.

Sean Bankier for his unwavering support, even in the testing circumstances of two theses being written while locked down in one small flat during a global pandemic.

Finally, I would like to thank my parents for their endless sacrifices, patience and support, without their help I would not be where I am today. Given how much they have had to hear about this PhD over the last four years they could probably give a decent lecture on GWAS.

Assistance

Thesis Drafting: My supervisors Peter Joshi and Jim Wilson commented on drafts of the chapters in this thesis, highlighting areas that required re-wording, further explanation, cut down, additional references and modifications of tables and figures.

Human Cohort Data: Called array and imputed genotypes were provided to me for subsequent analysis for ORCADES by Jim Wilson and for Croatia-Vis by Caroline Hayward. Phenotype data for clinical risk factors from the UK Biobank were downloaded and pre-processed by David Clark. The following omics assay datasets were QC'd by the indicated analysts and provided to me for subsequent analysis: UPLC IgG Glycomics data in ORCADES, Croatia-Vis and Croatia-Korčula by Lucija Klarić, DNA methylation data in ORCADES by Azra Frkatović and DNA methylation in GS:SFHS by Rosie Walker.

The work presented in Chapter 3 was submitted for publication as *"A catalogue of omics biological ageing clocks reveals substantial commonality and associations with disease risk"* by **Erin Macdonald-Dunlop**, Nele Taba, Lucija Klarić, Azra Frkatović, Rosie Walker, Caroline Hayward, Tõnu Esko, Chris Haley (Supervisor), Krista Fischer, James F Wilson (Supervisor), Peter K Joshi (Supervisor). Details of each author's contribution are listed in Chapter 3.

For the work in both Chapters 3 and 4 Peter Joshi extracted, pre-processed and performed quality control on the SMR01 electronic health record data and provided me with incident hospital admission data for groups of ICD10 codes for ORCADES and calculated Martingale residuals that I subsequently used for analysis.

The work presented in Chapter 5 was submitted for publication as *"Mapping genetic determinants of 184 circulating proteins in 26,494 individuals to connect proteins and diseases"* by **Erin Macdonald-Dunlop**, Lucija Klarić, Lasse Folkersen, Paul R.H.J. Timmers, Stefan Gustafsson, Jing Hua Zhao, Niclas Eriksson, Anne

Richmond, Stefan Enroth, Niklas Mattsson-Carlgren, Daria V. Zhernakova, Anette Kalnapenkis, Martin Magnusson, Eleanor Wheeler, Shih-Jen Hwang, Yan Chen, Andrew P. Morris, Bram Prins, Urmo Võsa, Nicholas J. Wareham, John Danesh, Johan Sundstrom, Bruna Gigante, Damiano Baldassarre, Rona J. Strawbridge, Harry Campbell, Ulf Gyllensten, Chen Yao, Daniela Zanetti, Themistocles L. Assimes, Per Eriksson, Daniel Levy, Claudia Langenberg, J. Gustav Smith, Tõnu Esko, Jingyuan Fu, Oskar Hansson, Åsa Johansson, Caroline Hayward, Lars Wallentin, Agneta Siegbahn, Lars Lind, Adam S. Butterworth, Karl Michaëlsson, James E. Peters, Anders Mälarstig, Peter K. Joshi (Supervisor), James F. Wilson (Supervisor). Details of author's contributions are listed in Chapter 5.

List of Abbreviations

AgeAccelGrim	GrimAge Acceleration
AMAR	Apparent methylation ageing rate
ASCVD	Atherosclerotic cardiovascular disease
BA	Biological Age
BH	Benjamini-Hochberg
BMI	Body Mass Index
CAD	Coronary artery disease
CD	Crohn's Disease
CHD	Coronary heart disease
ChIP-seq	Chromatic Immunoprecipitation-sequencing
ChronAge	Chronological Age
CI	Confidence interval
CNS	Central Nervous System
COJO	Conditional and Joint Association analysis
COPD	Chronic obstructive pulmonary disease
COVID	Coronavirus
C_q	Quantification Cycle
CRP	C reactive protein
C_t	Cycle Threshold
CVD	Cardiovascular Disease
DAG	Directed Acyclic Graphic
DBP	Diastolic blood pressure
DEXA	Dual-energy X-ray Absorptiometry
DNA	Deoxyribonucleic acid
DNAm GrimAge	DNA Methylation Grim Age
DNAm PhenoAge	DNA Methylation Pheno Age
DNAme	DNA methylation
DNase	Deoxyribonuclease
doi	Digital Object Identifier
EDTA	Ethylenediamine tetraacetic acid
EDU	Educational Attainment
EGCUT	Estonian Genome Center of the University of Tartu
EHR	Electronic Health Records
eQTL	Expression quantitative trait loci
ES	Embryonic stem cells
EWS	Exome-wide Sequence
FDR	False discovery rate
FEV1	Forced expiratory volume in 1 minute
FRS	Framingham risk score
GCTA	Genome-wide Complex Trait Analysis
GO	Gene Ontology
GP	General Practitioner
GRAMMAR	Genome wide rapid association using mixed model regression
GRM	Genomic relatedness matrix
GS:SFHS	Generation Scotland Scottish Family Health Study
GTEx	Genotype-Tissue Expression

GWAMA	Genome-wide Association meta-analysis
GWAS	Genome-wide Association Study
HD	Heart Disease
HDL	High density lipoprotein
HEIDI	Heterogeneity in dependent instruments
HIV1	Human immunodeficiency virus 1
HR	Hazard ratio
HRC	Haplotype Reference Consortium
HWE	Hardy Weinberg Equilibrium
IBD	Inflammatory bowel disease
ICD	International Classification of Diseases
IgG	Immunoglobulin G
IHD	Ischaemic Heart Disease
INDEL	Insertion or deletion
iPS	Induced pluripotent stem cells
ISLSP	Independently sampling from a latent set of complete predictors
IV	Instrumental variable
IVW	Inverse variance weighted
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LASSO	Least absolute shrinkage and selection operator
LC-MS/MS	Liquid Chromatography Tandem Mass Spectrometry
LD	Linkage disequilibrium
LDL	Low density lipoprotein
LDSC	Linkage disequilibrium score regression
LLOD	Lower Limit of Detection
LOD	Limit of Detection
LR	Lower Respiratory tract
MAF	Minor allele frequency
methQTL	DNA Methylation Quantitative Trait Loci
MI	Myocardial Infarction
ML	Maximum likelihood
MN	Malignant Neoplasm
mQTL	Metabolite Quantitative Trait Loci
MR	Mendelian Randomisation
MRI	Magnetic Resonance Imaging
MS	Mass spectrometry/Multiple Sclerosis
NA	Not Applicable
NCBI	National Center for Biotechnology Information
NMR	Nuclear magnetic resonance
NPX	Normalised Protein eXpression
OCA	Omics clock age
OCAA	Omics clock age acceleration
OLS	Ordinary least squares
ONS	Office for National Statistics
OR	Odds ratio
ORCADES	Orkney Complex Disease study
PBWT	Positional Burrows-Wheeler Transform
PC	Principal components
PEA	Proximity extension assay
PP	Posterior probability

pQTL	Protein quantitative trait loci
QC	Quality control
qPCR	Quantitative Polymerase Chain Reaction
QTL	Quantitative trait loci
RCT	Randomised Controlled Trial
REML	Restricted maximum likelihood
RNA	Ribonucleic acid
SBP	Systolic blood pressure
SCALLOP	Scandinavian collaboration for Olink plasma protein genetics
SD	Standard deviation
SE	Standard error
SLE	Systemic Lupus Erythematosus
SMR	Summary data-based Mendelian Randomisation
SMR01	Scottish Morbidity Records of Hospital Admission
SN LR	Suppurative & necrotic conditions of the lower respiratory tract
SNP	Single nucleotide polymorphism
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SuSiE	Sum of Single Effects
T2D	Type II Diabetes
TC	Total Cholesterol
TF	Transcription Factor
TSS	Transcriptional start site
UC	Ulcerative colitis
UKB	UK Biobank
UPLC	Ultra-performance liquid chromatography
VE	Variance Explained
WGS	Whole Genome sequencing
WHR	Waist-to-hip ratio

Contents

Tables & Figures

Chapter 1: Introduction

At its simplest, the aim of all scientific research is to gain a greater understanding of the world around us, how it works, and why it is the way it is. Physics often dedicates itself to the extremely small using quantum mechanics and the very large. In turn, chemistry concerns itself with elements and compounds, how they combine and interact. However, Biology aims to understand how all of these things come together to create living organisms, their inner workings and complexities and why they are the way they are.

It is only with our understanding of biology improving over the centuries that we have modern medicine. Without the knowledge of how our immune system works there would not have been 3 approved vaccines for COVID-19 developed[1–3] and 67.3 million doses distributed in the last 15 months in the UK[4]. Given that the theory that diseases were spread by noxious air, miasma[5], advanced by Hippocrates (c.460-377 B.C.E)[6] was still being used by scientists to explain epidemics of cholera[7] and the black death[8,9] in the 1850's and 14th Century respectively, this is an enormous advancement. However, despite the vast improvement since then there is still so much that we do not know.

Relative to the detailed mechanistic understanding that physics and chemistry have achieved, biology is still searching for an understanding of the end-to-end details of the mechanisms that support life. While we have extremely detailed knowledge of certain processes for example, how the human genome is copied and passed on during cell division and the sequence of steps that provide the body with energy, the Krebs cycle[10], these are often disparate. Our holistic understanding of biomedical systems remains vague in comparison. For example, there are effective medicines used today where we do not fully understand how they work, merely that they do, rather than having the detailed step-by-step comprehension that enables certain and accurate prediction of the effect of interventions and treatments. In 2019 there were still 18.6 and 10 million deaths due to heart disease and cancer, respectively[11], despite the billions of pounds of funding devoted to their research.

Nonetheless, improvement in modern medicine has more recently created new challenges for society and biological research. For example, people living longer has increased the prevalence of age-related diseases, highlighting the need to deepen our understanding of the underlying biology of ageing.

The only way to improve on this situation is to carry out further research. Only by increasing our understanding of the underlying biology will it be possible to establish: which pathways and mechanisms bring about ageing and diseases, design measures that predict individuals' risk of developing diseases in the future and devise effective treatments.

1.1 Promise of Genetics

Genetics was a promising opportunity for exploring variation between individuals. The idea that traits could be inherited, passed down through the generations was first demonstrated in the 19th Century by Mendel[12] however, it was the discovery of the structure of deoxyribonucleic acid (DNA), that carries information from one generation to the next in 1952[13], that was a pivotal step in deepening our understanding of molecular biology.

With the development of Sanger sequencing[14] in 1977, it became possible to determine the sequence of bases (nucleotides) that made up a short sequence of a sample's genetic code: adenine (A), guanine (G), thymine (T) and cytosine (C). However, this was an extremely inefficient and labour-intensive process, resulting in many different research groups each focussed on disparate, highly specific regions, genes or pathways of interest. Consequently, our understanding of DNA variation and its effect on traits as a cohesive whole progressed relatively slowly.

However, in 2003, due to the worldwide collaborative effort of the Human Genome Project, the first complete human DNA sequence (genome) became available[15]. This was hailed as the missing piece of the puzzle that would finally facilitate the long-awaited rapid progression of biological research[16], and in part it did. Having a complete human reference sequence, allowed the subsequent discovery of locations in the genome that vary between individuals, given that all Homo sapiens generally share 99.9% of their DNA sequence[15]. Individual bases where the alleles are known to vary between individuals, for example between A and C in a given population, are known as single nucleotide polymorphisms (SNPs). Together with the advancement of microarray technology, it became possible to measure the pair of alleles (one on each strand - genotypes) at hundreds of thousands of SNPs across the genome simultaneously[17]. With this information, it became possible to assess if the variant that an individual possesses at a given position in the genome is associated with a trait of interest.

Since the landmark publication in 2007[18], genome wide association studies (GWAS) have been the gold standard to uncover the underlying biology of complex traits. These studies systematically test if variation in the alleles at sites across the genome are associated with variation in the trait of interest (such as height or type II Diabetes mellitus)[19,20]. GWA studies were sold as the solution to the field’s previous difficulties, they would: allow the identification of regions of the genome that influence disease, provide insight into the underlying biological mechanisms and pathways that bring about these phenotypes and pinpoint potential opportunities for new treatments and interventions[16].

The continuing decline in price of genotyping arrays has meant that over 4,300 papers reporting on 4,500 GWAS have been published since 2007, spanning over 5,000 different phenotypes and reporting over 55,000 unique, associated genetic loci[21]. However, despite this, very few of these findings have been successfully translated into clinical use[22]. One of the reasons for this is that it is extremely difficult to infer causal genes (as the discovered causal regions often contain tens of genes) and thus end-to-end mechanisms of action of the associated genetic variants on the trait of interest[23–26]. Another reason is the many layers of separation between the DNA sequence level and end-point disease phenotype. It is thus natural to look to intermediate phenotypes, such as circulating biological molecules, to further unravel the end-to-end pathways and underlying complexities that result in some individuals developing disease and others ageing whilst disease free.

1.2 The Omics Era

Just as technology advanced for measuring genomics in the form of both genotyping and sequencing, there has been similar advancement in platforms capable of capturing large numbers of other types of biological molecules (omics). In addition to genomics, genome-wide gene expression levels (transcriptomics) and DNA methylation (epigenomics) which use sequencing technologies, high throughput assays quantifying the levels of often hundreds to thousands of circulating: plasma protein (proteomics), metabolite (metabolomics), lipid (lipidomics) and glycan (glycomics) levels are now available.

Transcriptomics is the study of the levels of RNA transcripts that are produced when the regions of the DNA that encode proteins are transcribed into RNA. These transcripts are then transported from the nucleus to ribosomes, so they can be translated into proteins. A snapshot of the levels of thousands of transcripts therefore provides a picture of the levels of all the genes expressed in a specific tissue at a specific time[27]. Their study allows the identification of genes that are differentially expressed in different cells, tissues, during diseases and in response to different treatments, offering the opportunity to form a more complete understanding of what is happening in the body.

The regulation of which genes are expressed over time and in which tissue is complex. They are partly regulated by epigenetic marks on DNA molecules, which vary by cell within individuals. One type of epigenomics, DNA methylation, has aroused considerable interest. DNA methylation is the reversible addition of methyl groups to cytosine residues, often in regions of the genome containing runs of GC dinucleotides (CpGs)[28], particularly promoters[29,30], and has been shown to regulate gene expression[31]. DNA methylation patterns have also been shown to change with environmental exposures and behaviours, such as tobacco smoking[32]. The ability to measure the levels of methylation at 850,000 CpG sites across the genome simultaneously[33], therefore provides the opportunity to provide a comprehensive picture by adding a substantial layer of information beyond the DNA sequence level.

Proteins are the building blocks of cells and are involved in all key cell processes, thus the ability to capture and study the abundance of hundreds to thousands (dependent of the platform) of proteins circulating in the plasma, proteomics, is invaluable[34–36]. The fact that there are numerous licenced drugs that target circulating plasma proteins[37–40] means proteomics assays are of particular interest in the search for novel drug targets or candidates for repurposing.

Glycans are sugar molecules that are added to proteins post-translation and are involved in functional regulation[41]. The diverse range of different glycans that are added to immunoglobulin G (IgG) antibodies, which are involved in the immune response, regulate IgG activity. The study of these glycans, IgG glycomics, offers the potential to understand how the immune system is regulated in response to different diseases, for example, IgG glycans have been shown to promote the shift between pro- and anti-inflammatory functions. Disruption of glycosylation has been associated with auto-immune[42] disorders as well as numerous other diseases[43]. This suggests further study of IgG glycomics has the potential to provide useful insight into pathways and mechanisms that lead to disease.

Metabolites are intermediate products of metabolic reactions; platforms to measure the levels of different metabolites circulating in the blood often capture overlapping measures with lipidomics assays due to the number of lipid fractions involved in the metabolic process. Large scale profiling of metabolite levels, metabolomics, therefore obtains a snapshot of circulating molecules that capture all major processes within the body[44].

Given the sheer number of different layers of biology that these omics capture beyond the DNA sequence level, they offer the opportunity to form a more comprehensive picture of what is happening inside the organism, especially those that are involved in regulation. Further, several of these omics biomarkers are much closer to end point phenotypes than DNA and, in the case of proteins, are themselves potential actors. Omics measures could therefore help to disentangle mechanisms for how associated SNPs influence disease phenotypes.

Each of the aforementioned omics assays quantify hundreds to thousands of features simultaneously. Due to the expense, there are often more features measured (p) than samples (n), these features are often highly correlated, providing challenges for statistical analyses. However, as these assays become more affordable, cohorts with increased number of samples with a number of these high-dimensional omics assays are becoming more abundant, giving rise to the field of multi-omics. Big Multi-omics data is an invaluable resource, and when paired with efficient computing and statistical methods such as machine learning, creates enormous potential for precision medicine research. It has the potential to increase our understanding of the complexities of biology that underpin ageing and disease, that are still lacking.

1.3 Ageing

A key area of research that has made extensive use of omics assays and a variety of statistical and machine learning approaches is ageing. It is the biggest risk factor for most late onset chronic diseases and contributes to morbidity and mortality.

Like many countries globally, the UK has an ageing population with 18.5% over the age of 65 (mid-2019) combined with a declining birth rate[45]. This is in part due to the steady increase in life expectancy of 3 years per decade (1970-2010)[46], which is at 79.4 years (males) and 83.1 (females) as of 2019 (ONS life expectancy 2017-2019[47]). However, this observed increase in life expectancy is not fully accompanied by an increase in the number of years expected to live disease free[48]. Interventions promoting healthy ageing could thus improve length and quality of life. If we had the ability to identify the prematurely aged or those at risk of premature ageing, this could drive personalised medicine, again improving quality of life.

Despite being aware of numerous changes that occur as we age, little is understood about which of these are causes or consequences of ageing[49]. That the changes to outward appearance such as greying hair, baldness, loss of skin elasticity and worsening of posture occur at different rates between individuals is apparent to us all. Less visible indicators such as a decline in eyesight and hearing and hypertension are also familiar to most. However, there are also molecular changes that occur with age such as: shortening of telomeres[50], genomic instability, DNA methylation pattern[51], deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, loss of proteostasis and altered intercellular communication that are all well documented[49].

Individuals of the same chronological age also vary enormously in their apparent frailty, health and number of co-morbidities, further suggesting that there may be differences in the underlaying rate at which people age. Even lifestyle factors that are known health risks such as smoking, stress and an individual's physical fitness have been shown to alter telomere length[52], DNA methylation patterns and apparent rate of ageing[53].

It has therefore been suggested that individuals may age at different rates, and that this could be measured using molecular markers[54,55].

The idea of this underlying biological age and the existence of biomarkers of ageing was postulated by Baker and Sprott[56], they suggested that chronological age may not be the best predictor of health, mortality or functional age, but that biological age may be. They reasoned that if ageing is the cumulative effect of basic biological processes, then it should be possible to observe it change throughout life and to detect biomarkers of biological age, and if true, these biomarkers would be better predictors of future functional capacity than chronological age.

In the last eight years high dimensional omics assays have become the ideal and favoured source of potential biomarkers of biological age.

1.4 Measuring Biological age: Biological Ageing Clocks

1.4.1 First Generation Clocks

The earliest attempts to construct models that predicted chronological age used DNA methylation data, as it has extensively documented age-related changes[57–64]. Bocklandt *et al.*[65] showed that two CpGs could explain 73% of the variance in chronological age, highlighting the fact that epigenetic markers could be used to construct models that estimate chronological age. More recent studies have gone a step further and built models that endeavour to estimate biological age.

Hannum *et al.* were the first to build an ageing "clock" that could be used to estimate a measure of biological age[55]. Again, using DNA methylation data, Hannum *et al.* went further and used the ratio of predicted age to chronological age to create the measure of "apparent methylomic rate of ageing" (AMAR). This measure could be used to identify individuals whose predicted age showed a larger than expected deviation from chronological age and hence was the first estimate of biological age. Hannum *et al.'s* clock is limited by the fact that it was trained only using DNA methylation data from adult whole blood, thus providing biased estimates for children and samples from different tissues.

In contrast Horvath's epigenetic clock, published the same year as Hannum *et al.*, was trained using methylation data from 51 different tissues, with samples ranging from foetal cord blood to centenarians[66]. Horvath's clock age, DNAme age, displayed four additional promising properties as a potential measure of biological age. First, embryonic stem (ES) cells and induced pluripotent (iPS) cells had a DNAme age of zero. Second, DNAme age was significantly correlated with cell passage number. Third, age acceleration, defined as the deviation between DNAme age and chronological age, exhibited a high heritability estimate. Fourth, DNAme age correlated with chronological age in chimpanzee tissue[66].

The high accuracy and broad applicability of Horvath's epigenetic clock means that it is one of the most extensively studied ageing clocks. Differences in DNAme age acceleration between cases and controls have been found in: neuropathology in the elderly[67], Down syndrome[68], Parkinson's Disease[69], Werner syndrome[70], Huntington's Disease[71], Alzheimer's Disease[72], physical and cognitive fitness[73], development[74], HIV1 infection[75], osteoarthritis[76], menopause[77] and centenarian status[78]. DNAme derived measures of age acceleration are statistically significant predictors of life expectancy[79]. The fact that DNAme age acceleration was associated with multiple age-related conditions and disease phenotypes suggests that biomarkers of ageing can be found and emphasises the potential of ageing clocks to stratify individuals for health risks. These results highlight the potential of ageing clocks as a valuable avenue of precision medicine research.

1.4.2 Other Omics Clocks

Since the first biological ageing clocks were published, studies have used a variety of different high-dimensional omics assays as sources of biomarkers and different statistical methods to build ageing clocks. Clocks trained on chronological age have since been constructed using telomere length[80], facial morphology[81], neuro-imaging data[82–85], metabolomics[86], glycomics[87], proteomics[54,88,89] and immune cell counts[90].

1.4.3 Limitations of Chronological Ageing Clocks

These studies uncovered limitations of omics ageing clocks trained on chronological age. First, that including enough predictors in the model can create a perfect predictor of chronological age. Lehallier *et al.* showed that the correlation between clock predicted age and chronological age increases with the number of proteins included in the model[91]. Further, it is possible to explain 100% of the variance in chronological age using DNA methylation data, by fitting all CpG sites simultaneously[92]. Returning estimates of biological age that equal chronological age is a major concern given that the purpose of estimating a measure of biological age is to be able to distinguish health outlook between individuals of the same chronological age.

Second, it is vital to ensure that these age acceleration measures derived from omics ageing clocks are capturing true underlying biological age, rather than being mere statistical artefacts. The most common approach used to assess this in previous omics clocks studies is to test if the omics clock age acceleration (OCAA) measure is associated with health outcomes for example, all-cause mortality, prevalent disease and incident disease. Often when OCAAs are associated with health outcomes, the effect sizes are small after adjusting for chronological age[93–95]. This difficulty in proving that OCAAs are capturing biological age encouraged a shift towards biological ageing measures that are more outcome focussed, rather than training on chronological age.

1.4.4 Second Generation Clocks

Levine *et al.'s* measure, DNAm PhenoAge, was the first to use a more sophisticated approach, regressing CpGs on a novel measure, phenotypic age, as the dependent variable[96]. Phenotypic age was estimated using regularised Cox regression, a model containing 42 clinical markers and chronological age was trained on 10-year all-cause mortality. Elastic net regression and DNA methylation data were then used to create DNAm PhenoAge, that outperformed previous epigenetic clocks by more accurately predicting all-cause and disease specific mortality.

More recently still, Lu *et al.*'s measure DNAm GrimAge comprises DNA methylation surrogates of 7 plasma protein levels, a DNA estimator of smoking pack years, chronological age and sex[97]. The resultant age acceleration measure outperformed previous epigenetic clocks as a predictor of all-cause mortality, time-to-cancer, time-to-coronary heart disease and was associated with number of co-morbidities.

GrimAge also remained prognostic of all-cause mortality after adjusting for traditional clinical risk factors.

1.4.5 Clock Comparisons

Comparisons of multiple different omics ageing clocks have been carried out, however, these studies focused on DNA methylation, traditional risk factors and frailty-based measures[93–95,98,99]. One common finding was that second-generation DNA methylation clocks, DNAm GrimAge and DNAm PhenoAge, were the most prognostic of mortality and incident disease, outperforming both first-generation epigenetic clocks and those built from other omics assays, trained on chronological age. However, after adjusting for chronological age, effect sizes of age acceleration measures still tended to be modest[93–95].

1.4.6 Summary

Despite all of the research into biological age, there are still many unanswered questions. There has been a slight shift, with more studies using mortality-based outcomes to train omics ageing models instead of chronological age. However, there has been little work done to fully establish the properties of different single omics ageing clocks trained on chronological age. Are the OCAAs from the numerous statistical approaches and sources of omics biomarkers actually capturing biological age? Is there only one single underlying biological age or are there multiple biological ages, each tracking susceptibility to different diseases or for different organ systems? There has been limited work to address this. Comparisons of multiple omics ageing clocks have been performed as mentioned previously, however, comparisons were often limited to few different omics sources. Commonly these clocks were only assessed based on their variance explained in chronological age and their ability to predict mortality or disease phenotypes. Few further investigations were made once a significant p-value for association with a health outcome had been established.

A comprehensive comparison and characterisation of multiple omics assays as potential sources of biomarkers of biological age would address these gaps in the knowledge. The uniquely broadly phenotyped cohort, the Orkney Complex Disease Study (ORCADES), annotated with 9 different omics assays would provide the ideal opportunity for this type of analysis.

Further characterisation of omics ageing clocks is essential as we need to understand what these OCAA measures are actually capturing, if the aim is to construct measures of biological age that have the potential to be clinically useful. If found to be so, omics ageing clocks could be used for personalised medicine and to potentially reverse the effects of ageing; there is already preliminary evidence for this being possible[100]. Not only could this prompt interventions that mean more people live healthier for longer, but it could also give us a greater understanding of the ageing process and determine if there is indeed an underlying rate of ageing.

1.5 Omics Biomarkers of Disease

In addition to biological ageing, omics measures from high-dimensional assays are a potential source of biomarkers for incident disease directly.

The ability to predict future health outcomes would aid precision medicine and drastically improve quality of life. Biomarkers or risk scores could be used: to predict the likelihood of successful outcomes after surgery or other procedures, to predict response to different therapeutic options, to gain insight into whether a disease is likely to recur, relapse or metastasise, to inform how regular preventative screenings are required to be and to give patients the opportunity to make lifestyle changes to lower their risk of developing morbidities in the future.

Motivated by the numerous potential benefits of effective biomarkers, this is an extensive area of study. Historically, research into biomarkers has focussed on traditional clinical measures to build risk scores, often for very specific clinical outcomes[101–103]. In fact, there are many whole sub-fields dedicated to optimising risk scores for several of the purposes outlined above: sub-phenotypes, disease recurrence[104], outcome post-surgery[105,106] and response to different therapies[107].

However, more recently, as assay technology and computing have advanced so have the sources of data used in developing risk scores. These include high-dimensional omics assays[108–111] and image analysis[112]. Statistical methods have also progressed and now stepwise regression[108], penalised regression[113], random forest[110], neural networks[83] and deep learning[112] are used in the construction of risk scores.

Omics assays that overlap with those measured in ORCADES have been associated with several different health outcomes. Menni *et al.* found that 46 IgG glycans were associated with the 10-year atherosclerotic cardiovascular disease (ASCVD) risk score[108]. Two eleven protein signatures have been shown to be prognostic of cervical[114] and ovarian[109] cancer respectively. Gisby *et al.* reported that the levels of 203 individual plasma proteins and a multivariable score were associated with clinical severity of COVID-19[110]. Pietzner *et al.* showed that plasma metabolite levels were prognostic for numerous incident diseases and of multimorbodity[111].

Given the unique range of omics assays, clinical phenotypes and incident diseases through electronic health records (EHR) that are available in ORCADES, it offers the potential to explore the suitability of these omics measures as biomarkers of health-related risk factors, clinical risk scores and incident disease. Such an investigation has the potential to provide greater insight into disease aetiology, as well as highlighting biomarkers that are prognostic of disease phenotypes which contribute towards the global burden of morbidity and mortality.

1.6 Leveraging Genetic Architecture of Protein Biomarkers

As mentioned previously, genome-wide association studies (GWAS), despite having discovered thousands of associated loci, have posited very few plausible mechanisms of action for these variants[23]. Similarly, recent techniques that allow testing for causal relationships between complex traits, using associated SNPs as instruments, again do not necessarily provide mechanistic insight[23]. However, although omics biomarkers might be consequences, they could also be causes of a mechanism and in any case offer insight into the underlying biology.

This, paired with the increase in availability of high dimensional omics data, means it is natural that the field of statistical genetics would take an interest in omics measures and apply the techniques used in studying complex trait genetics. GWAS and other common downstream analysis leveraging genetic information have been performed on proteomics[115–121], metabolomics[122], lipidomics[123], transcriptomics[124] and DNA methylation levels[125].

Proteomics are a natural starting point, as proteins circulating in the plasma are closer to end-point phenotypes than other omics biomarkers. Despite the inexorable link between transcript and protein level, the two measures are not always highly correlated[126,127], meaning that additional information can be gained by studying protein levels beyond transcript levels. Further, it is often protein dysfunction and dysregulation that lead to disease phenotypes and proteins circulating in the plasma are druggable targets[37–39,128–130].

It is only through the power of techniques used in statistical genetics that we are able to pose interesting research questions such as: how these layers of biology interact, how they are regulated and how does their dysregulation lead to disease. Given their potential impact on advancing our understanding, these techniques are therefore worth taking the time to introduce.

1.6.1 Genome Wide Association Studies

As mentioned previously, genome-wide association studies (GWAS) assess how variation in a trait of interest maps to a specific region in the genome. Historically this was done using linkage analysis[131], but is now achieved using genotype or sequencing data. As technology has advanced since the first human reference genome, the human genome project[15], there are now more comprehensive reference panels such as 1000 Genomes[132] and the Haplotype Reference Consortium[133] (HRC). Similarly, microarray technology has progressed in recent years from measuring hundreds of thousands to over a million SNPs across the genome[134]. Imputation of the allele dosages of SNPs not directly measured on the array using a reference is now a common strategy to increase coverage of the genome. This is possible using the knowledge from reference genomes of the patterns of variants that are correlated with one another - linkage disequilibrium (LD).

A GWAS therefore comprises millions of linear associations of the allelic dosage of single SNPs with a phenotype of interest while taking into account covariates such as relatedness of individuals and population stratification. This is done using linear regression for continuous traits or logistic regression for binary traits such as disease case-control status as shown:

$$\boldsymbol{y} = \mu + \boldsymbol{X\beta} + \boldsymbol{g} + \alpha SNP + \epsilon$$

Where $\boldsymbol{y}$ is an $n \times \mathbf{1}$ vector of trait values for n individuals, μ is the intercept, $\boldsymbol{X}$ is an $n \times p$ matrix of p covariates with the associated $\boldsymbol{\beta}$, a $p \times \mathbf{1}$ vector of fixed effects, $\boldsymbol{g}$ is a vector of random effects based on a genomic relationship matrix (GRM), that can be calculated from genotype data[135] and ϵ is an $n \times \mathbf{1}$ vector of residual deviations assumed to be distributed independently of the random genetic effects[136]. SNP is the allelic dosage of the SNP coded as number of effect alleles (0, 1 or 2) compared to a reference (e.g. AA, AG, GG, with reference to A) if genotyped and any value between 0 and 2 if imputed, α is the effect size of the SNP on y in outcome units per effect allele[136]. For each linear SNP-outcome association the hypothesis test is whether α differs from zero.

However, there are three reasons why GWA studies are not quite as simple as performing millions of single SNP associations. First, it is crucial to account for relatedness between the individuals in the sample, by including random effects from the GRM in the model. As to fail to do so ignores the covariance in $\boldsymbol{g}$ between relatives and the covariance between $\boldsymbol{g}$ and SNP as a result. Even in unrelated samples it is necessary to take into account population structure: this is most commonly done by including principal components of the genotypes as covariates. Failure to account for relatedness and wider population structure could result in confounded associations being used to draw incorrect conclusions about the biology of the trait of interest. However, there are now tools that account for relatedness between individuals[137–141] and allow GWA studies to be performed in related populations.

Second, these millions of linear associations are not independent, as each SNP is not independent, due to linkage disequilibrium (LD)[142]. LD is the property of the genome whereby alleles at SNPs in close physical proximity are correlated (because they co-segregate with each other). This is both a blessing and a curse as on the one hand, as often up to hundreds of SNPs will be in LD with a given common variant, it is therefore not necessary to have genotyped (or imputed) the causal SNP, just one in LD with it. On the other hand, all of the SNPs in LD with the causal variant will be associated with the trait of interest, making it more difficult to narrow down and determine the true causal variant.

Third, there is the issue of practicality, even with the power of modern computing, performing millions of mixed models is extremely computationally intensive. Rather than fit both fixed effects covariates and random components for every SNP, like some tools[143–145], it is more computationally efficient to split this into two separate steps. Step 1 being the correction of the trait value for covariates and random effects while step 2 involves performing per-SNP linear associations with the corrected trait values. There are now a variety of tools that use a two-step approach with different statistical methods to correct the trait values[137–141] which have massively improved the computational efficiency of performing GWA studies.

GWAS offer the opportunity to find genetic variants that are associated with the variation in the levels of omics biomarkers circulating in the blood, provide insight into how their levels are regulated and highlight regions of the genome in which to search for the underlying causal variants. With this technique, we are able to search for genes and pathways that underpin these different omics and to understand how these omics layers interact. The ability to answer these questions however depends on the power and therefore sample size of the GWAS.

1.6.2 Meta-analysis

Due to the expense of omics assays, cohorts with measures in large sample sizes are rare. In order to increase power to detect variants with modest effect sizes[146], GWAS run on the phenotype of interest in different cohorts are combined, thus increasing the sample size. This is an efficient strategy as it negates the need to share individual level data and with the use of a consistent reference panel across studies, genotypes can be imputed to provide a common set of SNPs for analysis.

The most common approach for genome-wide association meta-analysis (GWAMA) is to combine evidence from multiple studies weighted on the inverse variance of the estimated effect size for each SNP, and is implemented by the software METAL[147]. Inverse variance-weighted meta-analysed parameter estimates for each SNP are calculated as follows:

$$w_i = 1/SE_i^2$$

Where w_i the weight calculated for study i is based on the standard error of the effect estimate, SE_i, for study i. The meta-analysis effect estimate β is:

$$\beta = \sum_i \beta_i w_i \Big/ \sum_i w_i$$

Where β_i is the effect size estimate for study i. Meta-analysis standard error (SE) and P-value are calculated as follows:

$$SE = \sqrt{1 \Big/ \sum_i w_i}$$

$$P = 2\Phi(-|Z|)$$

Where Φ is the cumulative distribution function of the normal distribution and $Z = \beta/SE$.

As GWAMAs are an extremely effective way of increasing sample sizes available for studies, they have been used extensively to increase the number of trait-associated loci discovered for a diverse range of phenotypes[19,148]. This is particularly the case with proteomics, where the number of protein quantitative trait loci (pQTL) discovered has increased from 79 from a single cohort study[115] to 451 from a meta-analysis of the same 90 proteins[149]. GWAMAs therefore offer the potential to increase our ability to discover genetic loci involved in the regulation of the levels of omics biomarkers. The larger sample sizes increase power for downstream analysis, allowing the interrogation of potential mechanisms of action of associated genetic variants on the levels of omics biomarkers and the inference of causal relationships between omics biomarkers and disease.

1.6.3 Conditional Analysis

Both GWAS and GWAMAs estimate the effect size for each SNP in the study, on the trait of interest, however SNPs are not independent due to LD, arising from the fact that SNPs in close proximity are inherited together more often than would be expected if they were unlinked. Therefore any SNPs in high LD with a causal variant will also be associated with the trait of interest and could have p-values passing genome-wide significance (5×10^{-8}). Physical distance or LD clumping are commonly used to define significant loci: both of these approaches report the SNP with the lowest p-value in the region as tagging the signal, however, this still does not mean that the top SNP is causal[150]. This actually assumes that the top SNP captures the maximum amount of variation in the region by its LD with an unknown causal variant and that the other SNPs in the vicinity show association due to their correlation with the top SNP[150].

There are issues with this assumption, first, even if there is a single underlying causal variant, there is no guarantee that one of the genotyped or imputed SNPs may capture the overall variation at the locus[151,152]. Second, there may be multiple causal variants in a single locus and by reporting only one top SNP the variation explained by that locus may be underestimated[150].

Conditional analysis is a method commonly used to overcome these limitations, by performing association analysis while conditioning on the primary associated SNP at a locus, to test for the presence of conditionally associated SNPs. The tool, GCTA-COJO[150], performs approximate conditional and joint analysis using GWAS summary statistics and has become the gold standard method for discovering conditionally associated loci. Numerous GWAS on plasma protein levels have used GCTA-COJO to find secondary pQTL[116,149,153]. Conditional analysis facilitates the discovery of additional associated variants, allowing a more accurate picture of the genetic regulation of the trait of interest. This also offers further potential to find genes and pathways that lead to the end-point phenotype.

1.6.4 Heritability

As we aim to increase our understanding of how and why disease phenotypes arise, it is advantageous to determine how much of a given trait is determined by genetics and how much is driven by the environment. Heritability (h^2) is the estimate of the proportion of variation in a phenotype (V_P) that can be attributed to the additive effects of genetic variation (V_A), as opposed to environmental factors or chance.

$$h^2 = \frac{V_A}{V_P}$$

The knowledge that a particular disease has a high heritability, for example Mendelian diseases such as cystic fibrosis[154], means that an individual’s risk of developing said disease is determined at conception. Knowing this offers the opportunity for genetic counselling and will inform aspects of an individual at high risk's medical care. For example, how often they receive screenings, not only for the individual in question but also their close relatives. In contrast, the knowledge that a particular disease has a low heritability, and is in fact predominantly driven by environmental factors, offers the potential for changes in policy that may reduce the risk of disease across large sections of the general population, particularly if socio-economic status- or lifestyle-related factors contribute to disease risk.

Once the heritability of a trait has been estimated, it is interesting to deconstruct this further and assess how much of the genome is contributing to this heritability. Is the trait polygenic, meaning there are hundreds or even thousands of loci across the genome each contributing a small amount to the trait variation, as is the case with height[19], or are there very few loci or even a single locus that is entirely responsible for the variation in the (monogenic) trait, as with Mendelian disorders. An understanding of the genetic architecture of a trait in this way is useful, for example, in resource allocation when planning genetic studies, but also has further downstream applications.

There are numerous methods to estimate heritability from individual-level genetic data[145,155,156], however in practice it is common to employ methods that use GWAS summary statistics, as the larger sample sizes in publicly available meta-analyses from large consortia provide more power to produce more accurate estimates. However, the common method used to estimate heritability using GWAS summary statistics, Linkage Disequilibrium Score regression (LDSC)[157] assumes that the trait is polygenic[149]. While polygenicity is expected and has indeed been shown for many complex traits[158], Folkersen *et al.* highlighted that it may not be true of omics biomarkers such as plasma protein levels. Particularly as some proteins show a single extremely strong *cis*-association signal[149] (*cis* meaning proximal to the coding gene and *trans* being distant **Figure 1**).

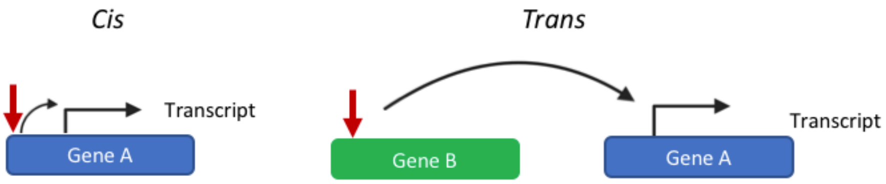


Figure 1. Cis & Trans Associated pQTL.

To overcome this assumption, Folkersen *et al.* estimated the heritability contributed by significant pQTL (pQTL component) and the remaining genome-wide SNP heritability (polygenic component), separately for each protein[149]. LDSC was used to estimate the polygenic component contributed by the SNPs not passing genome-wide significance. The pQTL component was calculated as the sum of the estimated variance in protein level explained by the lead SNPs:

$$h^2_{pQTL} = \sum 2p_i q_i \beta_i^2$$

Where β_i^2 is the estimated effect size of the lead variant of pQTL i, p_i is the minor allele frequency and $q_i = \mathbf{1} - p_i$. Folkersen *et al.* showed that the genetic architecture varies enormously between proteins, with some such as platelet derived growth factor subunit B (PDGF-B) being extremely polygenic but Interleukin 6 receptor subunit alpha (IL-6RA) being almost monogenic[149].

It is therefore extremely informative to assess both the total heritability of traits as well as how the contributions to that heritability are spread across the genome. In addition to providing an insight into epidemiology and giving us a better understanding of how the trait is regulated by genetics, this can also indicate targets for therapeutic intervention.

1.6.5 Genetic Correlations

It is possible to estimate how much genetic architecture two traits share, by calculating the genetic correlation between the two traits.

As an individual's phenotypic value (P) is equal to the sum of their genotypic value (G), the component of this value associated with genetic effects, and their environmental deviation (E), the component of phenotypic value not due to genetics.

$$P = G + E$$

The correlation of the genetic components of two traits (X and Y), r_{gxy}, is therefore:

$$r_{gxy} = cor(G_x, G_Y)$$

Genetic correlations between traits are often due to pleiotropy, the property where a single gene influences more than one trait. Genetic correlations capture the effect of all the segregating genes across the whole genome that affect both traits[159]. Positive correlations are observed when genetic variants act to increase both traits, in contrast negative correlations are observed when variants increase one trait but decrease the other[159]. This measure therefore indicates how SNP effects, genome-wide, align between two traits.

Practically, genetic correlations provide insight into the relationships between phenotypes and can be used to prioritise investigations into potential causal relationships. In terms of plasma protein levels, determining whether they are genetically correlated with disease phenotypes has the potential to inform epidemiology and increase our understanding of disease aetiology.

Estimating genetic correlations, however, is not a simple matter, as in reality it is not possible to accurately measure an individual's genotypic value (G). As the phenotypic component due to genetics (G) is itself a combination of the contributions from additive genetic effects (A), dominance effects (D), and genetic interaction effects (I):

$$G = \mathrm{A} + \mathrm{D} + \mathrm{I}$$

So, the genetic correlations quoted in the literature are actually estimates of the correlation of the additive genetic effects between two traits:

$$r_{Axy} = cor(A_x, A_Y)$$

However, these estimates are still commonly written r_g and referred to as genetic correlations despite being an estimate of genetic correlation due to additive genetic effects, as it is not possible to calculate true r_g. For consistency with the literature, I will only use the terms r_g and genetic correlations from this point on but note that they are referring to what is more accurately r_A.

There are methods that estimate r_g from GWAS summary statistics using the following approach[157,160,161]:

$$r_g = \frac{cov_G}{\sqrt{h_X^2 h_Y^2}}$$

Where the genetic correlation, r_g, is the covariance, cov_G, of the genetic effects of the two traits divided by the square root of estimated heritabilities of the traits X and Y. As genetic correlations are influenced by allele frequencies and therefore will vary between populations with different patterns of LD[159], it is necessary to take LD information into account when estimating r_g. The previous gold standard method for estimating genetic correlations from summary statistics, LD-score regression (LDSC)[157,160], uses LD information as does the recent method proposed by Ning *et al.*, high definition likelihood[161].

Genetic correlations between numerous traits have been reported[160,162], including previous pQTL studies, showing that plasma protein levels share genetic architecture with multiple phenotypes (Shen *et al.* Unpublished).

1.6.6 Mendelian Randomisation

While genetic correlations can indicate when two traits share architecture and that the genome-wide genetic effects correlate, they cannot tell us anything about the direction of the relationship between the two traits, they do not indicate causality.

The limited ability to assess causality between traits is a huge hurdle in biology. In order to form a comprehensive picture of how disease phenotypes arise, we want to know whether biomarkers, either molecular such as LDL cholesterol levels or complex traits such as blood pressure, are causes or consequences of disease.

Traditionally causality between two traits was only able to be assessed using randomised controlled trials (RCT), where participants are randomly assigned into either the control group or a group that received the intervention or exposure. Such studies are what is referred to as double blind, as both the participants and the researchers are unaware to which group participants have been assigned. In all other respects the groups are treated the same, therefore minimising sources of experimental bias.

Mendelian Randomisation (MR) is a technique that provides the opportunity to infer causal relationships between two traits using GWAS summary statistics and has revolutionised the field of genetic epidemiology. MR as we know it today, was first put forward in 2003[163,164] and enables the estimation of the causal effect of an exposure on an outcome, without the need to conduct an RCT. MR further allows a broader array of exposures to be investigated, as it would be neither possible nor ethical to subject research participants to exposures which are known health risks. Additionally, many apparently robust observational associations do not deliver the anticipated results when assessed using an RCT[165]. Unlike traditional epidemiological studies, which are hampered by confounding and reverse causation, the use of genetic data has several attractive advantages. Firstly, that as genotypes are assigned randomly when passed from parent to offspring during meiosis, they can be considered free from confounding from environmental factors that could influence exposure or outcome. Secondly, in accordance with the central dogma, as genotypes are not affected by potential exposures or outcomes, they are not subject to reverse causality. As subsets of individuals in a study population with differing number of exposure-modifying alleles at a SNP can be thought of as having been randomised to receive a different level of an exposure during their lifetime, there is effectively a naturally occurring RCT[166,167].

MR estimates the causal effect of an exposure on an outcome using SNPs as instrumental variables (IVs)[168] of the exposure. The two-sample MR approach[169] which uses GWAS summary statistics from two different studies for exposure and outcome is more commonly used, given the amount of publicly available GWAS summary statistics, particularly large meta-analyses from consortia. It is increasingly rare that a single study would have both outcome and exposure measured in a sample size large enough to rival those published. The TwoSampleMR R package and the associated database of exposure and outcome GWAS, MRBase, have become the most common tools for performing MR[170]. The basic method works as follows:

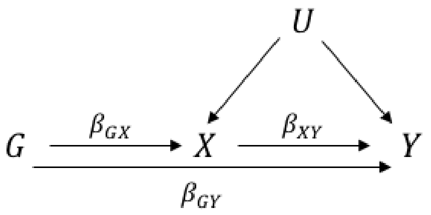


Figure 2. Mendelian Randomisation DAG.

As shown in **Figure 2** the effect of the IV (G) on the exposure (X) is β_{GX}, the effect of the IV (G) on outcome (Y) is β_{GY}, the causal effect of exposure (X) on outcome (Y) is estimated as the effect of IV on outcome divided by the effect of the IV on the exposure.

$$\beta_{XY} = \frac{\beta_{GY}}{\beta_{GX}}$$

This example, using a single SNP IV is the simplest MR method, the Wald ratio[171]. Inverse variance weighting (IVW) is used to combine ratio estimates from multiple independent (i.e. not correlated) IVs:

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{m} w_j \hat{\beta}_j}{\sum_{j=1}^{m} w_j}$$

Where w_j is the inverse variance of the ratio estimate for IV j, $\hat{\beta}_j$. The Wald ratio and IVW are the most common MR methods.

The MR approach relies on three assumptions[172] (**Figure 3**). First, that IVs are strongly associated with the exposure. Second, that IVs are not affected by unmeasured confounders of the exposure and outcome. Third, that IVs are not associated with the outcome via any other path than through the exposure. Any violations of these assumptions introduce bias into the results. For example, horizontal pleiotropy, when IVs affect the outcome via a different biological mechanism, other than through the exposure being considered, is a major source of biased causal effect estimates from MR[163,164]. Similarly, if there are differences in the single SNP causal effect estimates between multiple instruments, instrument heterogeneity, this could indicate that a proportion of the IVs selected are invalid, resulting in the overall causal effect estimate of exposure on outcome being biased[173].

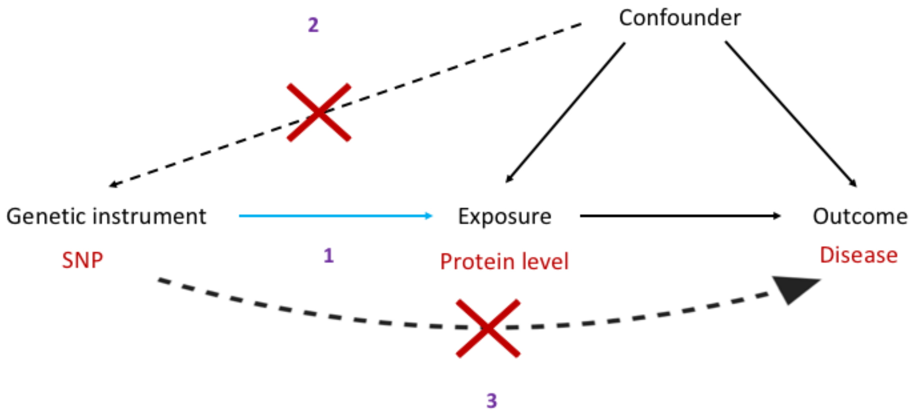


Figure 3. Assumptions of Mendelian Randomisation. *MR DAG with causal paths indicated by solid lines, assumptions numbered in purple, causal effects that violate assumptions in dashed lines, example of analysis of the effect of protein level on disease in red.*

As MR is extremely popular, new methods to overcome these limitations are constantly being developed such as MR-Egger which provides causal effect estimates even with invalid instruments. The MR-Egger intercept is also used to detect horizontal pleiotropy[174–176]. Maximum likelihood[177], weighted median[173] and weighted mode approaches are more robust to IV heterogeneity, due to allowing a proportion of IVs selected to be invalid.

Together these methods allow us to pose research questions that, without the ability to assess causality between traits, we were not equipped to answer.

Plasma protein levels are of particular interest when it comes to investigating potential causal relationships due to their: proximity to end-point phenotypes, potential to drive mechanisms that result in disease and the fact that they are druggable targets. For these reasons, dedicated research has been done on the use of pQTL as IVs, suggesting the following additional sensitivity analysis should accompany IVW MR[178]. Namely, that filters based on heterogeneity between instruments and MR-Egger intercept be applied when using multiple IVs to limit the chance of pleiotropy influencing results. Similarly, that evidence of reverse causality, here referring to the outcome having a causal effect on the exposure, should be assessed either by performing bidirectional MR if the full summary statistics for the exposure are available or the Steiger test for directionality[179] if not. Finally, that colocalisation (outlined below) should be used to provide additional evidence that exposure and outcome share a causal variant.

Previous pQTL studies have performed MR using pQTL as IVs with numerous outcomes. These studies have recapitulated reported causal relationships, replicated findings from randomised controlled trials[149] as well as highlighting several circulating plasma proteins such as PAPPA, F3 and SPON1 as novel drug targets[149] and opportunities for drug repurposing such as Denosumab, that targets RANKL, for treatment of Paget's disease[116].

These studies highlight the exciting discoveries possible using pQTL as IVs for MR. However, the number and robustness of causal relationships inferred depends on the strength of the instruments, which in turn depends on the power of the exposure GWAS. This emphasises the potential possible with even larger sample sizes. Well powered MR studies using pQTL as instruments offer the potential to infer causal relationships that would improve our understanding of disease aetiology, discover causal biomarkers of disease and pinpoint novel therapeutic targets for intervention.

1.6.7 Colocalisation

It is relatively simple to discover genetic variants significantly associated with variation in a trait of interest, however it is much more difficult to determine how these variants, via which genes and pathways, act to actually impact the trait. It is therefore of great interest to be able to determine if gene expression and a trait of interest share causal variants, although this can also be assessed for pairs of complex traits that are thought to be related.

This is of particular interest when studying proteomics, as the most direct biological mechanism for pQTL would be to alter the gene expression of the coding gene, thus also being an expression quantitative trait loci (eQTL).

There are several possible approaches to see if pQTL overlap with eQTL. The most straightforward is to see if the lead variant (top SNP) of the pQTL has been previously reported as being associated with expression of the relevant gene using publicly available eQTL datasets (e.g. GTEx)[180–182]. However, as mentioned previously, there is no guarantee that the top SNP is indeed the causal SNP for both traits, therefore formal tests for colocalisation are required.

The summary data-based Mendelian Randomisation (SMR) [183] method proposed by Zhu *et al.* identifies genes whose expression levels are associated with a complex trait using GWAS and eQTL summary statistics. SMR estimates the effect of gene expression (exposure, X) on trait (outcome, Y), β_{XY}, using a single SNP instrument (G):

$$\beta_{XY} = \frac{\beta_{GY}}{\beta_{GX}}$$

Where β_{GY} is the estimated effect of the SNP on outcome and β_{GX} is the estimated effect of SNP on exposure. A limitation of using single SNPs is that this approach is usable to distinguish between causality and pleiotropy (**Figure 4**). In order to distinguish between pleiotropy and linkage, the SMR software performs a Heterogeneity in dependent instruments (HEIDI) test. If the association is due to pleiotropy, all SNPs in LD with the causal SNP will have the same estimated β_{XY}. Therefore, testing against the null hypothesis of pleiotropy due to a single causal variant, is equivalent to testing if estimates of β_{XY} for multiple SNPs in the *cis*-eQTL region are heterogeneous.

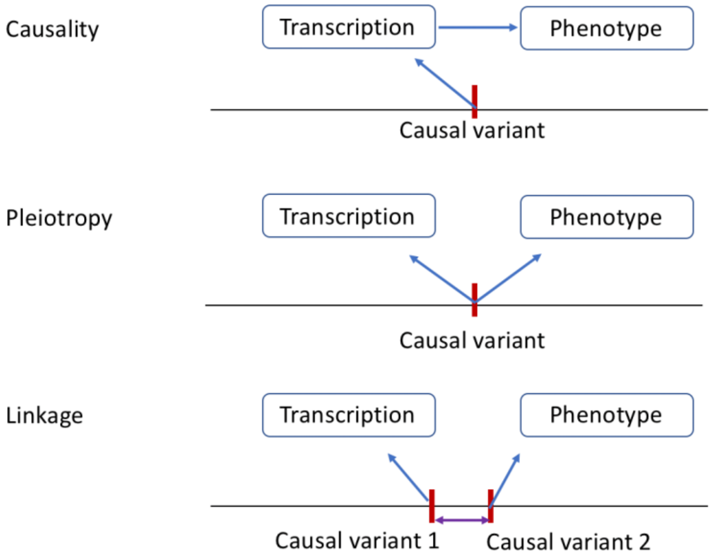


Figure 4. Causality, Pleiotropy or Linkage.

In contrast, the more recent colocalisation approach coloc[184] utilises a Bayesian method to assess whether two association signals from two different traits are consistent with a shared causal variant. Like SMR, coloc uses summary statistics rather than individual level data.

Using a set of (Q) variants common to both datasets for which minor allele frequencies, effect sizes, standard errors are known, coloc creates vectors of length Q of (0,1) values for each trait, where a value of 1 indicates that the variant is causally associated with the trait of interest. All possible pairs of vectors or “configurations” are assigned to one of the following five hypotheses:

H_0 = No association with either trait
H_1 = Association with trait 1, not with trait 2
H_2 = Association with trait 2, not with trait 1
H_3 = Association with trait 1 and trait 2, two independent SNPs
H_4 = Association with trait 1 and trait 2, one shared SNP

Using a Bayesian framework, coloc combines SNP level prior probabilities with the probability of the data at the configuration level, summed over all configurations to calculate a posterior probability for each of the 5 hypotheses. The posterior probability of H4 (PPH4) is of most interest as a large PPH4 suggests the two association signals are likely to share a causal variant and colocalise.

This approach assumes that the population samples used in the two studies contain unrelated individuals and that these samples are drawn from populations with the same ancestry, specifically that both allele frequencies and patterns of LD are shared. It also has limitations, firstly it assumes that the causal variant is included in the Q SNPs included in the analysis and secondly that there is only one causal variant in the region being considered.

These methods have been used in previous pQTL studies to look for evidence of pQTL acting via gene expression to cause the changes in the levels of circulating plasma proteins measured. Folkersen *et al.* found 125 associations between 96 genes and the levels of 54 proteins using SMR-HEIDI[149]. Sun *et al.* found that 78.5% of their testable *cis*-pQTL showed strong evidence (PPH4>0.8) of colocalisation with eQTL in at least one tissue[185], suggesting these pQTL influence the level of the protein circulating in the plasma by altering the transcript level.

1.6.8 Summary

Previous GWAS of plasma protein levels have utilised these methods leveraging genetic data that I have outlined. Thousands of genetic variants associated with variation in plasma protein levels[115–121], complexities of underlying genetic architecture, overlap with eQTLs[116,149], causal relationships with disease[116,149,153] and candidate novel therapeutic targets[149] have been reported. These studies highlight the potential further discoveries possible with even larger samples sizes.

In combining GWAS of protein levels in ORCADES with summary statistics from 17 cohorts from the SCALLOP Consortium, I will apply these methods to the largest GWAMA of 184 protein levels to date. This increased power will provide greater insight into the genetic regulation of protein levels and how they relate to disease as well as creating a resource of pQTL data available for future use in the field.

1.7 Aims

Despite the recent advance of technology that produced high-dimensional omics assays, their expense means that large cohorts with multiple omics assays are rare. I took advantage of the unique range of omics data available in the Orkney Complex Disease Study (ORCADES) to address gaps and broaden the scope of several research areas.

In this thesis I exploited this resource of multi-omics data to investigate "Omics Measures of Ageing and Disease Susceptibility" in the following three diverse areas:

1. How multiple omics assays compare as sources of potential biomarkers of biological age
2. Investigating whether multi-omics biomarkers are associated with health-related risk factors or prognostic of incident disease
3. Leveraging the genetic architecture of proteomics to investigate the relationships between plasma protein levels and disease

As such research could: elucidate mechanisms of biological ageing and disease; uncover biomarkers of incident disease and biological ageing, providing the opportunity for potential intervention or even prevention, therefore aiding precision medicine; infer causal relationships between circulating omics biomarkers and health outcomes; and discover potential novel therapeutic targets or candidates for drug repurposing.

Answering these questions offers the potential to better our understanding of the underlying biology of ageing and disease. Only by understanding how biological mechanisms work end-to-end are we able to design interventions that will help reduce the burden of disease. Only by understanding why we are the way we are can we hope to improve our quality of life.

Chapter 2: Data & Methods

2.1 Introduction

2.1.1 Data

This thesis used overlapping sets of cohorts for 3 different sets of analysis, some of which involved me processing individual level data, some of which were processed by the cohort analysts and meta-analysed by me.

For each cohort used throughout this thesis **Table 1** indicates which omics assay and to which of the three sets of analyses (Chapters) they contributed, as well as stating who carried out the pre-processing and quality control of the raw cohort level data, whether it was another analyst or me.

Cohort	Omics Assay	Chapter	QC	Analysis
ORCADES	Genetics	5	🟩	🟪
	DNA Methylation	3,4	🟩	
	PEA Proteomics	3,4,5	🟪	
	UPLC IgG Glycomics	3,4	🟩	
	NMR Metabolomics	3,4	🟪	
	MS Metabolomics	3,4	🟪	
	MS Complex Lipidomics	3,4	🟪	
	MS Fatty Acid Lipidomics	3,4	🟪	
	DEXA	3,4	🟪	
	Clinomics	3,4	🟪	
Croatia-Vis	Genetics	5	🟩	🟪
	PEA Proteomics	3,4,5	🟪	
	UPLC IgG Glycomics	3	🟩	
Croatia-Korčula	UPLC IgG Glycomics	3	🟩	🟪
	NMR Metabolomics	3	🟪	
UK Biobank	Clinomics	3	🟪	🟪
	DEXA	3	🟪	
GS:SFHS	DNA Methylation	3	🟩	🟪
EGCUT	Genetics	5	🟩	🟩
	PEA Proteomics	3,5	🟩	
	NMR Metabolomics	3	🟩	
	DNA Methylation	3	🟩	
ASAP	Genetics	5	🟩	🟩
	PEA Proteomics	5	🟩	

ARISTOTLE	Genetics	5	🟩	🟩
	PEA Proteomics	5		
Biofinder	Genetics	5	🟩	🟩
	PEA Proteomics	5		
COSM-C	Genetics	5	🟩	🟩
	PEA Proteomics	5		
Epihealth	Genetics	5	🟩	🟩
	PEA Proteomics	5		
Fenland	Genetics	5	🟩	🟩
	PEA Proteomics	5		
FHS	Genetics	5	🟩	🟩
	PEA Proteomics	5		
IMPROVE	Genetics	5	🟩	🟩
	PEA Proteomics	5		
INTERVAL	Genetics	5	🟩	🟩
	PEA Proteomics	5		
LifeLinesDeep	Genetics	5	🟩	🟩
	PEA Proteomics	5		
MPP-RES	Genetics	5	🟩	🟩
	PEA Proteomics	5		
NSPHS	Genetics	5	🟩	🟩
	PEA Proteomics	5		
PIVUS	Genetics	5	🟩	🟩
	PEA Proteomics	5		
SMCC	Genetics	5	🟩	🟩
	PEA Proteomics	5		
ULSAM	Genetics	5	🟩	🟩
	PEA Proteomics	5		

Table 1. Cohort Data used in Thesis. *Cohort: indicating the cohorts used throughout the analyses presented in this thesis. Omics Assay: the omics assay used for analyses per-cohort. Chapter: the chapter in which the indicated data was used. QC: Indicating whether another analyst (green) or I (purple) carried out the pre-processing and QC of the raw data. Analysis: indicating whether another analyst (green) or I (purple) performed the cohort level analysis.*

Due to manuscripts being included in this thesis, descriptions of the data and methods are spread across multiple chapters, so I will here briefly outline where descriptions of cohorts, omics assays and both general and analysis specific quality control (QC) and pre-processing can be found.

In this chapter I will also briefly describe the cohorts for which I handled individual level data and performed the cohort level analysis, these are the cohorts used for the multi-omic analyses in chapters 3 and 4. Here I will also describe the called and imputed genotype data for the ORCADES and Croatia-Vis cohorts that contributed to the meta-analysis discussed in chapter 5.

Descriptions of the omics assays used for analysis in chapters 3 and 4 (all of the omics listed in **Table 1** bar genetics) are described in the methods section of the submitted manuscript that comprises chapter 3. However, a detailed explanation of the data pre-processing and QC procedures that were common across all omics assays and used to create the final omics datasets used for these analyses is provided in this chapter.

Both general and analysis-specific QC and pre-processing will be described, and it will be explicitly stated if the procedure described was carried out by another analyst or myself.

The remaining 16 cohorts that provided summary statistics for the meta-analysis are described in the supplementary information for the manuscript that comprises chapter 5 (provided in the Appendix for chapter 5 in this thesis).

2.1.2 Methods

Similar to the cohort data, due to the inclusion of manuscripts in this thesis, descriptions of the methods used are also spread across chapters, so again I will indicate where the methods are located.

As both chapters 3 and 5 contain manuscripts, the details of analysis-specific methods for these chapters are in the relevant results chapter methods sections. To keep consistency between results chapters, I have also outlined analysis-specific methods in the methods section of chapter 4.

However, the background of common methods used across chapters 3 and 4 is described here. Specifically, I detail three different penalised regression methods, the limitations in ordinary least squares regression that they overcome and discuss why specific approaches were selected for use in these analyses.

I will here also outline: the preparation of data for two Olink proteomics panels, the procedure I used to perform cohort-level GWAS of these plasma protein levels, the QC of cohort-level GWAS summary statistics from all contributing cohorts and the meta-analysis protocol in more detail than is present in the methods section of the attached manuscript in chapter 5.

Details of the post-GWAS analyses performed are provided in the methods section of the attached manuscript in chapter 5, however the background and motivation behind these methods have already been outlined in the introduction (**1.6 Leveraging Genetic Architecture of Protein Biomarkers**).

2.2 Cohorts

2.2.1 ORCADES

The Orkney Complex Disease Study (ORCADES) is a family-based, cross-sectional study that seeks to identify genetic factors influencing cardiovascular and other disease risk in the isolated archipelago of the Orkney Isles in northern Scotland. In order to participate individuals were required to have at least two grandparents from Orkney. 2,078 participants between the ages of 16 and 100 years were recruited from 2005-2011[186]. There is decreased genetic diversity, alongside enrichment for rare variants, in this population compared to Mainland Scotland, consistent with high levels of endogamy historically. Fasting blood samples were collected, and many health-related phenotypes and environmental exposures were measured in each individual. Crucially for my purposes, it is also extremely densely annotated in terms of omics assays, having DNA methylation, proteomics, lipidomics, IgG glycomics, metabolomics (both using nuclear magnetic resonance (NMR) and Mass spectrometry (MS)) and body composition measures from whole body scans measured.

ORCADES was the main cohort used in this thesis. It is the primary sample for analyses in chapters 3 and 4, and additionally contributes to the larger meta-analysis in chapter 5. DNA was extracted and the resultant genetic data was used in chapter 5.

ORCADES Genetic Data

A total of 2,267 samples had DNA extracted and were genotyped across three different Illumina arrays. Samples were removed if they had a call rate <98% or were identified as ethnic outliers, duplicates, gender mismatches or excess identity-by-state (IBS) sharing incompatible with the pedigree. After sample quality control, 854 samples remained on the Hap300, 301 on Omni1 and 1,067 on the OmniX. The number of markers and quality control filters applied across the three arrays are shown in **Table 2**.

	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 2. Genotype Quality Control in ORCADES. *The number of variants pre- and post-QC and the filters applied to each of the three different genotyping arrays.*

These markers were then phased using Shapeit[187] v2.r873 and duohmm[188] software and imputed to HRC.r1-1 (without INDELS) using the Positional Burrows-Wheeler Transform (PBWT) algorithm[189] on the Sanger imputation server. The final dataset used for analysis comprised 12,696,745 SNPs (NCBI Build b37) for 2,215 samples. The preceding analysis was performed by others in the group and the resulting set of imputed genotypes passed to me.

2.2.2 Croatia-Vis

The CROATIA Vis study comprises 1,008 Croatian volunteers, aged 18–93 years, who were recruited from the villages of Vis and Komiža on the Dalmatian Island of Vis during 2003 and 2004. Participants underwent a medical examination and interview, led by research teams from the Institute for Anthropological Research and the Andrija Stampar School of Public Health, (Zagreb, Croatia). All subjects visited the clinical research centre in the region, where they were examined in person and where fasting blood was drawn and stored for future analyses. Many biochemical and physiological measurements were performed, and questionnaires of medical history as well as lifestyle and environmental exposures were collected. Akin to ORCADES, Croatia-Vis is a population isolate and has a similar study design. Several omics assays measured in Croatia-Vis overlap with those measured in ORCADES, specifically: UPLC IgG Glycomics and three Olink proteomics panels (inflammation 1 and cardiovascular 2 & 3). I had access to individual level data for this cohort as it is managed by a collaborator at the University of Edinburgh.

These omics assays were used in chapter 3 to replicate omics ageing clocks trained in ORCADES. Together with genetic data, proteomics data from two Olink panels in Croatia-Vis were used to perform genome-wide association studies that went on to contribute to the genome-wide association meta-analyses that are the subject of chapter 5.

Croatia-Vis Genetic Data

Extracted DNA samples from Croatia-Vis were genotyped using the Illumina HumanHap300v1 array and called using Beadstudio-Gencall v3.0. A total of 317,509 markers were filtered based on: sample call rate (>97%), SNP call rate (98%), a HWE threshold ($<1 \times 10^{-6}$) and a MAF threshold (≥ 0.01). The 272,930 markers meeting these criteria were used for imputation using the HRC reference. Shapet v2.r873 and duohmm software was used to phase the genotypes after which the PBWT (Sanger server) was used for imputation with HRC.r1-1 reference panel (without INDELS). Genotypes were originally NCBI genome build b35 but were lifted over to build b37 prior to imputation.

The final dataset used for analysis comprised 9,477,884 SNPs (NCBI Build b37) for 985 samples. The preceding analysis was performed by others in the group and the resulting set of imputed genotypes, passed to me.

2.2.3 Croatia-Korčula

The Croatia-Korčula sample is part of the larger group of CROATIA population isolates that contains Croatia-Vis. It is a family-based cross-sectional study comprising ~2,800 individuals from the Dalmatian Island of Korčula and used the same recruitment strategy and study design as Croatia-Vis. The study has two omics assays that overlap with those measured in ORCADES and was used in chapter 3 to replicate UPLC IgG Glycomics and NMR Metabolomics ageing clocks trained in ORCADES. Like Croatia-Vis, this cohort is managed by collaborators at the University of Edinburgh, and I was able to access individual level data and perform the replication analysis myself.

2.2.4 Generation Scotland

Generation Scotland: Scottish Family Health Study (GS:SFHS)[190] is a family-based genetic epidemiology study with DNA, socio-demographic and clinical data from ~24,000 volunteers aged 18-98 recruited between 2006 and 2011. The study was designed to create a resource for the study of the genetics of health, disease and quantitative traits that are important for public health. Adults from across Scotland from a range of socioeconomic status areas were invited to participate based on GP registry, with the additional criteria that they had at least one first degree relative who could also participate. The sample is 59% female with 87% of participants born in Scotland with 82% of parents and 75% of grandparents also born in Scotland. Participants completed a pre-clinic questionnaire before an in-person visit. Biological samples, cognitive function, personality traits, mental health data and lifestyle information were collected from participants as well as the study having linked electronic health records.

The family-based nature and increased kinship of the study allows the investigation of heritability, parent of origin effects, rare alleles and linkage mapping of health and disease related traits. In this respect it is similar to the highly related ORCADES sample. GS:SFHS has DNA methylation data for ~5,000 participants and was used in my analysis in chapter 3 to replicate two epigenetic ageing clocks trained in ORCADES. Individual level data for the specific markers selected for inclusion in my ageing clocks were extracted by Rosie Walker and I performed the replication analysis.

2.2.5 UK Biobank

The UK Biobank (UKBB) cohort (described in detail in Sudlow *et al.*[191]) is a prospective cohort established to investigate the genetic and non-genetic determinants of diseases of middle and old age. It comprises over 500,000 participants aged between 40 and 69 years at recruitment, which spanned 2006-2010. This age range was chosen to balance participants being old enough to be able to record incident health outcomes in the first few years of follow up, but also still young enough for the initial assessment to capture exposures before they are influenced by age-related morbidities. Participants were assessed at 22 assessment centres around the UK aiming to capture the heterogeneity in ethnicity, socio-economic status and the urban and rural mix that exists in the general population. Baseline measures including questionnaires, interviews, physical and functional measures, blood, urine and saliva samples were taken in person at assessment centres. Additional questionnaires have subsequently been sent to subsets of participants to complete online. The cohort has extensive phenotypic data available spanning questionnaires, physical measures, sample assays, accelerometery, multimodal imaging and longitudinal follow up including linked electronic health and primary care records. Since recruitment finished in 2010 the study has focussed on increasing the number of exposures and health-related outcomes it measures.

The UK Biobank is an extremely useful resource due to its large sample size and therefore power and the fact that it is publicly available. At the date of analysis omics data was not yet available for UKBB, however it did have a subset of ~5,000 individuals with DEXA imaging derived phenotypes and the traditional health-related risk factors that comprise the Clinomics set of phenotypes in ORCADES. The UK Biobank was used in my analysis in chapter 3 to replicate the DEXA and Clinomics ageing clocks trained in ORCADES.

2.2.6 Estonian Biobank

The Estonian Biobank[192] is a volunteer-based sample of the resident adult (18+) population in Estonia recruited between 2002 and 2019. The study was set up to establish a biobank with biological samples and health records from a large representative sample of the population to allow the investigation of genetic, environmental and behavioural background of common diseases. Individuals were invited to volunteer via their GPs as well as through promotional events. After the most recent round of recruitment the total sample size is >150,000, with genetic data and biological samples available for ~50,000 participants. Baseline measures included a standard health examination from a GP, blood samples, interview, questionnaires on lifestyle, medical history, personality and diet. The study also has extensive omics phenotyping including metabolomics, epigenetics, whole genome sequencing and longitudinal data is available in the form of linked electronic health records and the re-examination of a subset of participants.

The cohort has several omics assays that overlap with those measured in ORCADES and Estonian Biobank data was used in chapter 3 to replicate several omics ageing clocks trained in ORCADES. Specifically, the NMR Metabolomics, two DNA methylation clocks and a PEA Proteomics clock constructed from a subset of the Olink panels available in ORCADES. As discussed in section **2.3.2 QC of Omics Data in Replication Cohorts** of this chapter, analysis was performed on the Estonian Biobank data by Nele Taba using a quality control and analysis pipeline that I wrote, to avoid the requirement for access to individual level data by me.

The Estonian Biobank also contributed GWAS summary statistics for the 184 Olink proteins on the cardiovascular II and cardiovascular III panels to the GWAMA discussed in chapter 5.

2.3 Quality Control of Omics Data

2.3.1 QC of ORCADES Omics Data

For analyses in chapters 3 and 4, data from each omics assay in ORCADES was processed using the quality control (QC) pipeline outlined in **Figure 5**. For each assay in turn, single value omics measures (those that had the same value across all individuals in the sample) and outliers for the omics measures themselves were removed based on per assay z-score thresholds (**Table 3**). Omics data was then merged with outcome and covariate data. Age at venepuncture was the only outcome in analyses in chapter 3 whereas, in chapter 4 each assay was used to build models trained on 54 different outcomes spanning Martingale residuals for incident hospital admission for 44 disease blocks and 10 health related risk factors (Full list in **Supplementary Table 20**).

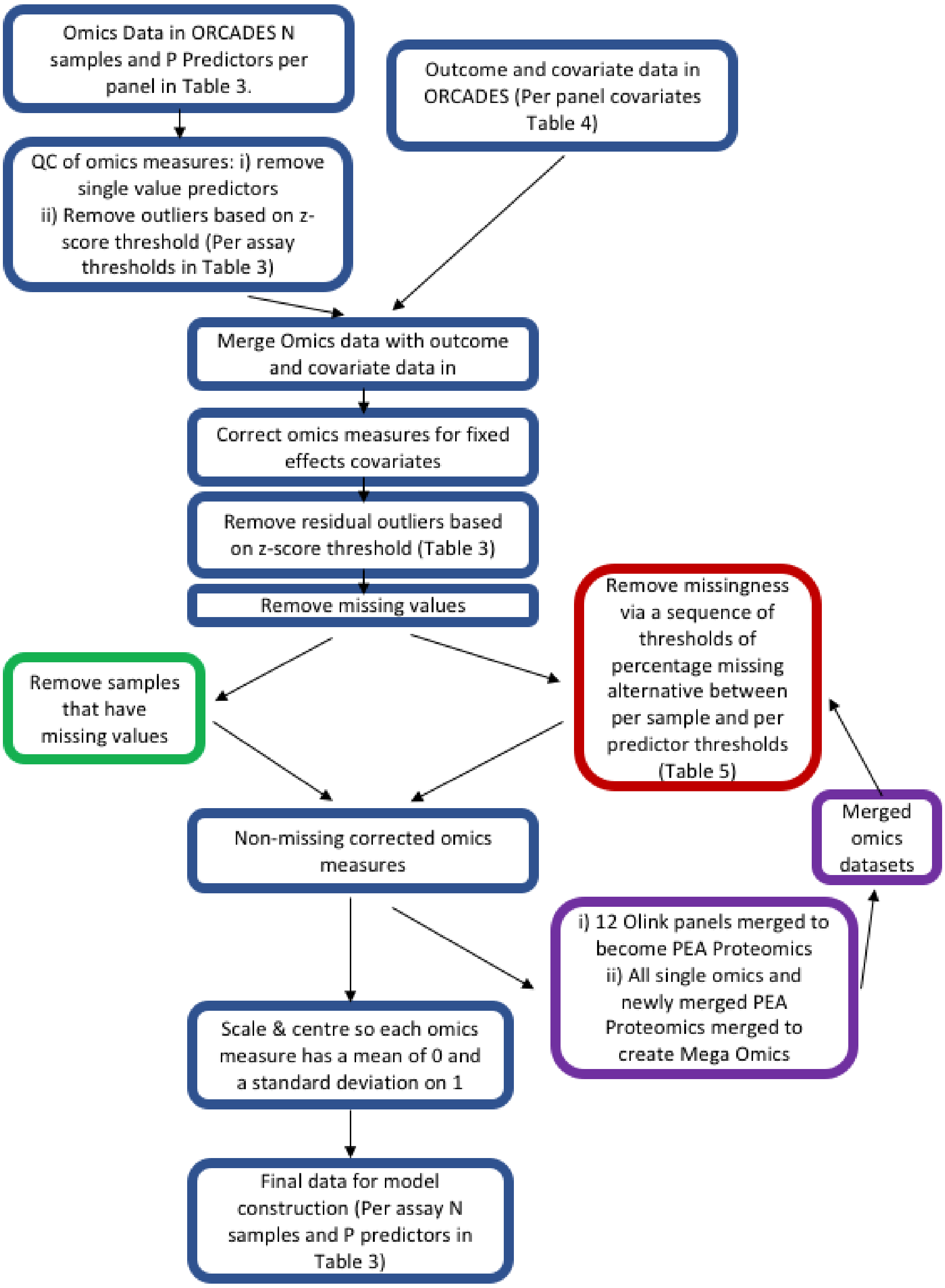

Omics Data in ORCADES N samples and P Predictors per panel in Table 3.
Outcome and covariate data in ORCADES (Per panel covariates Table 4)
QC of omics measures: i) remove single value predictors ii) Remove outliers based on z-score threshold (Per assay thresholds in Table 3)
Merge Omics data with outcome and covariate data in
Correct omics measures for fixed effects covariates
Remove residual outliers based on z-score threshold (Table 3)
Remove missing values
Remove samples that have missing values
Remove missingness via a sequence of thresholds of percentage missing alternative between per sample and per predictor thresholds (Table 5)
Merged omics datasets
Non-missing corrected omics measures
i) 12 Olink panels merged to become PEA Proteomics ii) All single omics and newly merged PEA Proteomics merged to create Mega Omics
Scale & centre so each omics measure has a mean of 0 and a standard deviation on 1
Final data for model construction (Per assay N samples and P predictors in Table 3)

Figure 5. Quality control pipeline for ORCADES Omics data. *Assays underwent the following steps as outlined, with assays with less missingness undergoing the NA removal procedure outlined in green and assays with more complex patterns of missingness the procedure in red. PEA Proteomics and Mega Omics assays were created by merging partly QC'd datasets as described and underwent the procedure highlighted in purple with NAs subsequently removed from the merged datasets according to the procedure outlined in red.*

QC for Biological Ageing Clocks

Omics measures were corrected for covariates using fixed effect linear regression (covariates fitted per assay in **Table 4**). For the assays indicated in **Table 3**, residual outliers were removed based on an assay level z-score thresholds that were chosen based on visualisation of the distributions. This was the procedure followed for the Hannum CpGs DNAme, Horvath CpG DNAme, NMR Metabolomics, Clinomics, DEXA, MS Fatty Acid Lipidomics, MS Metabolomics, MS Complex Lipids and UPLC IgG Glycomics datasets.

	N with Omics	N Predictors	Raw Z-score threshold	Residual Z-score threshold	N with Omics post QC	N Predictors post QC
DEXA	1,302	53	6	3	1,158	28
Clinomics	2,019	13	6	-	1,817	13
Hannum CpGs DNAme	1,052	62	6	-	1,035	62
Horvath CpGs DNAme	1,052	333	6	-	959	333
UPLC IgG Glycomics	2,030	77	6	-	1,937	77
NMR Metabolomics	2,015	225	5	3	1,645	86
MS Fatty Acid Lipidomics	1,000	44	6	4	954	33
MS Metabolomics	1,046	1,102	6	-	863	682
MS Complex Lipids	1,040	1,028	6	-	941	908
PEA Proteomics	1,057	1,102	6	-	805	886
Mega Omics		4,033	6	-	796	2,471

Table 3. QC Steps across Omics. *N with omics data: the number of samples in ORCADES with the relevant assay measured. N Predictors: number of predictors measured in each assay. Raw Z-score threshold: Z-score cut-off for raw omics measures. Residual Z-score threshold: Z-score threshold for omics residuals after correction for covariates. N with Omics post QC: the number of samples remaining after QC. N Predictors post QC: the number of predictors in the final QC'd dataset (the removal of missing samples and predictors to get post QC data is described later in this section).*

Omics Assay	Covariates
DEXA	Sex
Clinomics	Sex
Hannum CpGs DNAme	Sex
Horvath CpGs DNAme	Sex
UPLC IgG Glycomics	Sex
NMR Metabolomics	Sex, statin use
MS Fatty Acid Lipidomics	Sex, statin use, box number, box position
MS Metabolomics	Sex, statin use, day of assay, box number, row, col
MS Complex Lipids	Sex, statin use, day of assay, box number, row, col
PEA Proteomics	-
Mega Omics	-

Table 4. Covariates fitted per Omics Assay. *Sex and statin use were fit as binary variables. Box number: categorical variable. Date of assay: categorical variable. Row and column were fitted as ordered factors.*

Due to the algorithm used to construct penalised regressions models (discussed in detail in next section) requiring complete non-missing data, missing values (NAs) had to be removed. Given the relatively small sample size of ~1,000 I wanted to avoid losing too many samples and thus power. On the other hand, given that the sheer number of omics predictors measured is a distinguishing feature of ORCADES, I wanted to minimise the number of predictors removed from the analysis. I therefore removed missing values while attempting to maximise the number of both samples and predictors with complete non-missing data for my analysis.

For assays with relatively few missing values I removed samples based on a percentage missing threshold of 0, the procedure for removing NAs indicated in green in **Figure 5**. Hannum CpGs DNAme, Horvath CpGs DNAme and UPLC IgG Glycomics assays underwent this procedure. Assays with more complex patterns of missingness such as DEXA, Clinomics, NMR metabolomics, MS Fatty Acid Lipidomics, MS Metabolomics and MS Complex Lipids underwent the NA removal procedure outlined in red in **Figure 5**. This procedure involved removing missingness via a sequence of percentage missing thresholds, the thresholds alternating between being applied across samples and across predictors, details of the sequence of thresholds per assay are shown in **Table 5**. These thresholds were determined by manual examination of the visualisation of the proportion missing across either samples or predictors.

Sequence of Thresholds ↓

DEXA	**Clinomics**	**NMR Metabolomics**	**Fatty Acids Lipidomics**
0.3	0.3	0.08	0.3
0.01	0.05	0	0.05
0.01	0.04	-	0.04
-	-	-	0
MS Metabolomics	**MS Complex Lipidomics**	**PEA Proteomics**	**Mega Omics**
0.05	0.1	0.3	0.6
0.08	0.08	0.025	0.4
0.02	0.02	0.05	0.2
0.0075	0.007	0.004	0.1
0.015	0.01	0.005	0.05
0.007	0.004	0.003	0.02
0.01	0.007	0.003	0.02
0.005	0.003	0.002	0.005
0.0075	0.004	0	0.005
0.004	0.002	-	0.002
0.006	0.002	-	0
0.003	0.001	-	-
0.003	-	-	-
0.002	-	-	-
0.002	-	-	-
0.001	-	-	-

Table 5. Sequence of percentage missing thresholds applied to Omics Data. *The proportion of missing values threshold applied to samples (blue text) and predictors (red text) for the indicated omics assay, in the order that they were applied.*

As different Olink panels were assayed on different dates as panels became available over time, two different configurations of samples on plates were used. In order to minimise the effect of the position of the sample on the plate on protein level, I wanted to fit plate position as a covariate when processing the Olink data. The sample configuration differing across panels meant that each of the 12 panels of 92 proteins underwent the QC process separately. The raw omic Z-score threshold for all 12 panels was 6 standard deviations from the mean and there was no residual Z-score threshold applied. The same covariates were fitted for all panels namely: sex, season of venepuncture, time the sample had been in storage before assay (days), plate number, plate row and plate column. At the panel level, missing values were removed using the percentage missing threshold of 0 across samples approach (indicated in green in **Figure 5**). At this point the covariate corrected 12 panels were merged into the PEA Proteomics dataset which then underwent the sequential threshold procedure (approach in red in **Figure 5**) to end up with only those samples that had measures for the final set of proteins across all panels that make up the final non-missing dataset.

Similarly, to the PEA Proteomics dataset, the Mega Omics dataset was created by merging the 10 non-missing covariate corrected datasets, then removing missing values of the newly merged dataset using the sequential threshold procedure. As the creation of the PEA Proteomics and Mega Omics datasets merged the already QC'd assay data, the assay level covariates and Z-score threshold columns in **Table 3** and **Table 4** are not applicable.

Finally, the omics measures in the 11 non-missing covariate corrected datasets were scaled and centred to have a mean of 0 and a standard deviation of 1. These 11 QC'd datasets were then split into training and testing to build my standard and core models detailed in the methods section of chapter 3. It was also these QC'd datasets that were used to calculate principal components (PCs) of each omics assay in chapter 3.

QC for Omics Biomarker of Disease

The quality control procedure to create the 11 omics datasets used in the analyses in chapter 4 was identical, bar the inclusion of age at venepuncture as a covariate in addition to those listed in **Table 4**, for each assay. Age was not included as a covariate when creating the omics datasets for the ageing clocks analysis as I was interested in omics measures' relationship with age, whereas here the aim is to assess omics biomarkers association with risk factors and incident disease, independent of age.

The assay level z-score thresholds, missingness removal procedure and thresholds were the same as indicated in **Table 3** and **Table 5** respectively for all omics. The number of predictors across omics assays are shown in **Table 6**, the number of samples per assay across risk factors are shown in **Table 7** and across disease blocks in **Supplementary Table 16**.

Omics Assay	N Predictors
Clinomics	13
DEXA	29
Hannum CpGs DNAme	62
Horvath CpGs DNAme	333
UPLC IgG Glycomics	77
NMR Metabolomics	68
MS Fatty Acid Lipidomics	32
MS Metabolomics	682
MS Complex Lipids	908
PEA Proteomics	967
Mega Omics	2,534

Table 6. Number of Predictors available for each Omics Assay.

	FRS	BMI	Educational Attainment	HDL	Total Cholesterol	Cortisol	FEV1	Systolic BP	Creatinine	CRP
Clinomics	917	940	890	940	940	940	923	931	940	901
DEXA	984	1010	960	1033	1033	1032	991	1000	1033	989
Hannum CpGs DNAme	1081	1104	1088	1140	1140	1137	1083	1095	1140	1114
Horvath CpGs DNAme	909	934	886	957	957	956	917	924	957	919
UPLC IgG Glycomics	1766	1815	1694	-	-	1807	-	-	1815	1767
NMR Metabolomics	932	952	900	952	952	952	937	943	952	912
MS Fatty Acid Lipidomics	1806	1862	1778	1937	1937	1930	1766	1845	1937	1889
MS Metabolomics	842	861	814	861	861	861	846	854	861	827
MS Complex Lipids	1646	1692	1575	1693	1693	1688	1603	1677	1693	1647
PEA Proteomics	767	783	741	798	798	798	771	774	798	761
Mega Omics	776	793	749	-	-	793	-	-	793	759

Table 7. Number of Samples Across Omics Assays for Risk Factor Analysis. *FRS: Framingham risk score. BMI: body mass index. HDL: high density lipoprotein. FEV1: forced expiratory volume in 1 minute. Systolic BP: systolic blood pressure (mmHg). CRP: C-reactive protein.*

2.3.2 QC of Omics Data in Replication Cohorts

Quality control of the omics data in cohorts used for replication of models trained in ORCADES followed a similar procedure to that described above. Omics measures outliers were removed using the Z-score thresholds indicated in **Table 8**, single value predictors were removed, and assay-specific covariates were fitted as indicated in **Table 9**. Covariate corrected omics residuals outliers were removed based on Z-score thresholds (**Table 8**).

Cohort	Omics Assay	Raw Z-score threshold	Residual Z-score threshold	N with Omics post QC
UKBB	DEXA	6	6	3,740
UKBB	Clinomics	6	-	17,003
GS:SHFS	Hannum CpGs DNAme	6	-	5,048
EGCUT	Hannum CpGs DNAme	6	-	282
GS:SHFS	Horvath CpGs DNAme	6	-	4,950
EGCUT	Horvath CpGs DNAme	6	-	229
Korčula	UPLC IgG Glycomics	6	-	900
Vis	UPLC IgG Glycomics	6	-	382
Korčula	NMR Metabolomics	5	3	775
EGCUT	NMR Metabolomics	5	3	6,704
Vis	PEA Proteomics Subset 1	6	-	755
EGCUT	PEA Proteomics Subset 2	6	-	247

Table 8. Omics Data Description for Replication Cohorts. *Indicating the number of samples with omics data post-qc (N with omics post-qc), the z-score thresholds for raw omics measures (Raw z-score threshold) and covariate-corrected residuals (residual z-score threshold) for each replication cohort for each assay they contributed.*

Cohort	Omics Assay	Covariates
UKBB	DEXA	Sex, genomic ethnicity, batch, assessment centre, withdrawn
UKBB	Clinomics	Sex, genomic ethnicity, batch, assessment centre, withdrawn
GS:SHFS	Hannum CpGs DNAme	Sex
EGCUT	Hannum CpGs DNAme	Sex
GS:SHFS	Horvath CpGs DNAme	Sex
EGCUT	Horvath CpGs DNAme	Sex
Korčula	UPLC IgG Glycomics	Sex
Vis	UPLC IgG Glycomics	Sex
Korčula	NMR Metabolomics	Sex, statin use
EGCUT	NMR Metabolomics	Sex, statin use
Vis	PEA Proteomics Subset 1	Sex, site, plate number, plate row, plate column
EGCUT	PEA Proteomics Subset 2	Sex, season of venepuncture, time in storage (days), plate number, plate row, plate column

Table 9. Omics Covariates for Replication Cohorts. *The fixed effects covariates used to correct omics measures in each replication cohort for each assay they contributed. Withdrawn: binary variable indicating if the individual had withdrawn consent, only individuals that had not withdrawn consent by 28/07/2020 were included in the analysis. Genomic ethnicity: a binary variable indicating if the participant is genomically British, as all other cohorts used in the analysis contain only individuals of European ancestry.*

In contrast with the ORCADES QC pipeline, removal of missing values in replication cohorts was relatively simple, as only omics measures selected for model inclusion were required. Any samples that were missing any of the omics measures selected for model inclusion were removed. The remaining non-missing covariate corrected omics measures were scaled and centred to have a mean of 0 and a standard deviation of 1. These QC steps were performed by a pipeline that performed quality control, calculated predicted outcome and returned the Pearson correlation coefficient (95% confidence intervals and p-value) of predicted and observed outcome and the results (effect size estimate, standard errors, p-values and R^2) of a linear model fitting the predicted outcome on observed outcome. This meant that for the analysis of the Estonian Biobank (EGCUT) data, the pipeline was run by Nele Taba and these descriptive statistics were returned without any individual level data being disclosed. I performed the analysis in all of the other replication cohorts.

2.4 Penalised Regression

As mentioned in the introduction, penalised regression techniques are the most common methods used to construct biological ageing clocks. Penalised regression is used extensively in chapters 3 and 4 to build multiple omics ageing clocks and to predict incident disease using omics biomarkers respectively. Penalised regression techniques overcome certain limitations of ordinary least squares (OLS) regression. OLS is the most common form of linear regression used for prediction and modelled as follows:

$$y = \mathbf{X}\beta + \varepsilon$$

where y is an $n \times \mathbf{1}$ vector of the observed outcome for n samples, $\boldsymbol{X}$ is the $n \times p$ matrix of the values of p predictors for n samples, $\boldsymbol{\beta}$ is the $p \times \mathbf{1}$ vector of estimated effect sizes for p predictors and where ε is the unmodelled error. OLS estimates the effect sizes for the predictors that minimise the sum of the squared differences between observed values of the outcome (y) and values predicted by the model ($\hat{y}$). Under the assumption of normally distributed data, OLS is a method of maximum likelihood estimation as it estimates parameters that maximise the likelihood that the observed data is the most probable.

A limitation of OLS is that it performs poorly in situations where the number of predictors (p) approaches or surpasses the number of observations (n). This results in a loss of power as degrees of freedom (df) are lost with increasing p.

$$df = n - p - \mathbf{1}$$

Multicollinearity, multiple predictors being correlated with each other, also becomes more likely with increasing p. This is particularly an issue with omics assays, for example the NMR Metabolomics assay which contains a large number of cholesterol subfractions that are extremely intercorrelated. This redundancy between predictors creates inflation in the variance of the effect size estimates. Overfitting also becomes more likely with increasing p, particularly when $p > n$. With enough predictors it is possible to estimate effect sizes that perfectly predict the outcome in the training sample but will not be as predictive in other samples. In OLS having a large number of predictors also makes interpretation of results less clear, it is desirable to investigate relationships between fewer predictors and an outcome[193].

Previous strategies to overcome these limitations of OLS regression, by reducing the number of predictors fitted in the model include: best subsets regression[194], stepwise regression and selection based on p-value. However, these approaches also have limitations, firstly testing all possible subsets of predictors included in the model becomes computationally unfeasible as p increases. Secondly, stepwise procedures and selection of predictors based on p-values may be sensitive to small changes in the training data and therefore may be less effective in samples other than the training set.

The notion of penalised regression was first put forward in 1970 when Hoerl & Kennard described their method ridge regression[195]. Ridge, together with least absolute shrinkage and selection operator (LASSO)[193] put forward by Tibshirani in 1996 and elastic net[196] by Ziu & Hastie in 2005 comprise the suite of penalised regression techniques. These approaches work by introducing a penalty or constraint for including too many predictors, hence the term "penalised regression". These three techniques are also referred to as shrinkage methods as they reduce effect size estimates towards zero, utilising a shrinkage parameter λ. Rather than simply estimating parameters, θ, that maximise the log likelihood $\ell(\theta|x)$ and minimise the sum of squared residuals as in OLS regression, penalised approaches minimise the function:

$$M(\theta) = L(\theta|x) + \lambda P(\theta)$$

Where L is a loss function and is proportional to the residual sum of squares, M is the objective function whose value is to be minimised, P is the penalty function and λ

controls the trade-off between the two parts. The role of the penalty function is to penalise “unrealistic” effect size estimates, i.e. those that differ greatly from zero.

Ridge regression utilises the “L_2-norm penalty”, the sum of squared coefficients:

$$P(\beta) = \sum_{j=1}^{p} \beta_j^2$$

As the effect of the penalty is fine-tuned using λ, in cases where $\lambda = \mathbf{0}$, the estimated coefficients will be the same as OLS estimates, as the penalty will have no effect. As λ increases towards infinity, the coefficients will be shrunk towards zero. The L_2-norm penalty means that while ridge regression shrinks coefficients estimated towards zero, none actually become zero, thus all of the predictors presented will be present in the final model.

On the other hand, LASSO regression utilises the “L_1-norm penalty”, the sum of the absolute coefficients:

$$P(\beta) = \sum_{j=1}^{p} |\beta_j|$$

It often shrinks some coefficients to zero, therefore performing variable selection as well as shrinkage. Similarly, elastic net regression which utilises both the L_1-norm and L_2-norm penalties also performs both shrinkage and variable selection.

These methods also vary in the way they handle groups of highly correlated predictors. Where LASSO will only retain one of a group of highly correlated predictors with little regard for which one and ridge regression will retain all, elastic net encourages a grouping effect where sets of highly correlated predictors are retained or dropped from the model as a group[196].

This ability to produce sparse models makes LASSO and elastic net more appropriate for the construction of models that could be clinically useful. The ideal situation, both in terms of being more cost effective and minimally invasive to participants, would be to measure as few biomarkers as possible while achieving effective prediction of the desired outcome. For this reason, in chapters 3 and 4 I compare only LASSO and elastic net regression, not ridge regression. In chapter 3 I show that there was no difference between LASSO and elastic net regression for constructing multiple omics ageing clocks. Elastic net was chosen to be taken forward and used for all the results presented in chapter 3 due to precedent, more previously published ageing clocks having used elastic net[54,55,66]. In chapter 4 I present the results of predicting incident hospital admission for disease and health-related risk factors using both LASSO and elastic net and found however that LASSO was more effective. This finding is discussed in detail in chapter 4.

As penalised regression techniques are sensitive to the variability of the predictors, all omics measures were standardised (scaled to have a standard deviation of 1 and a mean on 0) prior to model fitting.

2.5 Genome Wide Association Studies

I performed GWAS on plasma protein levels from the cardiovascular II and cardiovascular III panels from Olink proteomics in the cohorts ORCADES and Croatia-Vis that contributed to the genome wide association meta-analysis (GWAMA) that is discussed in **Chapter 5: Genome-wide Association Meta-analysis of 184 Plasma Protein Levels**. GWAS were run on 183 proteins rather than 184 (2x92-proteins per panel) as in both cohorts the measures for the protein CCL22 were all NA, due to Olink swapping that protein out of their panel and replacing it with GP6 between the dates of initial assay and their returning below lower limit of detection values.

The decision to include individual protein measures that were below the limit of detection (LOD), the lowest quantity that can be distinguished from the background by the assay, was taken due to issues that arose in a parallel analysis of proteins on the inflammation 1 Olink panel by collaborators, because of high proportions of below LOD measures. Two strategies commonly used with <LOD data: imputing these values to a single value for example, LOD/2 or zero, or setting them to NA and considering these measures missing, both result in a truncated distribution which is not ideal for linear association analysis. However, there is information in these values, as despite the assay being unable to accurately determine the quantity, we know the true measure is between zero and LOD. Given that LOD values for assays tend to be conservative with clinical use in mind, rather than having a complete distribution of values for GWAS, the decision was taken to use all available information and include the below LOD values returned by Olink. Collaborators found that this approach overcame the issues with association results observed using a truncated distribution.

The values for protein measures returned from Olink Proteomics are in relative quantification units, referred to as Normalised Protein eXpression (NPX) units. These values are based on C_t (also known as C_q) values that indicate the number of cycles of amplification that are required, during the qPCR process, for the fluorescent intensity from the protein in the sample to be distinguishable from background levels. C_t values are therefore inverse to the amount of protein in the sample. In order to minimise the chance that technical differences rather than genuine biological differences in protein levels cause the values returned, normalisation is performed to minimise both intra- and inter-assay variation. Final NPX values are inverted compared to C_t values so that higher NPX value indicates a higher protein concentration. NPX values are on a $\log_2$ scale so that a 1 NPX unit change equates to a doubling of protein concentration. NPX values allow relative quantification as they indicate changes for individual protein levels across their sample set. The preceding normalisation was carried out in house by Olink and I received protein measures in NPX units and carried out subsequent quality control.

For the analysis in both ORCADES and Croatia-Vis, these raw protein NPX values, including those measures that were below the lower limit of detection, were inverse normal-rank transformed. A fixed effects linear model was then fitted to correct for the following covariates: age at venepuncture, sex, season of venepuncture

(ORCADES only), array (ORCADES only), plate number, plate row, plate column, time in storage days (ORCADES only) and the first 10 principal components of the genotypes.

Array was fitted as a covariate in the ORCADES analysis only, as multiple different arrays were used to genotype this cohort as outlined previously, this was not the case for Croatia-Vis. Similarly, season of venepuncture was not fitted in the VIS analysis as all of the blood samples were drawn over a two-month period (March-April 2013) and therefore does not vary in this sample.

Those fixed effects residuals along with a genomic relationship matrix were used to calculate GRAMMAR+ residuals using the "polygenic" function using the R package GenABEL[197]. The Kinship corrected residuals that were within 4 standard deviations of the mean were then inverse normal-rank normalised and used as the dependent variable for SNP associations. Residuals out with the z-score threshold were removed as outliers.

	ORCADES		VIS	
Phenotype QC	CVD2	CVD3	CVD2	CVD3
Total Individuals with genotype information	2027	2027	958	958
Individuals with Olink data	1057	1057	903	915
Individuals with Olink & covariate data	972	994	896	908

Table 10. Summary of Phenotype Sample Size. *Indicating the number of individuals with HRC-imputed dosages, with Olink measures and the final sample size of individuals with both Olink and covariate data for ORCADES and Croatia-Vis for the CVD2 and CVD3 panels.*

A summary of phenotype sample size is shown in **Table 10**, REGSCAN[198] v5 was used to perform SNP-phenotype associations with HRC imputed allele dosages, using an additive model.

2.6 Meta-Analysis

In order to increase power to detect variants with small effect sizes that are associated with plasma protein levels, I increased the sample size by meta-analysing. I combined the GWAS results from 16 cohorts from the SCALLOP Consortium using genome-wide association meta-analysis (GWAMA), thereby increasing my sample size from the ~2,000 individuals in ORCADES and Croatia-Vis that I had individual level data for.

Summary statistics from 18 cohorts were collected centrally on a secure server. The number of cohorts providing summary statistics for each protein are detailed in **Supplementary Table 37** and the imputation reference panels used by each cohort are indicated in **Supplementary Table 25**. All summary statistics used build b37. Harmonised cohort files prior to meta-analysis underwent the following quality control steps. Monomorphic SNPs and variants with missing allele frequency, effect size, standard error or p-values were removed from cohort level files. Similarly, any variants that had nonsensical information such as: standard errors of zero, infinite effect sizes, allele frequencies >1 or <0 or either effect or other alleles which contained any characters other than 'A', 'C', 'T', 'G', 'I' or 'D' (Insertions or deletions) were also removed. P-values provided, and those two-sided p-values calculated from z-scores (effect size/standard error) using the "pnorm" function in R, were compared and found not to deviate, suggesting there were not systematic errors in the cohort level p-values.

METAL software was used to perform the meta-analysis with the QC'd cohort level data. I used the inverse-variance-weighted meta-analysis (STDERR scheme) with the additional filter to include only variants with an imputation quality score >0.4. Wanting to utilise the diversity of the contributing cohorts and their imputation strategies, I did not use a minor allele frequency filter at this stage, however only included variants that were assessed in three or more cohorts. Minor allele filters were subsequently used in downstream analysis and are outlined in the appropriate sections.

Separate significance thresholds, pre-correction for multiple testing, were used for *cis* (1×10^{-5}) and *trans*-variants (5×10^{-8}). Rather than correcting the significance threshold for 184 traits, as the protein levels are correlated, I calculated the number of PCs required to explain 95% of the variance in the 184 protein levels and took this value as the number of independent traits tested, as done previously by Kettunen *et al.*[122]. I found that 85 PCs explained 95% of the variance in the levels of the 184 proteins in ORCADES (using the "prcomp" function in R), I repeated the analysis in CROATIA-Vis and again found that 85 PCs explained 95% of the variance. The thresholds for significance were therefore 1.18 x 10^{-7} (Bonferroni 1 x 10^{-5}/85) for *cis*- and 5.9 x 10^{-10} for *trans*-association variants.

I also used the "ANALYZE HETEROGENEITY" option when running METAL to assess whether the test statistics were consistent across samples. To minimise the effect that heterogeneity between cohorts had on my results, I used additional criteria for variants to be designated genome-wide significant. Only variants that had an I^2<30% (where I^2 describes the percentage of variation across studies that is due to heterogeneity between studies rather than chance[199,200]) or have both: i) effect direction consistent with the meta in at least 3 individual cohorts and ii) be nominally significant (p<0.05) in at least 3 individual cohorts, were eligible to be considered genome-wide significant.

Chapter 3: Biological Ageing Clocks

3.1 Introduction

3.1.1 Context

As outlined in the introduction, the idea of a measure that captures the underlying rate at which an individual ages and that is indicative of future health, beyond chronological age (chronAge), biological age (BA), is extremely attractive. Such a measure offers the potential to not only greater our understanding of the ageing process, but potentially allow interventions that may slow or even reverse ageing.

While second generation clocks such as DNAm PhenoAge and GrimAge have been shown to outperform previous ageing clocks by more accurately predicting mortality and health outcomes[96,97]. There has been insufficient work done to characterise the properties of ageing clocks trained on chronAge, given that there have been so many published and some have indeed been shown to be prognostic of future health outcomes beyond chronAge[93–95].

Additionally, the issue with clocks trained on chronAge, that it is possible to create a perfect chronAge predictor with certain omics assays[91,92], has been raised in previous studies however has been under explored. It is essential that this issue is addressed, as a measure that predicts chronAge with 100% accuracy will, by definition, be unable to distinguish health outlook between individuals of the same chronAge, thus defeating the purpose of trying to capture biological age.

Ageing clocks trained on chronAge have been built using a variety of types of biomarkers: epigenetics[55,66], proteomics[54,88,89], metabolomics[86], glycomics[87], neuro-imaging data[82–85], immune cell counts[90], facial morphology[81] and telomere length[80]. However, comparison between different omics clocks in a single sample have tended to be limited to a few types of omics, often epigenetic or those based on clinical risk factors or frailty[93–95,98,99].

I sought to utilise the extremely densely phenotyped ORCADES cohort to build the widest spread of different omics ageing clocks to date in a single sample and to compare these different assays abilities as biomarkers of age. I also aim to explore the potential of a single clock built using biomarkers from multiple different assays. I further sought to characterise the properties of these multiple omics ageing clocks trained on chronAge and attempt to determine if they are capturing something biological or are merely artefacts of the statistical method used.

I also assess the extent of redundancy between omics biomarkers both within and between assays to simultaneously address two issues. First, the issue of too many predictors in a model being a perfect chronAge predictor and second, that for a model to be practical and suitable for use in the clinic, it should ideally contain as few biomarkers as possible.

The implicit principal purpose of an ageing clock is to be indicative of the future health of individuals as they age. I therefore investigate if the age acceleration measures from these omics ageing clocks are associated with current health related risk factors, as these themselves are indicative of future health status. I also investigate whether they are prognostic of subsequent incident disease beyond chronAge.

3.1.2 Contributions

The idea for this project was conceived by Peter Joshi. Jim Wilson gathered all the study data and arranged and financed the lab assays. Lucija Klarić performed the pre-processing described in the Methods section for the UPLC IgG Glycomics data in ORCADES, Croatia-Korčula and Croatia Vis. Azra Frkatović and Rosie Walker performed the pre-processing of the DNA methylation data as described in the methods section, in ORCADES and GS:SHFS respectively. The remaining omics data were QC'd by me.

Development of the omics clocks, given the QC'd omics inputs was carried out entirely by me. I wrote a pipeline that will for any input omics dataset perform quality control (as described in **2.3 Quality Control of Omics Data**), specifically: the removal of outliers by a user-specified z-score threshold; correction for user specified covariates and removal of missing values according to a user configured bespoke script, in order to maximise both the number of test samples and number of predictors available. This pipeline splits the input dataset into training and testing according to a user-defined ratio. It also has the functionality to build penalised regression models incorporating the "glmnet" R package[201]. It gives the user the option to perform either: LASSO, ridge, elastic net with fixed (user-specified alpha) or elastic net with an alpha calculated via 10-fold cross validation in the training sample. This pipeline also creates principal components (PCs) of the input omics data, to allow for the model construction step to be run on PCs of the input omics.

I packaged a subset of steps from my pipeline that would perform basic quality control and predict chronological age in validation datasets, based on effect sizes derived in ORCADES. Using this pipeline Nele Taba replicated a PEA Proteomics, NMR Metabolomics and DNA methylation clocks in the Estonian Biobank. I performed the validation in GS:SFHS, Croatia-Vis, Croatia-Korčula and the UK Biobank.

I also performed the: construction of "Core" omics using a subset of biomarkers from each assay, the correlation analysis between omics clocks and analysis of the overlap between clocks in the information they provide about chronological age. I also performed the association of omics clock age acceleration and health-related risk factors analysis.

Peter Joshi extracted, pre-processed and performed quality control on the SMR01 hospital admission data for ORCADES. I provided him with omics age accelerations, and he tested their association with the subsequent incident hospital admissions for disease. I organised the test results and integrated them into the manuscript.

Subsequent to preparation of a complete first draft by me (other than his own description of the association method carried out by Peter Joshi), Peter Joshi and Jim Wilson contributed to the redrafting of the manuscript and all co-authors commented on the manuscript prior to submission.

The following manuscript has been placed on bioRxiv doi: https://doi.org/10.1101/2021.02.01.429117 and at the time of writing is being revised based on reviewers comments for resubmission at the peer-reviewed journal *Aging*.

3.2 Manuscript Submitted

A catalogue of omics biological ageing clocks reveals substantial commonality and associations with disease risk

Erin Macdonald-Dunlop[1], Nele Taba[2,3], Lucija Klarić[4], Azra Frkatović[5], Rosie Walker[6], Caroline Hayward[4], Tõnu Esko[2,7], Chris Haley[4], Krista Fischer[2,8], James F Wilson[1,4]*, Peter K Joshi[1]*

1: Centre for Global Health Research, Usher Institute, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, UK
2: Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Riia 23b, 51010, Estonia
3: Institute of Molecular and Cell Biology, University of Tartu, Tartu, Riia 23, 51010, Estonia
4: MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK
5: Genos Glycoscience Research Laboratory, Borongajska cesta 83H, 10000, Zagreb, Croatia
6: Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK
7: Program in Medical and Population Genetics, Broad Institute, 415 Main St, Cambridge, MA 02142, United States
8: Institute of Mathematics and Statistics, University of Tartu, Tartu, Narva mnt 18, 51009, Estonia

Abstract

Biological age (BA), a measure of functional capacity and prognostic of health outcomes that discriminates between individuals of the same chronological age (chronAge), has been estimated using a variety of biomarkers. Previous comparative studies have mainly used epigenetic models (clocks), we use ~1000 participants to create eleven omics ageing clocks, with correlations of 0.45-0.97 with chronAge, even with substantial sub-setting of biomarkers. These clocks track common aspects of ageing with 94% of the variance in chronAge being shared among clocks. The difference between BA and chronAge - omics clock age acceleration (OCAA) - often associates with health measures. One year's OCAA typically has the same effect on risk factors/10-year disease incidence as 0.46/0.45 years of chronAge. Epigenetic and IgG glycomics clocks appeared to track generalised ageing while others capture specific risks. We conclude BA is measurable and prognostic and that future work should prioritise health outcomes over chronAge.

Introduction

Age is a phenotype that we are all familiar with, and is a major risk factor for numerous diseases including the largest causes of mortality[49]. We all become acquainted with visible changes that accompany ageing, such as greying hair, baldness, loss of skin elasticity and worsening of posture, and that these vary noticeably amongst individuals of the same chronological age (chronAge). However, there are also molecular hallmarks of ageing such as telomere shortening, genomic instability and cellular senescence that also show variation in individuals of the same chronAge[49]. It has previously been hypothesised that an underlying biological age (BA), likely tagged by these molecular hallmarks, is what gives rise to age-related disease risk[96]. Measuring BA therefore has the potential to be more prognostic of health and functional capacity than chronAge and, as importantly, BA may be reversible[202], unlike chronAge[56].

Since this concept was proposed, there has been a push to construct models of BA, using a variety of both statistical methods and types of biomarkers; the resultant estimates we shall term omics clock ages (OCAs). The first OCAs were epigenetic clocks that used methylation levels of CpG sites across the genome - DNA methylation (DNAme) - to estimate chronAge using penalised regression[55,66]. The excess of OCA over chronAge being omics clock age acceleration (OCAA), hopefully measuring an underlying biological effect. DNAme's verification as a meaningful BA measure, rather than a mere statistical artefact, was confirmed when DNAme OCAA as calculated by Horvath's clock was shown to be associated with all-cause mortality[203]. Ageing clocks trained on chronAge have also been constructed using DNA methylation[55,65,66], telomere length[80], facial morphology[81], neuro-imaging data[82–85], metabolomics[86], glycomics[87], proteomics[54,88,89] and immune cell counts[90]. There has however, been limited comparison of the performance, for example accuracy and correlation, of different omics ageing clocks, particularly in the same set of individuals.

Moreover, there has been inadequate additional progress in demonstrating that the various OCA measures are actually tracking underlying BA beyond chronAge, and whether some clocks' OCAAs are more aligned to certain outcomes than others. For example, few significant associations of chronAge-trained OCAAs have been found with health outcomes other than mortality and those that do have low effect sizes[94,204–206].

The deep omic and health outcome annotation of the Scottish population-based Orkney Complex Disease Study[186] cohort (ORCADES) permits interrogation of the utility and limitations of BA clocks. Here, we compare the performance of 11 ageing clocks built from 9 different omics assays in the same set of approximately 1000 individuals in ORCADES, including whole body imaging and a clock based on the grand union of all the omics. Next, we assess the biological meaningfulness of the derived OCAA measures, by assessing their association with health-related phenotypes and incident hospital admissions (post-assessment) over up to 10 years follow-up.

The notion of BA raises fundamental questions. Is there one BA for a person, or a set of BAs, perhaps relating to different bodily systems[89,207]. Are measured (chronAge trained) OCAs tracking a single BA, with differences arising due to their focus and accuracy, or are they tracking different underlying BAs? This study aims to shed some light on these issues.

Results

Performance of Omics Clocks

We constructed eleven ageing clocks, training on chronAge, in the ORCADES cohort from assays already understood to be able to form effective ageing clocks[54,55,66,87], covering plasma Immunoglobulin G (IgG) glycans, proteins, metabolites, lipids, DNA methylation and a collection of commonly used clinical measures (such as weight, blood pressure, fasting glucose, etc), which we label Clinomics. To this we added two novel omics sets for clock construction: a DEXA whole body imaging set of body composition measures, and one based on all the omics assays considered simultaneously, which we term Mega-omics, as listed in **Table 11** (see Methods for assay descriptions). Rather than creating completely novel DNAme clocks when effective and extensively studied published clocks exist, our methylation clocks' potential predictor sets are the subsets of the CpG sites used in Hannum and Horvath's epigenetic clocks available on the Illumina EPIC 850k methylation array. With this caveat, all clocks were derived from scratch using the set of available predictors and elastic net regression.

We first assessed various forms of penalised regression: LASSO, elastic net with a fixed alpha of 0.5 and elastic net with alpha calculated via cross-validation, training clocks in 75% of the ORCADES cohort and evaluating in the remaining 25% (the testing sample). We found that clock performance in estimating chronAge was independent of penalised regression method used, across all the assays

(**Supplementary Figure 26**) and so elastic net regression with a fixed alpha of 0.5 only was employed in subsequent analyses.

Ages estimated by the model in the test set (i.e. OCAs) were highly correlated with chronAge for the majority of the omics clocks tested (**Table 11**), particularly PEA proteomics (r=0.93) and DNAme based (r=0.96 Hannum CpGs, r=0.93 Horvath CpGs) clocks (correlations in the training set in **Supplementary Figure 27**). Unsurprisingly, the mega-omics OCA had the highest correlation (r=0.97). Although all features were given equal opportunity to contribute to the mega-omics clock, those selected by the regression were predominantly DNAme- and PEA proteomics-based (34.6% CpGs, 31.8% PEA Proteomics, 20.6% MS metabolites, 13.1% other). We found that the MS Fatty Acids Lipidomics OCA had the lowest correlation with chronAge (r=0.45; **Figure 6**). The number of biomarkers available and then selected for model inclusion for each omics clock are indicated in **Table 11** (Full list of biomarkers measured in each assay in **Supplementary Table 17** and coefficients for all clocks in **Supplementary Table 18**).

Omic	N Individuals	N Predictors Available	N Predictors Selected	r
MS Fatty Acids Lipidomics	952	33	27	0.45
DEXA	1158	28	28	0.66
MS Complex Lipidomics	940	908	130	0.7
NMR Metabolomics	1643	86	81	0.74
UPLC IgG Glycomics	1937	77	50	0.74
Clinomics	1815	13	12	0.8
MS Metabolomics	861	682	181	0.81
DNAme Horvath CpGs	957	333	155	0.93
PEA Proteomics	805	886	203	0.93
DNAme Hannum CpGs	1033	62	50	0.96
Mega Omics	796	2471	214	0.97

Table 11. Multiple omics make accurate ageing clocks. *Indicating for each omics assay: N Individuals: the number of individuals in the ORCADES cohort that passes quality control, N Predictors Available: the number of predictors passing assay-level quality control and therefore available for selection for inclusion in the standard model, N Predictors Selected: the number of predictors selected for inclusion in the standard model, r: Pearson correlation of omics clock age (OCA) and chronAge. DEXA: Dual X-ray absorptiometry, DNAme: DNA methylation, CpG: cytosine nucleotide followed by guanine (5' to 3' direction), MS: mass spectrometry, NMR: nuclear magnetic resonance, PEA: proximity extension assay, UPLC: ultra-performance liquid chromatography, IgG: Immunoglobulin G. Within each omics assay, subject mean age at baseline*

*was 53-56 (SD~15) with an age range across clocks of 16-100, whilst the proportion female ranged from 55-61% (**Supplementary Table 19**).*

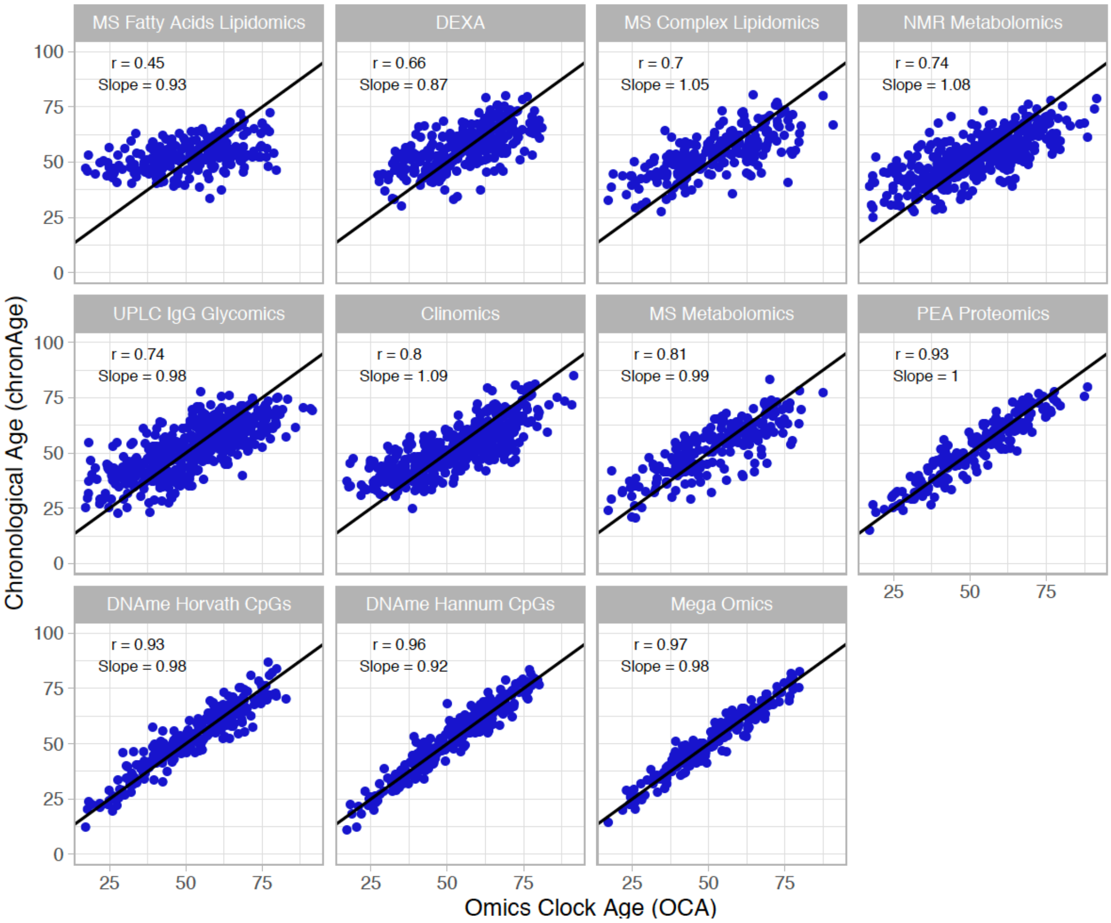


Figure 6. Multiple omics estimate chronological age, to varying degrees of accuracy, in a broadly unbiased manner. *The correlations of chronAge on the y-axis with ages estimated by the omics ageing clock (OCA) on the x-axis, in the ORCADES testing sample. Pearson correlation coefficient (r) and the slope of the regression of OCA on chronAge are indicated in each panel. Identity line indicated in black.*

Validation of Clock Performance in Independent Cohorts

We next used the clocks trained in ORCADES to estimate age in independent European cohorts to validate if they were more widely applicable beyond the Orkney population. We found that correlations between OCA and chronAge replicated to varying degrees in independent populations (**Supplementary Figure 28**). PEA proteomics and DNAme based clocks produced correlations of OCA and chronAge in the range of 0.89-0.98 in European cohorts replicating the range of 0.91-0.96 in ORCADES. UPLC IgG glycomics and Clinomics OCAs in independent populations showed a range of OCA-chronAge correlations of 0.56-0.62 compared to the 0.74-0.80 in ORCADES. Whilst the NMR metabolomics and DEXA did not replicate, with correlations of 0.26-0.55 in validation cohorts compared with 0.66-0.73 in ORCADES.

Accurate Performance of Clocks with Substantial Core Subset of Biomarkers

If the aim is to create BA clocks that have the potential to be clinically useful, it would be more efficient and cost effective to reduce the numbers of biomarkers that need to be measured in patients. To this end, we investigated the performance of our clocks using a reduced set of biomarkers. For each of our 11 omics clocks a "core" clock was constructed using only those biomarkers which were selected for model inclusion in >95% of 500 iterations of our clock construction procedure, as done by Enroth *et al.*[54] (See Methods for details). Comparable correlations of OCA and chronAge were achieved across all 11 clocks with a substantial subset or core of biomarkers (**Figure 7**), highlighting the potential for accurate OCAs with a small number of predictors (e.g. 30s-60s of biomarkers).

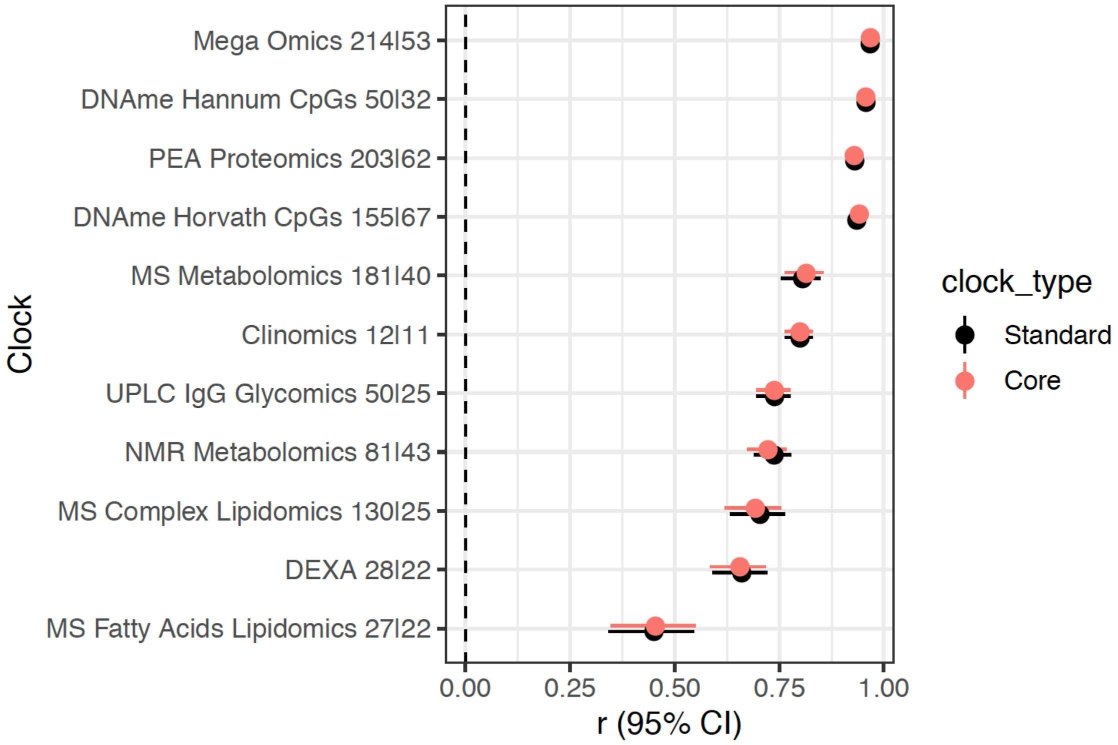


Figure 7. Substantial subsetting of biomarkers results in little dilution of accuracy. *Pearson's correlation (r) and 95% confidence interval of chronAge and OCAs from standard and core models for each omics assay indicated on the y-axis in the ORCADES testing sample. The number of predictors selected for inclusion in the standard and then core models are indicated in the y-axis labels (standard|core).*

Comparison of Biological Age Between Clocks

Omics Clock Age Accelerations (OCAAs) showed varying degrees of positive correlation between clocks (**Figure 8**). Unsurprisingly, the two DNAme based OCAAs were the most correlated with each other (r=0.73) and, in hierarchical clustering, formed a group on their own. The four clocks that are primarily constructed from lipid species and fractions, MS Fatty Acids Lipidomics, MS Complex lipidomics, NMR Metabolomics and MS Metabolomics clocks, all clustered together. The DEXA, Clinomics and UPLC IgG glycomics clocks formed a related group. Interestingly, the PEA Proteomics OCAA clustered between the DNAme and glycomics-DEXA-Clinomics-lipidomics cluster, on its own.

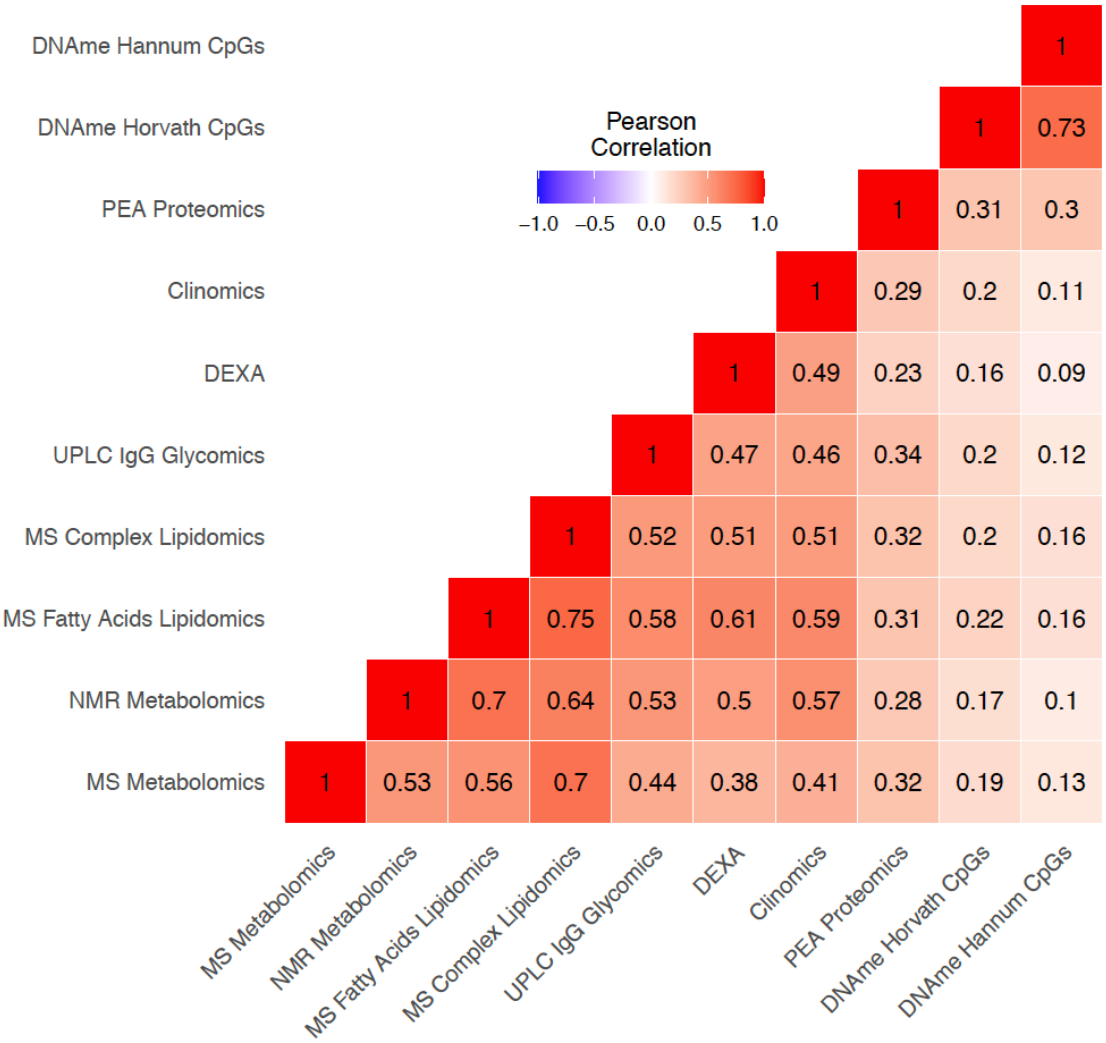


Figure 8. Variable positive correlations between different omics age accelerations. *Pearson correlation of OCAAs (omics clock age–chronAge) in ORCADES testing and training samples. Colour indicates the direction and the shade and number indicate the magnitude of the correlation. Rows and columns are ordered based on hierarchical clustering of the pairwise correlations.*

Proportions of Variance in Age Explained by Different Clocks

To determine if our different clocks are explaining the same or different variance in chronAge, we partitioned the variance in chronAge explained among our clocks. We calculated the unique variance in chronAge explained by each OCA as the squared part correlations of chronAge and OCA, while controlling for all other clocks. 93.9% of the variance in chronAge is explained by two or more clocks, whilst 4.1% remains unexplained by the 10 ageing clocks tested, with the remaining 1.9% being explained by one clock uniquely (**Supplementary Figure 29a**). The PEA proteomics and Hannum

CpG clocks explain the most variance in chronAge uncaptured by any other clock (0.59% and 0.46% respectively; **Supplementary Figure 29b**). Pairwise clock comparisons are shown in **Supplementary Figure 30**.

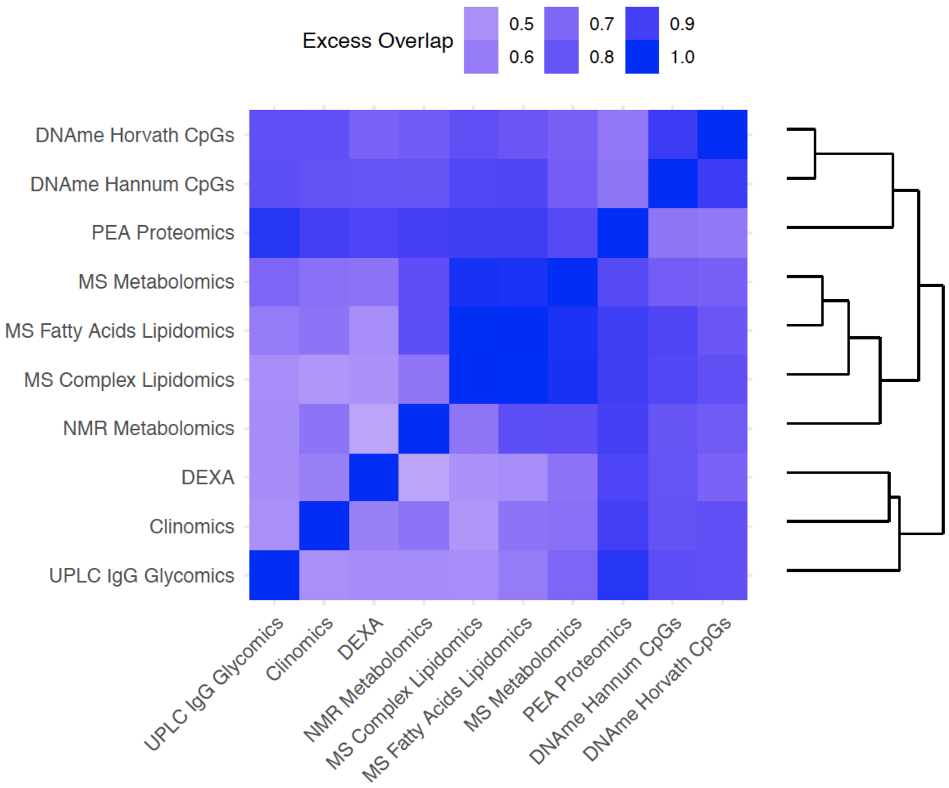


Figure 9. Bivariate analyses reveal that clock pairs tend to overlap more than expected by chance in the variance in ChronAge they explain. *The amount of excess overlap that would be expected by chance is indicated for each pair of clocks. This is the deviation of the observed variance in chronAge explained by a bivariate model containing a pair of OCAs and the variance expected to be explained by that pair, given that we know how much variance in chronAge they explain individually, if each of the clocks were independent samples from a set of latent complete predictors. This measure of deviation of observed from expected is scaled (See Methods for details) so that a value of 1 means that the second clock is adding no more information than the first, meaning that they overlap entirely in the information they provide about chronAge. A value of 0 would indicate the observed variance explained in chronAge is exactly what is expected if the two clocks were independently sampling. Negative values are possible on this scale but are not observed and would indicate disproportionately complementary components of chronAge were being tracked.*

Having found that clocks overlap in the information they provide about chronAge, we tested to see if, together, pairs of clocks jointly explained a different proportion of variance in chronAge than would be expected if the clocks were each independently sampling from a latent set of complete predictors of chronAge (ISLSP). This analysis should reveal whether the clocks were tracking complementary dimensions of ageing:

situations where the pair of clocks overlapped less than expected if they were independently sampling (negative values on this scale). Strikingly, excess overlap was found across all pairs of clocks (**Figure 9**), with the lowest excess overlap value measured at 0.41 (comparison of the NMR Metabolomics and DEXA clocks): all 10 omics clocks, considered pairwise, track more common rather than complementary aspects of chronAge.

The most overlapping were the MS Fatty Acids Lipidomics and the MS Complex Lipidomics clocks (excess overlap of 0.98; note on our scale, a clock shows 1.00 excess overlap with itself, whilst ISLSP would show 0.00). These two clocks formed a cluster with the clocks derived from NMR and MS Metabolomics (which both contain many lipid features). Similarly, the two DNAme-based clocks clustered tightly together with an excess overlap of 0.91. As these clocks are extremely accurate, a large amount of overlap in variance explained is inevitable; they are tracking common aspects of ageing.

OCAAs compared to chronAge as predictors of disease risk

We next sought to test the effect of OCAAs compared to chronAge on risk factors and post assessment disease incidence, as measured by hospitalisation in the ORCADES cohort, where the outcome was thought *a priori* to associate with age. For risk factors we chose body mass index (BMI), systolic blood pressure (SBP), cortisol, creatinine, C-reactive protein (CRP), forced expiratory volume in 1 second (FEV1), Framingham Risk Score, and total cholesterol. For diseases we chose five International Statistical Classification of Diseases and Related Health Problems (ICD)-10 Chapters: II (Neoplasms - codes C), IV (Endocrine, nutritional and metabolic diseases - codes E), IX (Diseases of the circulatory system - codes I), and X (Diseases of the respiratory system - codes J). The ICD-10 blocks used and their codings are listed in **Supplementary Table 20**.

In order to compare OCAA and chronAge, we first quantified the effect of chronAge on disease and risk factors (**Supplementary Figure 31** & **Supplementary Figure 32**).

All 8 risk factors and 32/44 disease blocks were taken forward as they were significantly associated with chronAge (beta>0, FDR<10%) and had >5 incident cases (disease blocks). The effect of chronAge on (standardised) risk factors appeared to vary by trait, whereas for diseases, it appeared that the effect of chronAge (on the hazard ratio scale) might be similar across diseases, with a consistent doubling of risk every 14 years.

We tested for risk factor and disease associations with OCAA, using chronAge and sex as covariates. Results were then rescaled to be per year of chronAge effect, by dividing the observed effect of OCAA by the effect of chronAge on the outcomes, as identified at the previous step. This was taken trait-by-trait for risk factors, and a single effect for all disease groups and chapters: -0.0492 $\log_e$ HR.

Despite limited power for detecting OCAA-disease associations, 6/352 tests were statistically significant (FDR<10%) as were 31/88 OCAA-risk factor associations. We also found evidence of enrichment of positive effects of OCAA on both risk factors (85%) and disease (74%), with 35% and 23% being nominally significant (one sided p<0.05), respectively. Across clocks, the inverse variance-weighted mean effect of one year of OCAA on risk factors/disease was the same as 0.45/0.46 years of chronAge (SE~0.01, note here and elsewhere ~ denotes indicative, see Methods for details). For risk factors, as might be expected, this was strongly influenced by an average effect of 1.23 years for Clinomics OCAA (0.16 without Clinomics). Interestingly, the mean effect across all diseases of one year's DNAme Hannum/Horvath CpGs OCAA was similar to one year of chronAge (ratio: 1.03/0.85, SEs ~0.18), but the effect on risk factors was much lower (ratio: -0.03/-0.01, SEs ~ 0.06). Complete results are shown in **Supplementary Table 21** and inverse variance-weighted effects are shown in **Supplementary Figure 34**.

In general, only associations with the Clinomics OCAA passed FDR, however both DNAme OCAAs and the UPLC IgG Glycomics OCAA were nominally associated with eleven ICD10 blocks, one more than Clinomics (**Figure 10**). In contrast, the PEA proteomics clock (r=0.93 with chronAge) showed only one nominally significant disease-OCAA association. Looking at disease groupings, E70-E90 Metabolic disorders and J09-J18 Influenza and Pneumonia showed the most nominal associations across all OCAAs. Curiously, on the other hand, C34-C44 Melanoma and C51-59 Malignant Neoplasms of the female genital organs, showed generally negative associations with OCAAs.

The greater statistical power for risk factors results in considerably more significant associations at FDR<10% (**Figure 11**). Once more, Clinomics, as might be expected, has the greatest number of significant associations, however NMR metabolomics and UPLC IgG Glycomics OCAAs are nearly as broadly predictive. Mega-omics, MS and NMR Metabolomics OCAAs show positive associations with all risk factors. It should be noted that while the Clinomics OCAA showed most significant FDR<10% associations with diseases and risk factors, its predictors (e.g. cholesterol, FEV1 and SBP) are often close to and designed to predict clinical endpoints and overlap with the risk factors considered here. Similarly, metabolite and lipid-based clocks contain cholesterol subfractions. All OCAAs were associated positively with BMI and total cholesterol. We found strong associations between OCAAs and the marker of inflammation CRP (often with effect sizes >1), meaning OCAA had a larger effect than chronAge. Overall, the averaged effect of OCAA on risk factors as a proportion of the effect on diseases was large for MS Fatty Acid Lipidomics/Clinomics/PEA proteomics (69%/230%/291%) suggesting they are directly tracking the risk factors we considered. Conversely, this proportion was small for Hannum CpGs/Horvath CpGs/UPLC IgG Glycomics (-3%/-1%/29%), suggesting they are prognostic of incident disease and therefore track more generalised ageing (**Supplementary Figure 34**).

We wanted to check if observed OCAA-health associations were driven by the associations of health with smoking and of OCAA with smoking. Our analysis fitting smoking status as a confounder suggests they were not (**Supplementary Figure 35a & b**).

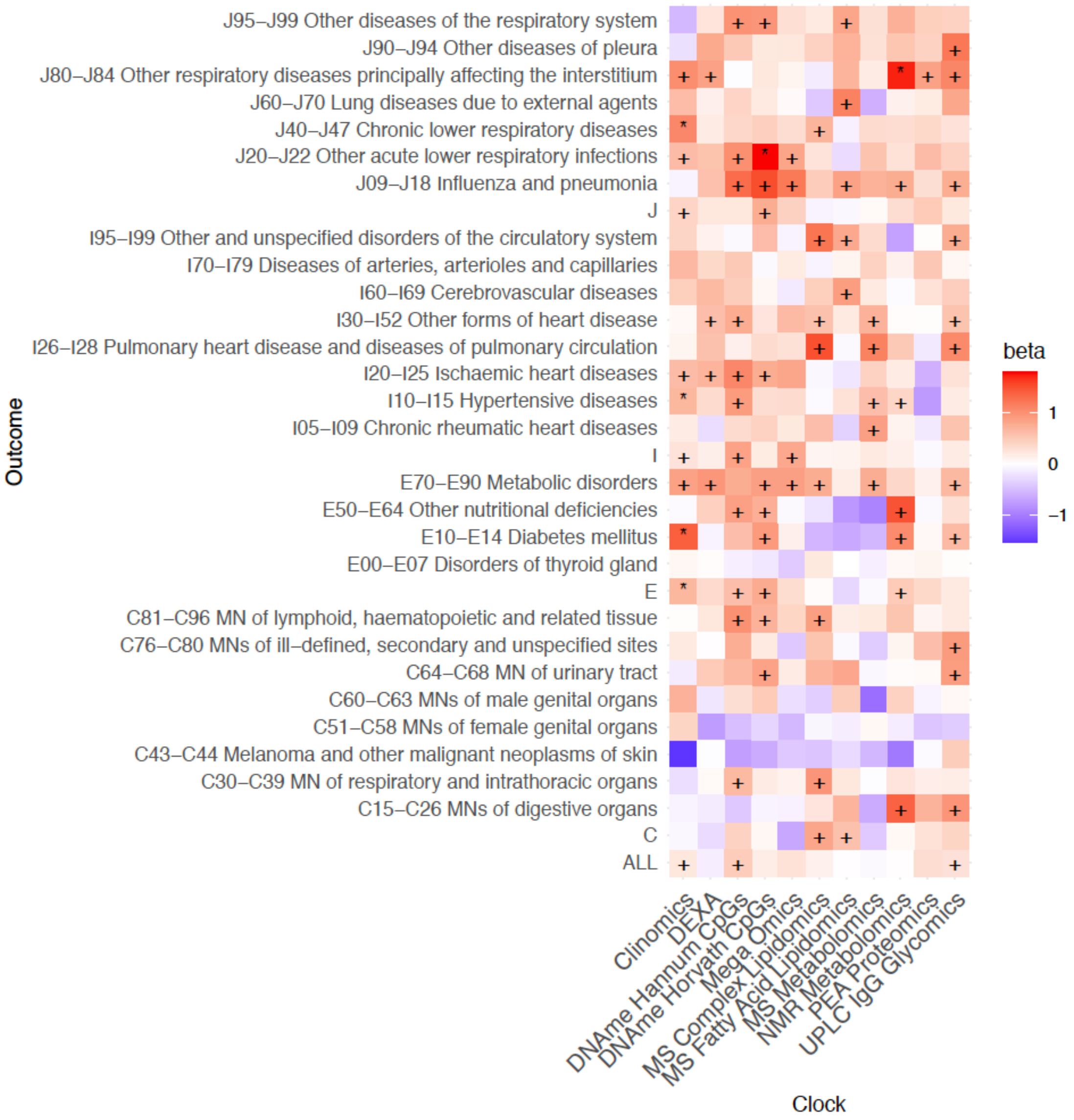


Figure 10. Positive age acceleration associations observed with increased disease risk. Associations with rates of hospitalisation. *+/* Association nominally/FDR<10% significant in the frequentist test that OCAA has a positive effect on outcomes. Beta: the relative effect of a year of OCAA to a year of chronAge on disease (initially measured in $\log_e$ hazard ratios, effect sizes are unitless after division). A value of one indicates that a year of OCAA is equally as deleterious as a year of chronAge and is indicated in salmon colour. To facilitate reading, note the DNAme Horvath CpGs-BMI beta is 1.02 and the DNAme Hannum CpGs-C81-C96I beta is 1.00. Clock: the omics clock on which OCAA was measured. Disease group: the set of diseases (defined by ICD10 codes) which were tested for first incidence after assessment against the clock, already prevalent cases were excluded (Case numbers for each disease block in **Supplementary Table 21**).*

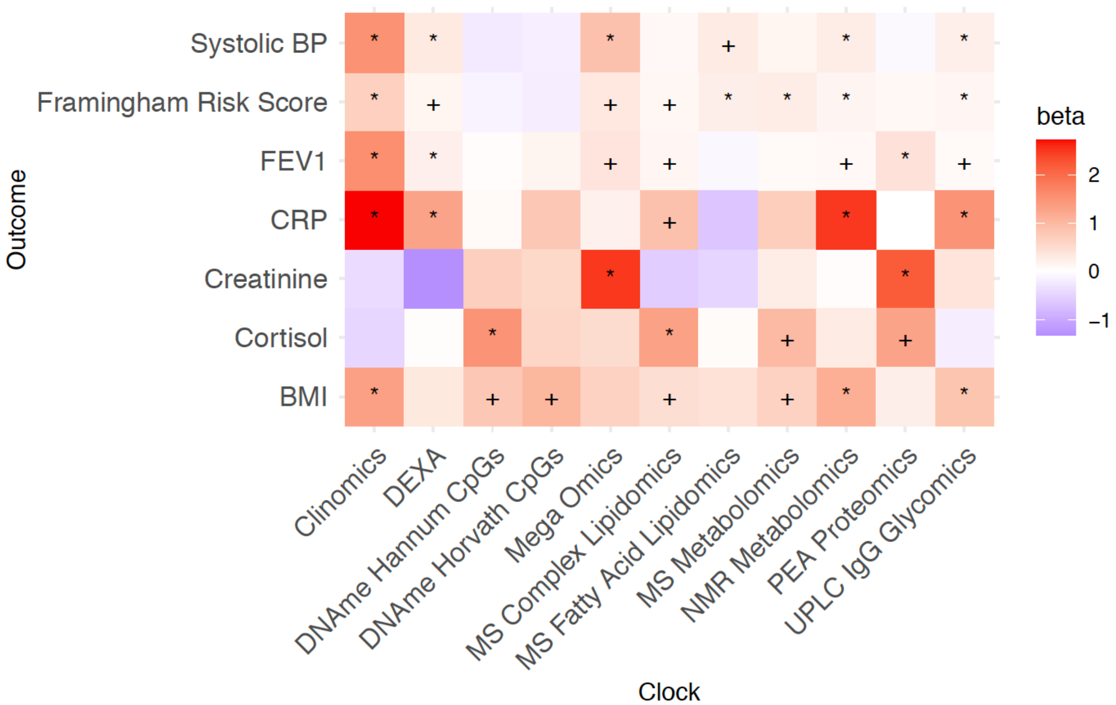


Figure 11. Positive age acceleration associations observed with increased disease risk. Associations with disease risk factors. *+/* Association nominally/FDR<10% significant in the frequentist test that OCAA has a positive effect on Risk factors. Beta: the relative effect of a year of OCAA to a year of chronAge on risk factor (effect sizes are unitless after division). A value of one indicates that a year of OCAA is equally as deleterious as a year of chronAge and is indicated in salmon colour. Total cholesterol, which showed a particularly large effect from MS lipidomics OCAA, is excluded here to aid visualisation (the effects on cholesterol can be seen in* ***Supplementary Figure 37****).*

Comparison of predictive abilities of different OCAAs for risk factors and disease

In order to determine which OCAAs could draw more meaningful distinctions between subjects in terms of health outcomes, we repeated the previous analysis using standardised OCAAs. As in principle, two OCAAs could have the same association effect size on disease, but one might be much more prognostic for the population as a whole than the other if it had much larger variation in its range. We found that the standardised Clinomics OCAA showed the greatest predictive power, with an IVW-average effect across all risk factors of 0.39 compared to the range of 0.05-0.12 for the other clocks, with Hannum and Horvath CpGs OCAA smaller still, at -0.014 and 0.018, respectively (SEs ~0.01, in all cases). Conversely, FEV1 was the risk factor on which standardised OCAA had the largest effect (0.20, SEs ~0.01, IVW-averaged across clocks), whilst standardised OCAA had the smallest effect on creatinine/reversed cortisol (0.02/0.04, SEs ~0.01).

Standardised OCAA effects on disease showed an even more uniform pattern (**Supplementary Figure 34**): the IVW-average effect across diseases was between 0.11 (MS Metabolomics) and 0.24 (Clinomics), except for the 0.016 of PEA Proteomics (SEs ~0.04). Despite limited power, the disease group showing the most sensitivity to standardised OCAA across clocks was J80-J84 (Other respiratory diseases principally affecting the interstitium; 0.76, SE~0.16), perhaps consistent with the FEV1 finding. Although predictive of risk factors, Clinomics OCAA does not appear unusually predictive of disease. Lung function appears particularly sensitive to both ageing (**Supplementary Figure 32**) and OCAA.

Clocks built from few omics principal components are effective predictors of health outcomes

Finally, we reduced dimensionality and assessed the underlying information about ageing being captured by different omics at the assay level, rather than simply the predictors selected for model inclusion. We constructed clocks using a few principal components (PCs) of omics measures as predictors and repeated the previous analyses with their (standardised) OCAAs, estimating chronAge (**Supplementary Figure 38**) and predicting health outcomes (**Supplementary Figure 39** & **Supplementary Figure 40**). The pattern was striking, the IVW-mean effect sizes across all risk factors of 3 PC OCAAs were more than double our standard OCAAs (**Supplementary Figure 39**). For all OCAAs, bar DNAme-based, including more omics PCs in the clocks reduced their ability to estimate distinctions in risk factors. IVW-mean effects on diseases were generally similar for the 3 PC and standard OCAAs, except for the PEA Proteomics OCAA, where 3 PCs- based clock outperformed the standard clock by a factor of 10. Overall, OCAAs derived from a few omic PCs appeared equally predictive as our standard OCAAs for diseases and more predictive for health risk factors.

Discussion

We have performed the most exhaustive comparison of different omics assays as potential biomarkers of age to date. We have shown firstly, it is possible to construct ageing clocks that produce highly accurate estimations of chronAge with a wide variety of omics biomarkers (correlation of OCA with chronAge ranged 0.66-0.97). Secondly, ageing clocks built using PEA proteomics, DNAme, UPLC IgG glycomics and clinical risk factors in ORCADES were able to estimate chronAge in independent populations. Thirdly, it is possible to achieve as highly accurate estimations of chronAge using a substantial subset of core biomarkers from each assay compared to our standard clocks. Despite finding only modest positive correlations between our OCAAs, we showed that different clocks overlap in the variation they explain in chronAge, more than would be expected by chance if they were independently sampling from a latent set of complete predictors. We found associations of OCAAs with total cholesterol, Framingham Risk Score, C-reactive protein and systolic blood pressure. We found 6 statistically significant (FDR<10%) individual associations and strong evidence of enrichment of association of OCAA with incident disease collectively across our tests (20% were nominally significant p<0.05). We found more variation in OCAA predictiveness across risk factors, than across diseases. Overall, we estimated that one year of OCAA has an effect of 0.46/0.45 years of chronAge on risk factors/disease incidence and showed that OCAA based on clocks built using a few principal components of omics were as prognostic as those presented with all available features.

The correlation of our PEA proteomics, DNAme, UPLC IgG glycomics OCAs and chronAge were similar to published models[54,55,66,87]. Unsurprisingly, DNAme-based clocks built in ORCADES were able to estimate age in both Scottish (Generation Scotland) and Estonian Biobanks (EBB), as the Hannum and Horvath epigenetic clocks have been used successfully in numerous populations. We showed for the first time that clocks built from Olink PEA-based proteomics replicate (in EBB and Croatia-Vis), while clocks built using the SOMAlogic[89] proteomics platform have been shown to replicate across populations previously. Our UPLC IgG glycomics clock also replicated in an independent population, mirroring the applicability of published GlycanAge measures[87]. Conversely, our NMR metabolomics and DEXA clocks had much lower correlation with chronAge in EBB and UKB. The success of these clocks appears to be study-specific: differences in lifestyle and environmental factors that change with age between the populations of the Orkney Islands and general populations in the UK and Estonia are a plausible cause. This finding serves as a warning as to the generalisability of ageing clocks to new populations.

For a measure of BA to be clinically useful and efficient, accurate biological age estimation based on as few predictors as possible is ideal. We substantially reduced the numbers of biomarkers from each assay that were included in our clocks and showed no dilution of performance across all of our clocks. Enroth *et al.*[54] showed that this was possible with a protein-based clock, however, we reduced the number of proteins by a larger factor and achieved the same accuracy estimating chronAge. This high performance with a substantial subset of predictors has not previously been shown systematically across nine different types of biomarkers.

The extremely high correlations with chronAge reported, such as the r = 0.97 of the Mega-omics OCA, highlight an issue that has been discussed in prior work: that if enough biomarkers were included in the model, it would be possible to perfectly estimate chronAge and, by definition, fail to detect (distinct) BA. Lehallier *et al.*[89] showed that correlation between OCA and chronAge increases with the number of proteins included in the model. Further, it is possible to explain 100% of the variance in chronAge using DNAme data in large samples[92]. A perfect age predictor would give no information about variation between individuals of the same age and even those which are near perfect will have too little variation in the OCAA to be indicative of health status or outcomes beyond chronAge[208]. We found this trend in our results, that the most accurate estimators of chronAge: Mega-omics, PEA proteomics and MS metabolomics OCAAs were not strongly associated with subsequent hospital admissions, nor DNAme-based OCAAs with risk factors. Of course extremely accurate estimators of chronAge do have their uses, for example in a forensic context[209], but are not useful in terms of BA. This does not mean the assays themselves cannot be used to estimate BA but highlights a limitation of training ageing clocks on chronAge.

A useful BA must be an indicator of health status or outcomes beyond chronAge. We found DNAme-based OCAAs were better estimators of incident disease than risk factors, consistent with the known performance of Horvath's epigenetic clock. Several groups have shown Horvath's DNAme OCAA to be associated with subsequent all-cause mortality[78,203,210–212]. Differences in Horvath's OCAA between cases and controls have been found for numerous disease phenotypes[67,69–78]. In contrast, Horvath's OCAA has been found not to be associated with common risk factors including: LDL cholesterol and CRP[206], a finding we confirmed. We found that Clinomics and lipid based OCAAs were better at predicting risk factors than disease, whereas the opposite was true for DNAme and UPLC IgG Glycomics OCAAs. The similarity between the predictors in the Clinomics and lipid-based clocks and some of the risk factors could be driving these associations. In contrast, DNAme and UPLC IgG Glycomics being prognostic of incident disease beyond chronAge suggests they are more likely to be capturing underlying BA.

It is perhaps not surprising that the Clinomics OCAA showed the strongest evidence of association with disease - it used common clinical measures thought to be prognostic of health. Nonetheless, the pattern is a reassuring proof of concept. The overall enrichment of OCAA-disease and -risk factor association, strengthens the case for the notion of BA, trackable through omics markers. Previously, it has been shown that GlycanAge is associated with risk factors[87] and that IgG glycans (i.e. not an OCAA, rather the glycan levels themselves) are effective predictors of incident type 2 diabetes and cardiovascular events[108,213,214]. However, we are the first to show UPLC

IgG glycomics OCAA to be prognostic of incident disease and highlight this is not simply due to tracking the risk factors we considered.

As by definition, having a BA of +1 indicates that the individual has the same functional capacity and risk of age-related disease as the average individual that is one calendar year older than them, indicating the effect of true BA is the same as 1 year of chronAge. Our estimate that the mean effect of 1 year of OCAA on disease incidence is the same as 0.45 years of chronAge is important. BA thus appears to be real and measurable and have effects of similar magnitude to chronAge, albeit our estimates are significantly diluted compared to chronAge, possibly due to OCAA capturing only some aspects of BA, reflecting the types of assay and tissue, rather than BA itself. Better measures of BA seem worthy of pursuit, as do interventions that can reverse well-measured BA. The negative association between Melanoma and other malignant neoplasms of skin (C43-C44) and OCAAs for many clocks, contrasts with the trend of positive OCAA-disease block associations, suggests that there may be more than one BA. If replicated, this will highlight that skin BA and other BAs need not closely align, and we speculate this finding might also generalise across other organs.

A strength of our work was the sheer number and range of assays and therefore omics ageing clocks whose performance we compared in the same individuals, whereas previous comparisons have been limited to DNAme-based clocks[93,94,215] or DNAme, clinical risk factors and frailty measures[95]. We have tried to validate our omics ageing clocks trained in ORCADES in independent populations where available, to illustrate their wider applicability. A limitation faced by previous studies was the narrow age range of individuals in the training sample, for example Lee *et al.*'s epigenetic clock trained in a pregnancy cohort produced extremely accurate estimations of chronAge for individuals under 45 but underestimated age in older individuals[216]. Our clocks avoid this limitation due to the wide age range (16-100) of individuals in the ORCADES cohort.

The novel assessment of excess overlap between clocks is a strength of this work, as it has not previously been shown that, across multiple different omics assays, OCAs overlap more than would be expected by chance if they were ISLSP, indicating these clocks are tracking more common rather complementary aspects of ageing. A further strength is the regularisation of effect sizes - we have measured the effect of OCAA per effect of year of chronAge - giving a natural and understandable scale. Another strength is its scope, with many clocks tested against many age-related diseases. Of course, this is also a weakness, as it reduces power after compensation for multiple testing. Nonetheless, the essentially agnostic view taken of individual disease groupings and clocks does mitigate the risk of publication bias.

A limitation of this work is the relatively small sample size, both in terms of the number of individuals with multiple omics assays and within that, the number of incident hospital admissions over the follow-up period. Due to the low number of deaths in our sample we are as yet unable to test for the association of OCAA on mortality, as in previous studies. As the omics data available for ORCADES is cross-sectional, we were unable to comment on the variation of OCAA within individuals over time. However, we were able to investigate the prognostic ability of single time point OCAAs on hospital admissions over a 10-year follow up. The nature of our sample, a population isolate, means there is potential for local factors to influence our results. We have shown this is not the case for several of our omics clocks' accuracies (**Supplementary Figure 28**), as they were successfully replicated in additional populations, however, it could contribute to the poor replication seen for the DEXA and NMR metabolomics clocks. The use of hospitalisation as a measure of incidence is a limitation, particularly acute for diseases normally treated in the community such as type 2 diabetes and influenza. Nonetheless, we are likely to have captured the most severe cases and have tested whether this severity associates with OCAA and presumed frailty, giving rise to more severe experience of the disease. Secondly, the correlated nature of the assays and of the disease outcomes mean our tests have not been independent, although this means the FDR corrections have been conservative. A more powered study might also try to disentangle individual markers especially those retained in our core omics clocks and consider their biological plausibility as sitting on the causal pathway.

Of course, association does not imply causation. Although the use of a prospective cohort has reduced the risk of reverse causation, undiagnosed cases (at baseline) might still have contributed to the effects we observe, although confounding where a latent set of underlying traits is influencing disease susceptibility and the biomarkers is perhaps more likely. Nonetheless, even in the absence of causation, OCAA does appear to often be a biomarker of disease and underlying BA.

In conclusion, our work has strongly further evidenced the existence of BA as distinct from chronAge, whilst highlighting a substantial part of the OCAA is noise. The data also suggested there may be more than one type of BA, as measured by different clocks and giving rise to differing amounts of disease susceptibility, most strongly implied by our evidence that skin age and heart age may move in opposite directions. We also highlight that some OCAAs (e.g. PEA proteomics) may capture specific risks and consequently associate with health, whilst others (e.g. DNAme and UPLC IgG glycomics) may capture more generalised ageing. Our observation that clocks derived from few PCs of omics are less accurate in estimating chronAge but better able to predict risk factors, suggests that the search for BA should be pursued through salient features of biology. This supports the recent success of ageing clocks trained on all-cause mortality based measures[96,97], DNAmePhenoAge[96] and GrimAge[97], which have been shown to be more prognostic of health and mortality outcomes than DNAme clocks trained on chronAge directly[94,95,204,217]. We therefore suggest that the focus of future research should continue to shift to clocks trained on mortality, or more ideally all-cause morbidity, that are prognostic of subsequent health outcomes rather than accurate chronAge estimators.

Acknowledgements

We thank Archie Campbell for assistance with the electronic health records from ORCADES. The Orkney Complex Disease Study (ORCADES) was supported by the Chief

Scientist Office of the Scottish Government (CZB/4/276, CZB/4/710), a Royal Society URF to J.F.W., the MRC Human Genetics Unit quinquennial programme "QTL in Health and Disease", Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). DNA extractions were performed at the Edinburgh Clinical Research Facility, University of Edinburgh. We would like to acknowledge the invaluable contributions of the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney. We acknowledge support from the MRC Human Genetics Unit programme grant, "Quantitative traits in health and disease" (U. MC_UU_00007/10).

The CROATIA-Vis and, CROATIA-Korčula studies were funded by grants from the Medical Research Council (UK), from the Republic of Croatia Ministry of Science, Education and Sports (108-1080315-0302; 216-1080315-0302) and the Croatian Science Foundation (8875); and the CROATIA-Korčula genotyping was funded by the European Union framework program 6 project EUROSPAN (LSHGCT2006018947). We thank the staff of several institutions in Croatia that supported the field work, including Zagreb Medical Schools, the Institute for Anthropological Research in Zagreb, the recruitment team from the Croatian Centre for Global Health, University of Split and all the study participants. We are grateful to the Helmholtz Zentrum Munchen (Munich, Germany), AROS Applied Biotechnology, (Aarhus, Denmark) and the Edinburgh Clinical Research facility, University of Edinburgh (Edinburgh, United Kingdom) for SNP array genotyping. Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006] and is currently supported by the Wellcome Trust [216767/Z/19/Z]. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, University of Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" (STRADL) Reference 104036/Z/14/Z). We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants and nurses. Ethical approval for the GS:SFHS study was obtained from the Tayside Committee on Medical Research Ethics (on behalf of the National Health Service).

This research was funded by the European Union through the European Regional Development Fund (N.T.) and SP1GI18045T (T.E., N.T.), grants PUT1660 (T.E., N.T.) and PUT1665 (K.F., N.T.) of the Estonian Research Council. The Estonian Biobank study was approved by the Ethics Review Committee on Human Research of the University of Tartu and the Estonian Genome Center, University of Tartu scientific committee. Written informed consent was obtained from participants in accordance with the Declaration of Helsinki. The work of L.K. was supported by an RCUK Innovation Fellowship from the National Productivity Investment Fund (MR/R026408/1).

We thank the UK Biobank Resource, approved under application 19655. We acknowledge funding from the UK Medical Research Council Doctoral Training Programme in Precision Medicine (MR/N013166/1).

Author Contributions

E.M.D: Project conception, methodology, manuscript drafting and editing, formal analysis, visualisation. N.T: Performed validation analysis in Estonian Biobank. L.K: Processed and performed QC on the UPLC IgG Glycomics data in ORCADES, Croatia-Vis & Croatia-Korčula and supervised A.F. A.F: Processed and performed QC on the DNA methylation data in ORCADES. R.W: Preparation of the DNA methylation data in GS:SFHS. P.K.J: Project conception, methodology, formal analysis, manuscript drafting and editing, project supervision. J.F.W: Project conception, manuscript drafting and editing, project supervision. All other authors commented and approved the manuscript prior to submission.

Competing Interests

P.K.J is a paid consultant for Humanity Inc. and Global Gene Corporation.

Methods

Cohort Data

Analyses were predominantly carried out using the Orkney Complex Disease Study (ORCADES)[186], a population-based isolate cohort that is extensively characterised in terms of both traditional phenotypes, omics assays and mean 12 years of follow up via linked electronic health records (EHR). The additional cohorts, Croatia-Vis and Croatia-Korčula[218,219], were used to validate omics ageing clocks trained in ORCADES. Croatia-Vis was used to validate a clock trained in ORCADES using a subset of proteins (those measured on the Olink CVDII, CVDIII and INFI panels) referred to as protein subset 1 and the UPLC IgG glycomics clock. Replication of the NMR metabolomics and UPLC IgG glycomics clocks trained in ORCADES was carried out in Croatia-Korčula. The Estonian Biobank[192] (EBB) cohort was used to validate a clock trained using a subset of proteins (those measured on the Olink CVII, CVDIII, INF1 and ONCII panels) referred to as protein subset 2 as well as the NMR Metabolomics clock. Both EBB and the Generation Scotland: Scottish Family Health Study (GS:SFHS)[190], a family-based cohort comprising volunteers across Scotland, were used to assess two DNAme-based ageing clocks. Finally, the UK Biobank[191] (UKB) was used to test the Clinomics and DEXA clocks trained in ORCADES.

Omics Assays

Dual X-ray absorptiometry (DEXA): Whole body imaging was performed on the Hologic fan beam DEXA scanner (GE Healthcare). Measures of body composition were derived from the DEXA scans using APEX2 software for bone, lean and fat tissue and

APEX4 software for android, gynoid, visceral and lean fat mass content. 28 measures in the following broad categories: bone mineral density, bone mineral content, fat or lean mass percentages for head, trunk and limbs were selected for analyses. These were measures that did not use chronAge in their calculation and were also available in the UK Biobank. Measures were removed as outliers based on a z-score cut-off of 6 then pre-corrected for sex. Residuals were additionally subject to a threshold by removing outliers with a z-score cut-off of 3.

DNA Methylation: The Illumina EPIC 850K array was used to measure DNA methylation levels in ORCADES. Quality control was carried out using the meffilQC pipeline[220] and minfi package[221]. Samples were excluded as outliers: if >1% of probes had a detection p-value > 0.01, due to failure of sex concordance, if samples showed evidence of dye bias or failed median methylation signal z-score cut-off of 3. Probes were removed as outliers if the detection p-value was >0.01 in >1% of samples or had a bead count of <3 in at least 5% of samples. The *preprocessNoob* function in the “minfi” package was used for array normalisation to remove unwanted technical variation. M values were corrected for the technical covariates: plate number (as a random effect), season of venepuncture, year of venepuncture, plate position and 10 principal components of the control probes (as fixed effects) using GCTA-REML[145].

Instead of creating novel DNA methylation clocks when there are landmark clocks available in the literature, we constructed clocks based on Hannum and Horvath’s original epigenetic clocks, to compare with our other omics. As ORCADES used the Illumina EPIC 850k chip rather than the earlier 450k/27k chips used by Hannum and Horvath, our methylation clocks are subsets of Hannum and Horvath’s clocks. It has been shown that imputing probes that are absent from the 850k chip but present in the 450k/27k set leads to underestimation of both published ageing measures[33]. Thus, for our clocks named Hannum CpGs and Horvath CpGs we presented 62/71 and 333/353 of sites, respectively, that were present on the 850k chip to the penalised regression algorithm for model selection. Residuals from REML within a z-score threshold of 6 were then corrected for sex.

NMR Metabolomics: The high throughput NMR metabolomics assay of EDTA plasma (Nightingale Health Ltd., Helsinki, Finland) quantified 225 metabolomics measures in molar concentration units. The measures include amino acids, ketone bodies, low molecular weight metabolites and numerous lipid and lipoproteins subclasses. In both ORCADES and Croatia-Korčula, metabolite measures were removed as outliers based on a z-score cut-off of 6, pre-corrected for sex and the use of statins as a binary variable. Residuals were additionally removed as outliers with a z-score cut-off of 3.

MS Fatty Acids Lipidomics: Shotgun lipidomics and liquid chromatography tandem mass spectrometry (LC-MS/MS) was used to quantify the molar concentrations of 44 fatty acids as described previously[222]. Fatty acid measures were removed as outliers based on a z-score cut-off of 6, pre-corrected for sex, box number, plate position and use of statins.

UPLC IgG Glycomics: The glycan data have previously been described in detail by Krištić *et al.*, for the ORCADES[87], Croatia-Vis and Croatia-Korčula[218,219] studies. Raw glycan measures were total area normalised and batch corrected using the "ComBat" function of the sva package[223] in R. The normalised glycan measures were excluded as outliers based on a z-score threshold of 6 and pre-corrected for sex.

PEA Proteomics: 1,102 proteins were measured using a proximity extension assay method (Olink Bioscience, Uppsala, Sweden)[224] from EDTA plasma in 12 x 92-protein panels designated by the manufacturer: cardiovascular 2, cardiovascular 3, inflammation 1, metabolism, cardiometabolic, cell regulation, development, immune response, organ damage, oncology 2, neurology and neuro-exploratory. Measures for all twelve panels are available for 1,057 individuals in ORCADES, with subsets available in Croatia-Vis (inflammation 1, cardiovascular 2 and cardiovascular 3) and EBB (inflammation 1, cardiovascular 2, cardiovascular 3 and oncology 2). PEA proteomics-based OCAs were re-derived using these subsets to allow comparison across populations. NPX values of proteins (on the log2 scale) including those non-missing below the lower limit of detection (LOD), were removed as outliers with a z-score cut-off of 6. These measures were then pre-corrected for the following covariates via fixed effects linear regression: sex, season of venepuncture, time the plasma sample was in storage between collection and assay (days), plate number, plate row and plate column.

Clinomics: This dataset consisted of 13 selected clinical measures that are routinely measured during visits with general practitioners and clinicians: albumin, fasting plasma glucose, calcium, uric acid, high density lipoprotein cholesterol (HDL), total cholesterol (TC), triglycerides, height, weight, forced expiratory volume in 1 second (FEV1), and diastolic (DBP) and systolic blood pressure (SBP).

MS Metabolomics & MS Complex Lipidomics: Non-targeted metabolomic and lipidomic features were detected and quantified using Metabolon as described previously[225]. The HD4 dataset comprised measures of 1,143 biochemicals while the complex lipids dataset measured 1,052 biochemicals, these were treated as two separate omics assays referred to as MS Metabolomics and MS Complex Lipidomics respectively. Measures were removed as outliers with a z-score cut-off of 6. These measures were then pre-corrected for the following covariates via fixed effects linear regression: sex, statin use, assay run day, plate number and plate row and plate column.

EHR: The ORCADES cohort has record linkage to hospital admission records (Scottish Morbidity Records: SMR01). The first occurrence of any hospital admission with ICD10 diagnosis, was taken as incidence. NHS Scotland records moved from ICD9 to ICD10 in April 1996, so diagnoses since ~12 years prior to assessment were captured. The disease groupings analysed included each ICD10 block within 5 Chapters thought a priori to associate with age II (Neoplasms - codes C), IV (Endocrine, nutritional and metabolic diseases - codes E), IX (Diseases of the circulatory system - codes I), and X (Diseases of the respiratory system - codes J). For Chapter II only C codes (malignant) were included in our analyses. Chapters as a whole were also analysed, as were all the diseases from these chapters simultaneously. Incident disease was defined as the time of first hospital admission with a diagnostic code recorded (in any position in the admission record) for any disease within the grouping being analysed. For each disease grouping, subjects with recorded admission prior to the date of venepuncture were then excluded entirely in the subsequent analysis, as already prevalent.

Quality Control of Omics Measures

Outliers were defined based on z-score thresholds that varied between omics datasets depending on the distributions of the raw measures. Omics measures were pre-corrected for known batch effects and covariates (specified above) using fixed effects linear regression or other specified methods. A second pass z-score threshold on the residuals was used to detect further outliers for a subset of assays and all missing values were removed. The residuals produced from covariate correction were then scaled and centred to have a mean of zero and a standard deviation of one to ensure that effect sizes of any variables included in the models were comparable.

Clock Construction

Per Omics Assay: The individuals in the ORCADES cohort were split into 75% training, 25% testing. For the analysis comparing clock performance across omics platforms the testing 25% of samples were taken preferentially from the pool of individuals that possess measures for all of the omics platforms. Tenfold cross validation in the training sample was used to select the shrinkage parameter, λ, for the penalised regression that was estimated to produce the model with the minimum mean squared error. Models were constructed using three different procedures implemented using the glmnet[201] and caret packages in R with chronAge at venepuncture as the dependent variable: i) least absolute shrinkage and selection operator (LASSO) regression ii) elastic net regression with an alpha of 0.5 iii) elastic net regression with alpha selected using 10-fold cross validation in the training sample. We found no difference in performance between the three methods, so construction using elastic net regression with an alpha of 0.5 was used throughout the analyses presented. This model was then used to estimate chronAge in the testing sample and an independent out of cohort sample if available.

As stochasticity is present in the procedure, the variables selected for model inclusion will vary depending on the individuals selected to be in the training sample. Clock construction was repeated 500 times and the features selected for inclusion and the correlation between chronAge and age estimated by the model were recorded. This was done to ensure that the model performance results presented here are representative, and not an outlier due to individuals at extreme ends of distributions contributing to the training sample and a rare model being used to draw conclusions (data not shown).

Mega-Omics: model that was presented with all of the features from all of the omics platforms. The dataset itself was created by merging all of the corrected omics measures (residuals) after platform level quality control, again standardising all features to have a mean of zero and standard deviation of one. The clock was created using the same construction procedure outlined above.

Core Models: were constructed per omics assay. The elastic net regression algorithm was presented with only those predictors that were selected for model inclusion in >95% of the 500 iterations of clock construction for the relevant omics platform. This reduced set of predictors then underwent clock construction as described above.

Principal Component Clocks: To ensure that the differences in variance explained in chronAge by different omics clocks is not due to the discrepancy between the number of features available and hence the number of features selected for model inclusion across omics types. But rather is a genuine difference in the information about ageing captured by different omics; clocks were built using principal components (PCs) of the relevant omics platform as features. The first 3, 5, 10 and 20 PCs were extracted from the covariate corrected scaled and centred omics data at the platform level using the *prcomp* function in R. These PCs were then presented to the elastic net algorithm and clocks built.

Correlation of OCAAs

Pairwise Pearson correlations between 10 of our OCAAs were calculated, Mega-omics OCAA was excluded from this and all between clock comparisons as it contains predictors spanning multiple assays.

Partitioning Variance Explained in ChronAge

The unique variance in chronAge explained by each clock, sr_i^2, was calculated as the squared part correlation of chronAge (Y) and age estimated by clock i while controlling for all of the other k clocks. Part correlations were calculated using the *spcor.test* function in the "ppcor" package in R[226]. The portion of variance in chronAge explained by all of the k clocks together, the R^2 from the following model:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots b_kX_k$$

Where Y is chronAge and $X_{1\ldots k}$ are age estimated by clocks 1 to k, was used to partition the total variance of chronAge further into that which remains unexplained by the 10 clocks ($\mathbf{1} - R^2$) and that which is explained by overlapping clocks:

$$\mathbf{1} - (\mathbf{1} - R^2) - \sum_{i=1}^{k} sr_i^2$$

To gain a more detailed insight into the relationship between clocks we carried out pairwise comparisons. Following the same procedure as outlined above, the unique variance in chronAge explained by each clock in the pair is the squared part correlation of chronAge and age estimated by one clock while controlling for age estimated by the other clock in the pair. The variance remaining unexplained by either of the clocks was $\mathbf{1} - R^2$ of a bivariate model. The overlap, calculated by subtraction, is specifically the variance in chronAge explained by both of the clocks in the pair. This is unlike overlap calculated in the previous step, where we were only able to state that this variance was not unique to a particular clock but unable to deconstruct further.

Assessing the Overlap between Clocks

We assessed whether the combined variance in chronAge explained by pairs of clocks deviated from what would be expected by chance if both clocks were independently sampling from a latent set predictors (ISLSP) of chronAge. The combined variance in chronAge explained by both clocks together was calculated as the multiple R^2 from a bivariate model, with chronAge being the dependent variable and the estimated ages from the two clocks in the pair the independent variables. The variance explained in chronAge (v_i) by each clock (i) individually was the univariate R^2 from the regression of estimated age on chronAge. The expected variance in chronAge explained by two clocks by chance (E) was calculated as follows:

$$E = \mathbf{1} - (\mathbf{1} - v_1)(\mathbf{1} - v_2)$$

The idea being that the variance in chronAge not already explained by the first clock is $\mathbf{1} - v_i$. With the null hypothesis that the two clocks are independent samples from the latent set of complete predictors and thus explain partly overlapping information about age. The expected left unexplained after the addition of the second clock is thus $(\mathbf{1} - v_1)(\mathbf{1} - v_2)$.

To allow for the comparison of the deviation of observed variance explained in chronAge ($\boldsymbol{O}$) from expected (E) across pairs of clocks, this deviation was re-scaled. As the magnitude of v_i effects the possible range of values $\boldsymbol{O}$ could take. The theoretical minimum variance explained (E_{min}) by two clocks is the variance explained by the larger of the two clocks alone (the second clock only providing information already captured by the first). The theoretical maximum (E_{max}) is $v_1 + v_2$ or 1 if $v_1 + v_2 > \mathbf{1}$ (the clocks are explaining entirely non-overlapping variance). Comparisons containing clocks with high v_i will have a much smaller range of possible $\boldsymbol{O}$ than those with low v_i so directly comparing the magnitude of the deviation of observed from expected is not ideal. The results presented are on a scale of excess overlap calculated as follows:

$$\frac{E - O}{E - E_{min}}$$

With a value of 0 meaning that the observed variance explained equals that expected by chance if the clocks were independent. A value of 1 denoting that no additional variance was explained with the addition of the second clock. Negative values are possible and mean that the two clocks overlap less than expected and track separate aspects of chronological ageing, but in practice, we see that the clocks always track more common aspects than would be expected under the null hypothesis, albeit to varying degrees.

Association with health-related phenotypes & Incident Disease

OCAAs were tested for association with health-related risk factors and age-related incident diseases, as measured by hospital admission.

Association with chronAge: We first tested whether the risk factors and disease outcomes were associated with chronAge. For incident disease: time from assessment to incidence or to study end (the date when SMR01 records were extracted: December 2017, around ten years after assessment) was modelled using a Cox proportional hazard model[227] and the Surv function in the “survival” package in R. Subjects with prevalent disease were excluded. The baseline hazard was dependent on time since assessment, and hazards ratios dependent on chronAge and sex. We used time since assessment as the determinant of base hazard rather than chronAge, so that we could determine which groupings had stronger age-related effects and compare the effects of OCAA to those of chronAge. P-values for association with chronAge (and later OCAA) were calculated using a one-sided test, with H_1 being that chronAge increased risk.

Association with OCAAs: with standardised risk factors (units of phenotypic standard deviation) were carried out using linear regression with chronAge and sex fitted as fixed effects covariates. To restrict the burden of multiple testing we only tested the association of OCAAs on risk factors or disease blocks which showed a statistically significant association (effect size >0) with chronAge at outset (Benjamini-Hochberg FDR<10%) and had >5 incident cases (disease blocks). We tested the effect of OCAAs on each disease grouping using the same model as for chronAge, including chronAge and OCAA as effects. OCAA was not standardised but observed effect sizes were rescaled (divided) by the effect of chronAge, using the same model, enabling a comparison of the effect of one year's OCAA with one year's chronAge, with a value of 1 denoting the same effect. False discovery rate was again determined using the Benjamini-Hochberg method (FDR<10%).

Across both risk factors and disease, we found that large estimated effects arose in the context of large SEs. To facilitate visualising the results we had most confidence in we applied a shrinkage method, imposing a prior assumption on the distribution of beta (mean 0, SD 1) to the likelihood of our observed beta, shrinking resultant estimates with larger SEs more towards 0.

Individual tests of association generally had limited power due to multiple testing and the low variance of OCAA (compared with chronAge). We therefore considered the results of each OCAA across multiple outcomes by inverse variance weighting (IVW) observed results for individual outcomes. The covariance amongst outcomes and predictors, mean that the independence assumption for meta-analysis (or sign testing) is violated. Whilst this should not bias estimates, their precision will be overstated. We consider these results to be descriptive, and not conformable to formal testing. We use "~" to denote SEs calculated under the violated independence assumption, but still consider these useful to give a sense of magnitude. Conversely, for the same reason, the formal tests we perform (FDRs) are likely to be conservative.

We repeated these analyses with standardised OCAAs to compare the prognostic ability of different OCAAs at a population level, across risk factors and diseases and with our PC clocks OCAAs.

3.3 Conclusion

By performing the most exhaustive comparison of different types of omics assays as potential sources of biomarkers of biological age, we found that it is possible to build models that produce OCA estimates that are highly correlated with chronAge from a wide range of omics assays. We showed systematically across omics, that a substantial subset of biomarkers is required to achieve the same performance as with standard models and that omics ageing clocks overlap in the information they provide about age, with 94% of the variance in chronAge not being unique to one clock. Finally, we showed that omics clock age acceleration estimates (OCAAs) are associated with health-related risk factors and are prognostic of incident disease over and above chronAge.

This work highlighted several issues for the field that require further discussion. First that lots of sources of data could be used to produce accurate chronAge estimates, given enough predictors. On the one hand, this may be useful as numerous high dimensional omics assays are expensive and the flexibility to use only one or the ones that are already measured in the sample is reassuring. On the other hand, too many predictors will exacerbate the issue of overfitting discussed previously and will make ageing clocks study population specific, thus reducing their potential clinical applicability. Larger sample sizes and greater diversity in the sample used for model training will increase the chance of producing ageing clocks that will be effective across populations.

Second, that there is biomarker redundancy both within and between omics assays in terms of chronAge estimation. Our core models highlight the predictor redundancy within assays. The finding that pairs of OCAs overlap in the variance in chronAge that they explain, more than would be expected if they were independently sampling from a complete set of latent predictors, emphasises the redundancy between assays. Although taking steps to minimise the number of biomarkers included in BA clocks has been done before[54], we were the first to show that it is possible systematically across 9 different assays. Further, it is not common practice in the field and would be beneficial given the desire for ageing clocks to be clinically useful.

Third, we quantified the proportion of our OCAA that may be capturing noise rather than true underlying BA. As mentioned in the discussion, due to training on chronAge, by definition the effect on functional capacity and risk of age-related disease of one additional year of underlying BA is the same as 1 year of chronAge. The estimate that the mean effect of one year of OCAA on incident disease is the same as 0.45 years of chronAge rather than 1, highlights how much our estimate is attenuated due to OCAA containing noise. This is not usually addressed in the field and has not previously been quantified. Routine quantification of the proportion of OCAA that is capturing noise rather than potentially true BA, would be a substantial step forward in evidencing whether these models are indeed effective biomarkers of underlying BA.

Fourth, all of the omics assays bar DEXA scans were carried out on blood samples. This is an advantage in that it is minimally invasive and convenient to measure, however does mean that we did not consider tissue specific or multi-tissue clocks, which may prove useful to create system or organ specific ageing models. If such datasets become available, it would be invaluable to the field to capture the potentially different BA of different organs and tissues.

Despite showing that OCAA are associated with health-related risk factors and prognostic of incident disease beyond chronAge, effect sizes were modest. In our case this may be due to low sample size, but in general this is the case for clocks that are trained on chronAge[93–95].

As mentioned previously, promising results have been found using the ageing clocks DNAm PhenoAge and GrimAge that have been trained on all-cause mortality based measures[96,97] rather than chronAge. With DNAm PhenoAge, shown to be more prognostic of all-cause mortality, number of co-morbidities, probability of being disease free and increase in physical functioning problems than either Hannum or Horvath's epigenetic clocks[96]. In turn, AgeAccelGrim has also been shown to be prognostic of incident coronary heart disease, time-to-congestive heart failure, hypertension and type 2 diabetes as well as to outperform other DNAme-based OCAAs in predicting time to any cancer.

This trend that ageing clocks trained on mortality measures rather than chronAge, are more indicative of future health status was also reported by three groups comparing multiple DNAme based ageing models. Maddock *et al.* found that OCAAs from GrimAge and DNAm PhenoAge were associated with physical performance, cognitive performance and subsequent decline in performance. These associations were not found with OCAAs based on Hannum and Horvath clocks[94]. Upon fitting 9 different DNAme- or clinical risk factor-based OCAs simultaneously, Li *et al.*[95] showed that GrimAge, Horvath's DNAmAge and frailty index (FI)[228] based OCAA were prognostic of all-cause mortality, independent of chronAge and the 6 remaining OCAAs. When comparing GrimAge, DNAm PhenoAge, Horvath DNAm Age, Hannum DNAm Age and a DNAme based estimator of telomere length (DNAm TL)[229], Hillary *et al.* showed that only GrimAge and DNAm PhenoAge were statistically significantly associated with incident disease (COPD, type 2 diabetes, and heart disease) after correction for multiple testing[204], while none of the clocks trained on chronAge were prognostic.

Our work therefore supports the shift of focus of the field to concentrate on ageing clocks trained on mortality, or more ideally multiple morbidity, that are prognostic of incident disease rather than those that are merely accurate predictors of chronAge. In the next chapter, I investigate the potential of the multiple omics assays discussed in this chapter, as potential sources of biomarkers of health-related risk factors and incident disease directly, rather than their derived OCAAs.

Chapter 4: Biomarkers of Incident Disease

4.1 Introduction

In the previous chapter we sought to understand how omics profiles varied with age and the degree to which OCAA is predictive of disease, inferring that successful prediction meant that OCA is biomedically meaningful beyond chronAge, and not just a regression artefact. Whilst such an approach is hoped to capture generalised ageing, as age is the most significant risk factor for late-onset diseases, it seems likely that individual predictions of specific diseases would be more accurate if a more direct approach were taken: basing predictions on the omics measures directly, rather than OCAA.

Historically, research into biomarkers and risk scores has focussed on traditional clinical measures to build risk scores for clinical outcomes, for example incidence[101–103], or recurrence[104] of diseases, outcome post-surgery[105,106] and response to different therapies[107].

But recently, as assay technology and computing have advanced, so have the range of predictors used beyond conventional clinical measures to include high-dimensional omics assays[108–111] and image analysis[112]. These studies have used a variety of statistical methods to create multivariable scores, ranging from stepwise regression[108] and penalised regression[113] to more sophisticated machine learning techniques such as: random forest[110], neural networks[83] and deep learning[112].

Several of the omics assays that are available for ORCADES have previously been shown to be biomarkers for several different risk scores and outcomes. Menni *et al.* showed that 46 IgG glycans from the same ULPC assay as in ORCADES, were associated with 10-year atherosclerotic cardiovascular disease risk score (ASCVD)[230], a leading measure of CVD risk. PEA Proteomics have also been found to be significant disease biomarkers. Scores containing 11 plasma protein levels prognostic of cervical cancer[114] and ovarian cancer[109] have been reported by Berggrund *et al.* and Enroth *et al.,* respectively. Gisby *et al*. found 203 plasma protein levels associated with the clinical severity of COVID-19 patients as well as using random forests to create a score that predicts COVID-19 severity[110].

Plasma metabolite levels from the MS Metabolomics assay have been shown to be biomarkers for numerous incident diseases and of multimorbidity[111]. This study by Pietzner *et al.* emphasised that 65.6% of metabolite levels were significant biomarkers for multiple incident disease phenotypes and highlighted relationships between risk factors and incident disease that were mediated by metabolites, thus indicating actionable shared pathways.

As ORCADES has neither a large sample size in terms of number of individuals with omics measures nor incident disease cases, I instead, aim to take advantage of the breadth of phenotypes available. This is a unique opportunity to investigate omics measures as biomarkers of health outcomes in a curated dataset with 10-year follow-up for many diseases, however this required caution with regards to multiple testing and a rigorous training-testing split. This therefore has the potential to yield an interesting exploratory analysis, revealing perhaps which omics assays or biomarkers and diseases are most tractable, as well as what increases in sample size or years of exposure might be required in future analyses.

I also intend to investigate the relative importance of the omics biomarkers selected for inclusion in risk scores, to identify and highlight the most promising predictors of subsequent incident disease.

With this in mind, in this chapter I: investigate the potential of these omics measures as biomarkers of health-related risk factors and incident diseases directly, without going through an OCAA; use the multiple omics assays and electronic health record data available in ORCADES and use the pipelines set up for the analysis in the previous chapter. This investigation of whether omics make suitable biomarkers for risk factors/disease either on their own or in composite models is thus likely to be of great value.

4.2 Methods

4.2.1 Data

Cohort Data

Analyses described in this chapter were carried out using data that has been previously described in **Chapter 2: Data & Methods** and **Chapter 3: Biological Ageing Clocks**. Namely, the same samples, omics assays, electronic health records and other phenotypic information, subject to the same quality controls.

Analyses were carried out using record linkage of the ORCADES cohort to hospital admission records, Scottish Morbidity Record: SMR01. The same disease blocks and risk factors were considered as in **Chapter 3: Biological Ageing Clocks** (list of disease blocks **Supplementary Table 20**) with the addition of educational attainment measured in years of schooling completed.

Omics Data

The same omics assays: DEXA, NMR Metabolomics, MS Fatty Acids Lipidomics, UPLC IgG Glycomics, PEA Proteomics, Clinomics, MS Metabolomics, MS Complex Lipidomics, and two sets CpGs (DNA methylation), subsets of those used in Hannum and Horvath's epigenetic clocks. The complete list of biomarkers per assay is in **Supplementary Table 17**.

Each omics assay underwent assay level quality control as described in **Chapter 2: Data & Methods Table 3**. The only difference in assay level quality control from the previous chapter, is that all assays were also corrected for chronological age at venepuncture, in addition to the previously stated covariates. As in contrast to

Chapter 3: Biological Ageing Clocks, where I was interested in omics measures' relationship with age, here the aim is to assess omics biomarkers association with risk factors and incident disease, independent of age. A complete list of biomarkers for each assay passing quality control for this analysis is in **Supplementary Table 22**.

Martingale Residuals

As the pipeline for quality control and construction of penalised regression models that I have used throughout this thesis is set up to work with quantitative traits, the use of time-to-event data is not ideal. However, Therneau *et al.*[231] demonstrated an approach to calculate covariate-corrected residuals from a Cox proportional hazard model that can then be used as a quantitative trait. These residuals named "Martingale residuals" are calculated for each individual as the difference between the number of observed events during the study period and the number of events expected given the model/values of the covariates in the study population. For this analysis Martingale residuals were calculated by Peter Joshi using an already established group pipeline as follows. Cox proportional hazard models for time to first hospital admission for each of the 44 disease blocks were fitted with age at venepuncture and sex as covariates (as described in chapter 3). Martingale residuals were then scaled by the proportion of events in the population, any linear association effect on this scale estimates the log hazard ratios[231,232] in the Cox model.

4.2.2 Univariate Associations of Omics measures with Martingale residuals

Univariate linear regressions were performed for each of the 3,302 QC'd omics measures on the Martingale residuals (units $\log_e$HR) of 44 disease blocks and 10 standardised risk factors (units of phenotypic standard deviation), with age at venepuncture and sex fitted as covariates. As the risk factors: HDL cholesterol, total cholesterol, FEV1 and systolic blood pressure are also Clinomics predictors they were not included in the 3,302 omics biomarkers, reducing the total number of tests from

178,524 ($3{,}306 \times 54$) to 178,308. Omics-outcome associations that passed 5% FDR using the Benjamini and Hochberg method[233] were considered significant. Plots were generated using the circlize R package.

4.2.3 Penalised Regression

Each omics assay was then considered in turn, the individuals in ORCADES with both assay measures and outcome measures were split into training and testing. For analyses where the outcome was a risk factor, the sample was split randomly into 75% training and 25% testing. For analyses with Martingale residuals as an outcome, cases and controls were still split randomly but separately into 75% training and 25% testing, to ensure that the case-control ratio was consistent between training and testing samples.

Two penalised regression models were built for each combination of omics assay and outcome: LASSO and elastic net with a fixed alpha of 0.5 using the glmnet package in R[201]. The linear association of the outcome predicted by the penalised regression model on observed outcome was used to compare performance across all omics and outcomes.

Biomarkers selected for model inclusion for each assay-outcome combination using LASSO regression and their coefficients are indicated in **Supplementary Table 23**.

4.2.4 Controlling Multiple Testing

Given the number of omics assays, 11, and outcomes, 54 (44 diseases and 10 risk factors), and the limited power due to small sample size and low numbers of cases for many of the disease blocks, I sought to limit the number of tests I performed to maximise the chance of achieving statistically significant results.

Diseases and risk factors were considered as two separate experiments, and for each, the following procedure was followed to identify which analyses would be sufficiently powered and would be taken forward for formal statistical testing.

Models predicting each of the outcomes were built using each of the different omics assays in the training sample, and the fit of predicted outcome on observed outcome assessed. In order to determine which analyses were expected to be powered in the testing sample, I calculated the p-value expected in the testing sample. This was possible as, if the model is not overfit, the effect size in the testing sample is expected to equal the effect size in the training sample. Because the testing sample is one third of the size of the training sample, and under the assumption that the observed training effect is the true effect, the expected t-test statistic in the testing sample is:

$$t_{testing} = \frac{\beta_{training}}{SE_{testing}}$$

Where the standard error expected in the testing sample is:

$$SE_{testing} = \sqrt{3} \times SE_{training}$$

Under these assumptions, expected two-sided p-values in the testing sample were then calculated using the “pnorm” function in R. An iterative process was then used to select analyses that were expected to pass a 5% FDR significance threshold. All n possible analyses were ranked based on expected p-value, and expected q-values were calculated using the Benjamini and Hochberg method[233]. This process was repeated using the top $n - 1, n - 2 \dots n - k$ analyses (ranked by expected p-value), until all of the top k analyses expected q-values were less than 5%. These were the k analyses that were plausibly powered to detect statistically significant signals, taking into account correction for multiple testing.

Importantly, no use of the testing data had been made at the point of determination of this list. This list was therefore being taken forward to an independent testing set, and so whatever the merits of the selection criteria for the list, the FDR measured in the test set is valid. The naïve expectation was of course that every test would be passed, but should this turn out not to be true, even to a material extent, it does not invalidate the independent test set FDRs, but indicates that the predicted p-values from training were not always borne out, perhaps due to chance or overfitting.

Thus, for the $\mathbf{k}$ analyses taken forward for formal statistical testing, those whose observed q-values in the testing sample passed an FDR 5% significance threshold are presented as statistically significant results. For associations with both diseases and risk factors, standardised effect sizes are presented (both score and outcome in standard deviation units). This means for interpretation, effect sizes for score-outcome associations indicate the change in outcome (in outcome standard deviation units) for every standard deviation increase in the score.

4.2.5 Score Profiling

As discussed in the introduction, I sought to understand the relative importance of each predictor within each score. Scores were of course a linear combination of the predictors. However, as the predictors within omics assays are not orthogonal, simply taking the squared multivariable or univariable standardised effect sizes of the predictors on the score, is not an appropriate way to partition the variance contribution to the score amongst the predictors. At its simplest the sum of such measures will not equal one.

In an attempt to minimise the contribution estimates being down to an artefact of the method, I used two different approaches to partition the variance explained in the score amongst the predictors. First, I used the hierarchical partitioning algorithm implemented by the hier.part package in R[234]. This package is based on the algorithm proposed by Chevan & Sutherland in 1991 that averages over all possible orders of variables, producing an estimate of the independent effects of each variable as a proportion of the total R^2[235]. The implementation of this approach by the hier.part package however, is optimised for models with 10 predictors or fewer. For this reason, I was only able to use this approach to partition the variance explained in scores with <10 predictors, so I also used an alternative approach for all of the scores of interest, allowing comparison between approaches for those with <10 predictors.

As for scores with large numbers of predictors it is not feasible to average over all possible orders, I used an iterative approach to rank predictors based on their univariate R^2, to determine an order of importance of predictors. This approach I have termed "iterative ranking" consisted of first ranking the predictors based on the estimate of their univariate variance explained, when fit in a model against the score. The predictor with the largest estimated R^2 (best single predictor) was then fit against the score and the residuals of this linear model calculated. These residuals conceptually being what remains of the score after having the variance explained by the first predictor removed and are considered as the "corrected outcome". The remaining predictors were then, one at a time, fit in a linear model against the corrected outcome and ranked based on the univariate R^2. The predictor with the largest R^2 from this step was then considered the second-best predictor. This process of finding the next best predictor based on ranking of univariate R^2, fitting this next best predictor against the previously corrected outcome and calculating residuals, fitting the remaining predictors one-by-one against the new corrected outcome and ranking again continues until an order of importance (the length of the number of predictors in the score) is determined. The predictors were then fit against the score, one-by-one, in this determined order of importance and the additional multivariate R^2 added by the addition of each predictor created the profile for the score.

I produced profiles for 5 Clinomics scores for: I20-I25 Ischaemic heart disease, I10-I15 Hypertensive diseases, E65-E68 Obesity and other hyperalimentation, E10-E14 Diabetes mellitus and "E" all block E metabolism related disorders and 4 scores for E10-E14 Diabetes mellitus constructed from: DEXA, Mega-Omics, MS Metabolomics, NMR Metabolomics and PEA Proteomics. I was also interested to see how the profile for my Clinomics score for FRS compared to the FRS used in clinical practice. For all of the scores discussed previously, the total multivariate R^2 of the fit of predictors on score was equal to one however, FRS is not a linear combination of its components

(Formula for FRS **Equation 1** [101]) and contains components that are not part of my Clinomics dataset. This meant that the total R^2 of a linear model fitting the components of FRS that overlap with Clinomics against FRS was <1. So in order to compare like with like, I considered the proportion of the total R^2 that was additionally explained by the addition of each predictor in the determined predictor order.

$$\begin{aligned}\boldsymbol{RiskFactors} = {} & [\mathbf{ln}(age) \times AgeFactor] + [ln(TotalChol) \times TotalCholFactor] \\ & + [\mathbf{ln}(HDLChol) \times HDLCholFactor] + [ln(SysBP) \times SysBPFactor] \\ & + Cig + DM - AvgRisk\end{aligned}$$

$$\boldsymbol{FRS} = \mathbf{100} \times \left(\mathbf{1} - RiskPeriodFactor^{e^{RiskFactors}}\right)$$

Equation 1. Formula for Framingham Risk Score.

Where age is age at venepuncture, TotalChol is total cholesterol (mg/dl), HDLChol is HDL cholesterol (mg/dl), SysBP is systolic blood pressure (mmHg) and the coefficients for individuals of European ancestry are in **Table 12**.

Coefficient	Men	Women	Note
AgeFactor	3.06117	2.32888	
TotalCholFactor	1.1237	1.20904	
HDlCholFactor	-0.93263	-0.70833	
SysBPFactorUntreated	1.93303	2.76157	not on HTN treatment
SysBPFactorTreated	1.99881	2.82263	on HTN treatment
Cig	0.65451	0.52873	if current smoker
DM	0.57367	0.69154	if T2D
AvgRisk	23.9802	26.1931	
RiskPeriodFactor	0.88936	0.95012	

Table 12. Coefficients for Framingham Risk Score for European Ancestry Individuals. *Indicating the coefficient for the above formula for both sexes of European ancestry. HTN: hypertension. T2D: type II diabetes. These coefficients are optimised for individuals between the ages of 30-74 who have not had previous cardiac events such as myocardial infarction and strokes.*

4.3 Results

4.3.1 Martingale Residuals vs Cox Proportional Hazard Models

Prior to assessing the potential of omics biomarkers to directly predict incident disease (using incident hospital admissions as a proxy), I first investigated whether using Martingale residuals of incident disease with my data was a valid approach, rather than fitting Cox proportional hazard models and using time to event data throughout subsequent analyses.

To do so, I repeated the analysis in **Chapter 3: Biological Ageing Clocks Figure 10**, that fit Cox proportional hazard models of OCAA with age (at venepuncture) and sex as covariates, against time to hospital admission. This time fitting linear models testing each of the 11 OCAAs against Martingale residuals of 44 disease blocks. Again, I included age and sex as fixed effects covariates and assessed the concordance between the effect size estimates and standard errors from the two approaches.

Visual inspection of the effect size estimates from both approaches appear to concur across both omics assays and disease blocks (**Supplementary Figure 41**). This, together with the consistency of effect size estimates over 3 different concordance measures across all OCAA-disease block associations (**Table 13**), led to the use of Martingale residuals for incident disease as the outcome and the use of linear models for the remaining analyses presented in this chapter.

Concordance Measure	Beta	Standard Error
X1/X2	0.891	0.914
X1-X2	-0.0195	-0.0403
abs(X1-X2)	0.0953	0.0763

Table 13. Concordance of Effect Size Estimates from Cox models and Martingale residuals. *Indicating the mean estimate of concordance across all OCAA-disease block associations for each measure when X is Beta (effect size in units of log_eHR/standard deviation of OCAA) and standard error are X, estimates from the Cox model are estimate 1 and those from Martingale residuals are 2. For example, 0.891 is the mean of $\beta_{Cox}/\beta_{Martingale\ residuals}$ estimates across all OCAA-disease block associations.*

4.3.2 Univariate Associations of omics biomarkers and outcomes

To determine if single omics measures are potential biomarkers of health outcomes, I performed linear associations fitting age and sex as covariates for 3,302 omics predictors against the 54 outcomes. I found 8,526 (4.78% of all tests) significant (5% FDR) biomarker-outcome associations between 2,686 single omics biomarkers and 54 outcomes (**Figure 12**), with evidence of enrichment of associations as 12.77% of the tests were nominally significant ($p<0.05$). The ratio of monounsaturated fatty acids to total fatty acids (MUFA/FA) from the NMR Metabolomics assay was significantly associated with the most outcomes (12: 7 risk factors and 5 diseases). (**Supplementary Figure 42**). Additionally, 7 biomarkers had 11 significant outcome associations including plasma glucose levels, weight, total trunk mass, three metabolites from the MS Metabolomics assay and tumour necrosis factor receptor superfamily member 6B (TNFRSF6B) protein level. In contrast, 629 biomarkers (across 7 omics assays) were only associated with one outcome and 616 associated with no outcomes.

All omics assays had at least one significant biomarker-outcome association, with Clinomics showing the highest number of associations relative to the number of measures in the assay and the Hannum and Horvath subsets of CpGs the least (**Table 14**). All biomarkers in the Clinomics, NMR Metabolomics and MS Fatty Acid Lipidomics assays had a significant association with at least one outcome. Interestingly, despite only 33.6% of the Horvath subset of CpGs having significant outcome associations, they were associated with the greatest number of different outcomes (44/54). This is in direct contrast to the MS Fatty Acid Lipidomics assay, 100% of whose biomarkers were associated with only 11 different outcomes.

Omics Assay	N Biomarkers	N Associations	N Associations Per Biomarker	Percentage of Biomarkers with Associations	N Outcomes	Percentage of Tests Significant
MS Complex Lipidomics	908	3318	3.65	97.58	41	6.77
MS Fatty Acid Lipidomics	32	144	4.5	100	11	8.33
PEA Proteomics	1102	2776	2.52	83.39	53	4.66
MS Metabolomics	682	1460	2.14	80.94	53	3.96
DNAme Hannum CpGs	62	46	0.74	51.61	22	1.37
DNAme Horvath CpGs	333	150	0.45	33.63	44	0.8
NMR Metabolomics	68	356	5.24	100	18	9.69
UPLC IgG Glycomics	77	91	1.18	62.34	14	2.19
Clinomics	9	57	6.33	100	20	11.73
DEXA	29	128	4.41	96.55	19	8.17

***Table 14. Omics Assay Level Associations with Health Outcomes**. N Biomarkers: number of biomarkers in omics assay included in association analyses. N Associations: number of significant (5% FDR) biomarker-outcome associations. N Associations Per Biomarker: number of significant (5% FDR) biomarker-outcome associations divided by the number of biomarkers in the assay. Percentage of Biomarkers with Associations: the percentage of biomarkers that were significantly associated with at least one outcome. N Outcomes: the number of different outcomes that biomarkers in each assay were associated with. Percentage of tests significant: the percentage of significant (5% FDR) biomarker-outcome associations out of those tested for that assay, calculated as (N Associations/(N Biomarkers*N Outcomes tested))*100.*

Outcome	Description	N Associations	N Assays	Percentage of Significant Tests
all	ALL	35	6	1.06
bmi	BMI	1231	10	37.28
bp_sys	SBP	331	8	10.02
c	Cancers	3	2	0.09
c00.c14	MN Lip/Throat	42	4	1.27
c15.c26	MN Digestive	11	5	0.33
c30.c39	MN Respiratory	17	4	0.51
c43.c44	MN Skin	17	6	0.51
c45.c49	MN Soft Tissue	13	4	0.39
c50.c50	MN Breast	11	3	0.33
c51.c58	MN Female Genitals	6	3	0.18
c60.c63	MN Male Genitals	6	4	0.18
c64.c68	MN Urinary Tract	61	7	1.85
c69.c72	MN eye/brain/CNS	98	6	2.97
c73.c75	MN Thyroid	12	3	0.36
c76.c80	MN 2nd Site	10	2	0.30
c81.c96	MN Lymphoid	89	6	2.70
cortisol_nmol_l	Cortisol	386	8	11.69
creat	Creatinine	692	8	20.96

crp	CRP	480	10	14.54
e	Metabolic/Endocrine	128	8	3.88
e00.e07	Thyroid	42	7	1.27
e10.e14	Diabetes	140	6	4.24
e15.e16	Other Glucose	28	5	0.85
e20.e35	Other Endocrine	5	4	0.15
e50.e64	Other Nutritional	18	4	0.55
e65.e68	Obesity	50	7	1.51
e70.e90	Metabolic	71	6	2.15
edu	EDU	65	4	1.97
fev1	FEV1	284	9	8.60
frs	FRS	945	10	28.62
hdl	HDL	1256	10	38.04
i	Vascular	27	5	0.818
i05.i09	Chronic HD	12	4	0.36
i10.i15	Hypertensive	76	7	2.30
i20.i25	IHD	15	6	0.45
i26.i28	Pulmonary	9	5	0.27
i30.i52	Other HD	31	5	0.94
i60.i69	Cerebrovascular	15	4	0.45
i70.i79	Arteries	14	3	0.42
i80.i89	Veins	23	4	0.70
i95.i99	Other Circulatory	9	4	0.27
j	Infectious	28	5	0.85
j00.j06	Acute Respiratory	33	6	1.00
j09.j18	Flu/Pneumonia	32	5	0.97
j20.j22	Acute LR	86	4	2.60
j30.j39	Upper R	12	5	0.36
j40.j47	Chronic LR	19	5	0.58
j60.j70	Lung	10	3	0.30
j80.j84	Other Respiratory Int	103	5	3.12
j85.j86	SN LR	18	4	0.55
j90.j94	Other Pleura	11	4	0.33
j95.j99	Other Respiratory	9	3	0.27
totchol	TC	1351	10	40.91

Table 15. Outcome Level Associations with Omics Biomarkers. *Outcome: disease block (ICD10 code chapter) or health-related risk factor. Description: description of disorders covered. N Associations: number of significant (5% FDR) biomarker-outcome associations. N Assays: number of omics assays that the outcome has significant (5% FDR) associations with. Percentage of tests significant: the percentage of significant (5% FDR) biomarker-outcome associations out of those tested for that outcome, calculated as (N Associations/(N tests))*100. FRS: Framingham risk score. BMI: body mass index. EDU: educational attainment. HDL: high density lipoprotein cholesterol. TC: total cholesterol. SBP: systolic blood pressure. FEV1: forced expiratory volume in 1 minute. CRP: c-reactive protein. HD: heart disease. CNS: central nervous system. IHD: ischaemic heart disease. MN: malignant neoplasm. LR: lower respiratory tract. Upper R: upper respiratory tract. SN LR: Suppurative & necrotic conditions of the lower respiratory tract. Other Respiratory Int: other respiratory diseases affecting the interstitium.*

As 76.6% of the significant biomarkers were associated with more than one outcome, I looked at the pattern of associations across outcomes (**Figure 12**). Total cholesterol, HDL cholesterol and BMI each had >1000 significantly associated omics biomarkers (**Table 15**, **Supplementary Table 24**). Diabetes Mellitus (E10-E14) and all metabolic/endocrine disorders combined (E) were the disease blocks with the most associations (140 and 128 respectively). Interestingly, all cancers (C) had the fewest associations (3) however, all of the disease blocks containing specific cancer subsets had >3 significant associations.

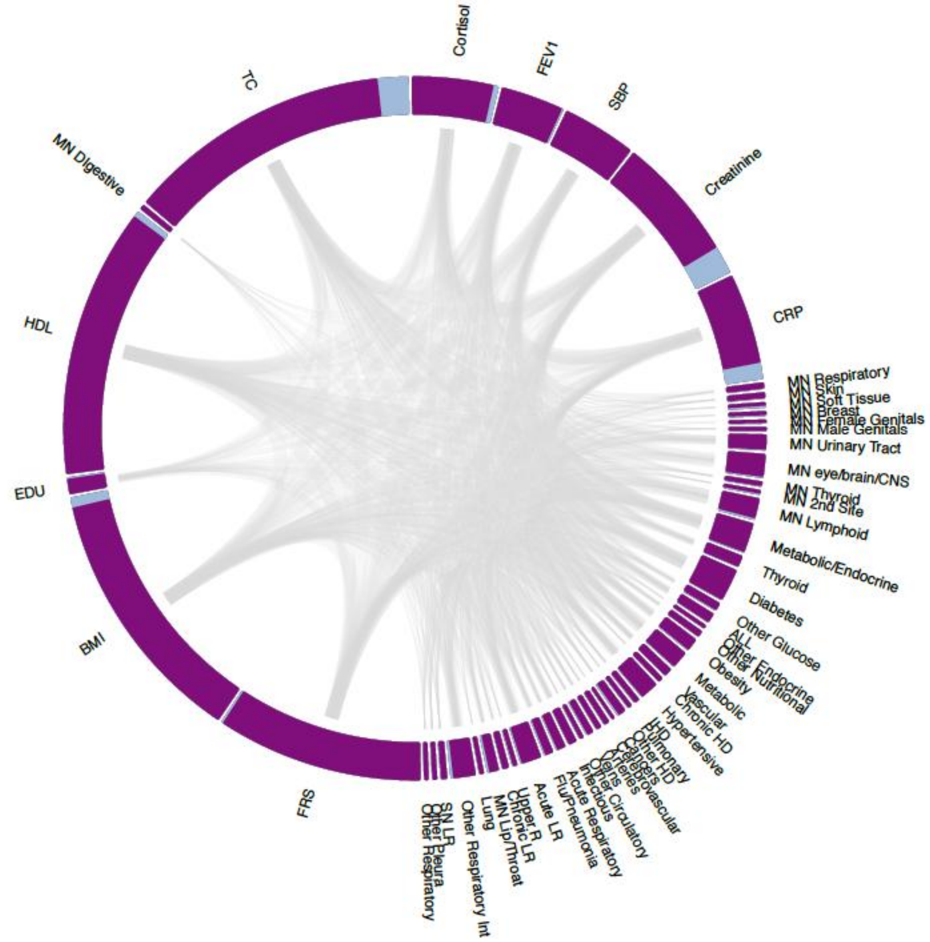


Figure 12. Connectivity between Risk Factors and Incident Disease Blocks based on Associated Omics Measures. *Each segment represents an outcome (risk factor or disease block). The size of the segment indicates the number of single omics biomarkers that the outcome was significantly (5% FDR) associated with. Each segment is split to show the number of associations with at least one other disease (purple) or biomarkers that were uniquely associated with that outcome (blue). Lines connecting two outcomes indicates that they are associated with shared omics biomarkers, with the thickness of the line depending on the number of biomarker associations they share. FRS: Framingham risk score. BMI: body mass index. EDU: educational attainment. HDL: high density lipoprotein cholesterol. TC: total cholesterol. SBP: systolic blood pressure. FEV1: forced expiratory volume in 1 minute. CRP: c-reactive protein. HD: heart disease. CNS: central nervous system. IHD: ischaemic heart disease. MN: malignant neoplasm. LR: lower respiratory tract. Upper R: upper respiratory tract. SN LR: Suppurative & necrotic conditions of the lower respiratory tract. Other Respiratory Int: other respiratory diseases affecting the interstitium.*

In terms of biomarker specificity, outcomes ranged from malignant neoplasms of mesothelial and soft tissue (C45-C49), with the largest percentage of disease specific associations (69.2%), to Malignant neoplasms of the digestive organs (C15-C26) with all significantly associated biomarkers also being associated with at least one other outcome. **Figure 12** highlights how interconnected different outcomes are, with the connections (in grey) indicating two outcomes were significantly associated with common biomarkers. The width of the connection denotes the number of biomarkers shared. HDL and total cholesterol shared the most biomarkers (891) with HDL and BMI also sharing >800. However, 221 outcome pairs share only 1 common biomarker. BMI and Creatinine levels had the most connections with other outcomes (50) with malignant neoplasms of female genital organs the having the least with 4 (**Supplementary Figure 43**). As all outcomes were significantly associated with >1 biomarker, I next assessed multivariable omics models.

4.3.3 Penalised Regression Models

Selection of Penalised Regression Method

To see which penalised regression approach was the most effective for creating omics prediction models for the outcomes, I compared LASSO and elastic net regression with an alpha of 0.5. For each of these two methods, I constructed models from 11 omics assays trained on 54 health related risk factors and incident disease blocks. Given the aim to create clinically useful prediction models, ridge regression was excluded as a potential approach as it does not produce sparse models, and models with hundreds of predictors are not suitable for my purposes. **Figure 13** highlights that LASSO was the more effective method, as the effect sizes from the regression of predicted outcome on observed outcome were consistent between the training and testing samples across the majority of outcomes. Elastic net regression, however, was extremely inconsistent across outcomes, suggesting possible overfitting to the training samples. Results for Clinomics and DEXA scores are shown here as examples, full results across omics are included in **Supplementary Figure 44,** however, the pattern of LASSO having consistent effect size and direction estimates across training and testing samples continued across all omics assays. Based on this evidence LASSO was the method taken forward and used in subsequent analysis.

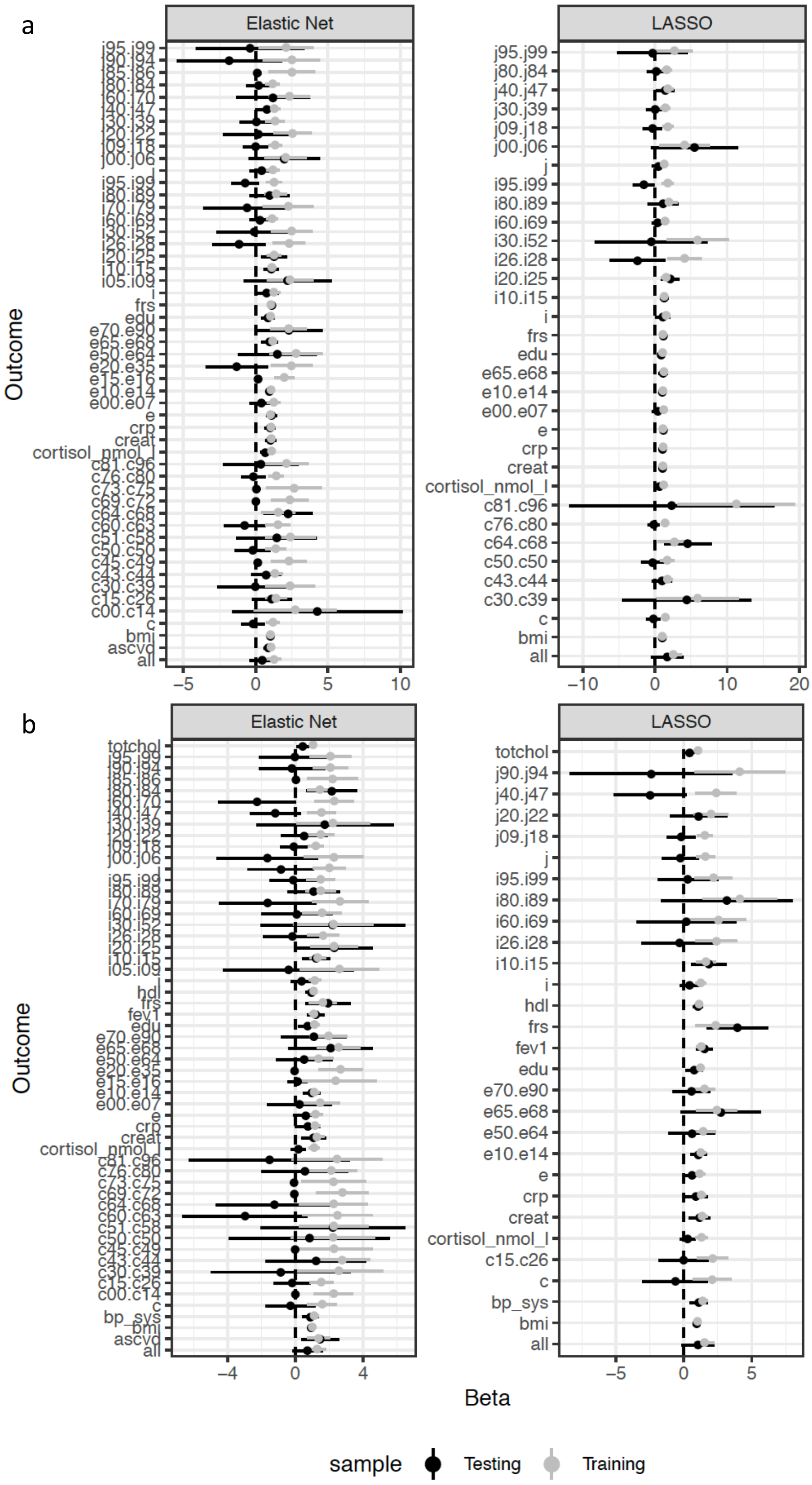


Figure 13. Comparison of LASSO and Elastic Net Regression. *Effect size and 95% confidence intervals from regression of outcome predicted by the model and observed outcome in training and testing samples. These estimates are across outcomes (y axis) and between methods (panels) for a) Clinomics and b) DEXA. Results for all 11 omics in* ***Supplementary Figure 44****.Not all 54 possible outcomes are listed for each omics assay-method pair. For a subset of pairs, using optimised lambda from the*

cross-validation in the training sample, the algorithm selected to include no predictors, as none outperformed the sample mean in predicting the Martingale residual of the outcome. This could be due to the limited number of cases for these disease blocks. Therefore, these pairs are not shown and were not taken forward.

Controlling Multiple Testing

Once LASSO was chosen as the penalised regression approach for the analysis, I considered the issue of multiple testing. A number of aspects of this study design make multiple testing a concern i) the limited sample sizes for each omics assay ii) that these small sample sizes will be reduced further during model construction when split into training and testing iii) the low number of cases in many of the disease blocks (mean number of cases for each disease block across omics assays in **Figure 14**, numbers of cases and controls across all omics-disease blocks in **Supplementary Table 16**) and iv) the number of outcomes considered in our exploratory analysis.

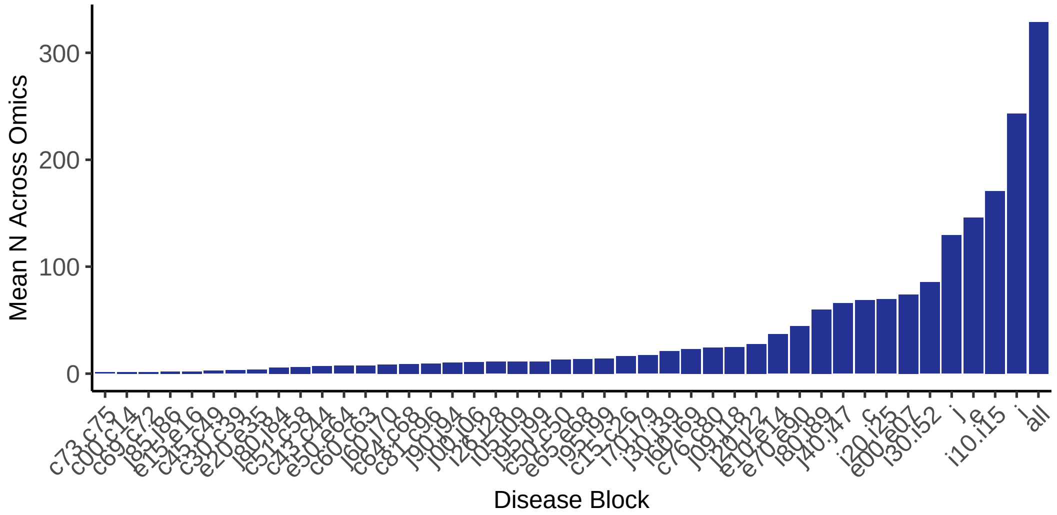


Figure 14. Mean Number of Disease Cases Across Omics. *Mean number of cases available across the 11 omics assays for each of the 44 disease blocks.*

In order to maximise the robust statistically significant results possible with my data, I limited the number of formal statistical tests performed. Only tests that were expected to pass 5% FDR in the testing sample, based on observed effect sizes in the training sample, were taken forward for formal testing (See Methods for details).

Testing was limited to 82 tests of omics scores trained on risk factors and 144 for omics scores trained on Martingale residuals for incident diseases.

Omics Models Associated with Risk factors and Incident Disease

The number of predictors selected for model inclusion for each omics assay-outcome pair are shown in **Supplementary Table 23**. 69 omics scores were significantly (5% FDR) associated with health-related risk factors and 12 with disease blocks (**Figure 15**, Training and Testing comparison in **Figure 16** and **Figure 17**).

First, assessing the effectiveness of omics scores across outcomes: unsurprisingly, Clinomics scores were significantly associated across multiple outcomes, as measures included in the Clinomics assay were selected for their clinical use to indicate health status and prognosis. Further, the only outcomes in **Figure 15** that Clinomics scores were not associated with (grey cross on white background), were those that were not run to avoid circularity as Total cholesterol, systolic blood pressure, HDL cholesterol and FEV1 were predictors available for selection in the Clinomics assay. Interestingly, UPLC IgG Glycomics score was significantly associated with all 10 risk factors but with low effect sizes. Mega-omics scores were associated with BMI, plasma Cortisol, Creatinine and CRP levels with standardised effect sizes 0.999 (0.02), 0.89 (0.04), 0.93 (0.07) and 0.73 (0.07) respectively.

Under the perpendicular view, looking at risk factors across omics scores: BMI was associated with 9 different omics scores, with Mega-omics, Clinomics, PEA Proteomics and DEXA scores having standardised effect sizes >0.88. This effect size of 0.9 of the PEA Proteomics score on BMI is of particular interest as, unlike DEXA and Clinomics, protein levels are not *a priori* BMI related. The Mega-omics score's effect size of 0.998 on BMI contained only three biomarkers however, they were height, weight and the level of FBP1 (Fructose-Bisphosphatase 1) protein that is involved in glucose metabolism. Total and HDL cholesterol were each associated with the same 7 omics scores, with larger effect sizes evident from scores that were built from predominantly lipid-based omics assays.

Several other large effect sizes were found, the association of Creatinine levels and Mega-omics (Beta: 0.92, SE: 0.07) and CRP levels with both PEA Proteomics (Beta: 0.74, SE: 0.05) and MS Complex Lipidomics (Beta: 0.76, SE: 0.06) scores.

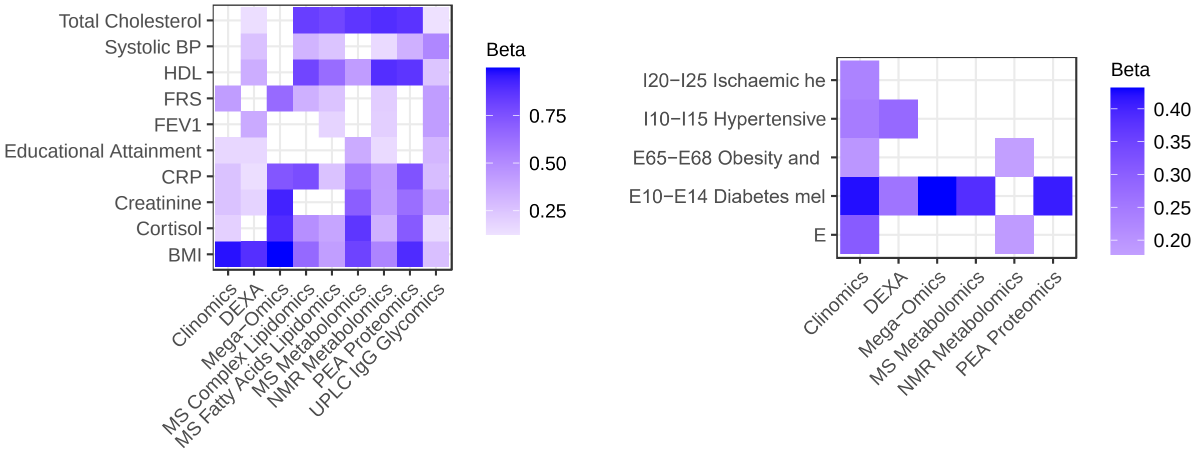


Figure 15. Omics Scores Predict Risk Factors and Subsequent Incident Disease. *Beta: the standardised effect size (both score and outcome in standard deviation units) of the omics scores (x axis) on the outcomes on the y axis. Results for health-related risk factors on the left and disease blocks on the right. Only results that passed 5% FDR significance threshold are shown. FRS: Framingham Risk Score. E10.E14: Diabetes mellitus. E: All block E metabolism related disorders. I10.I15: Hypertensive diseases. E65.E68: Obesity and other hyperalimentation. I20.I25: Ischaemic heart diseases.*

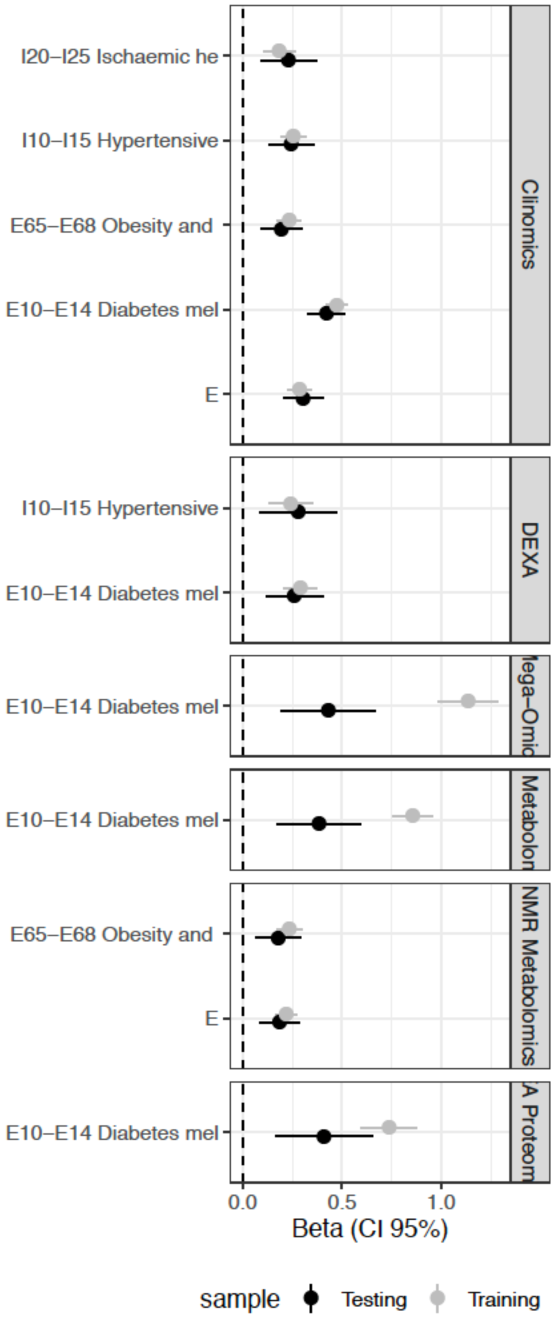
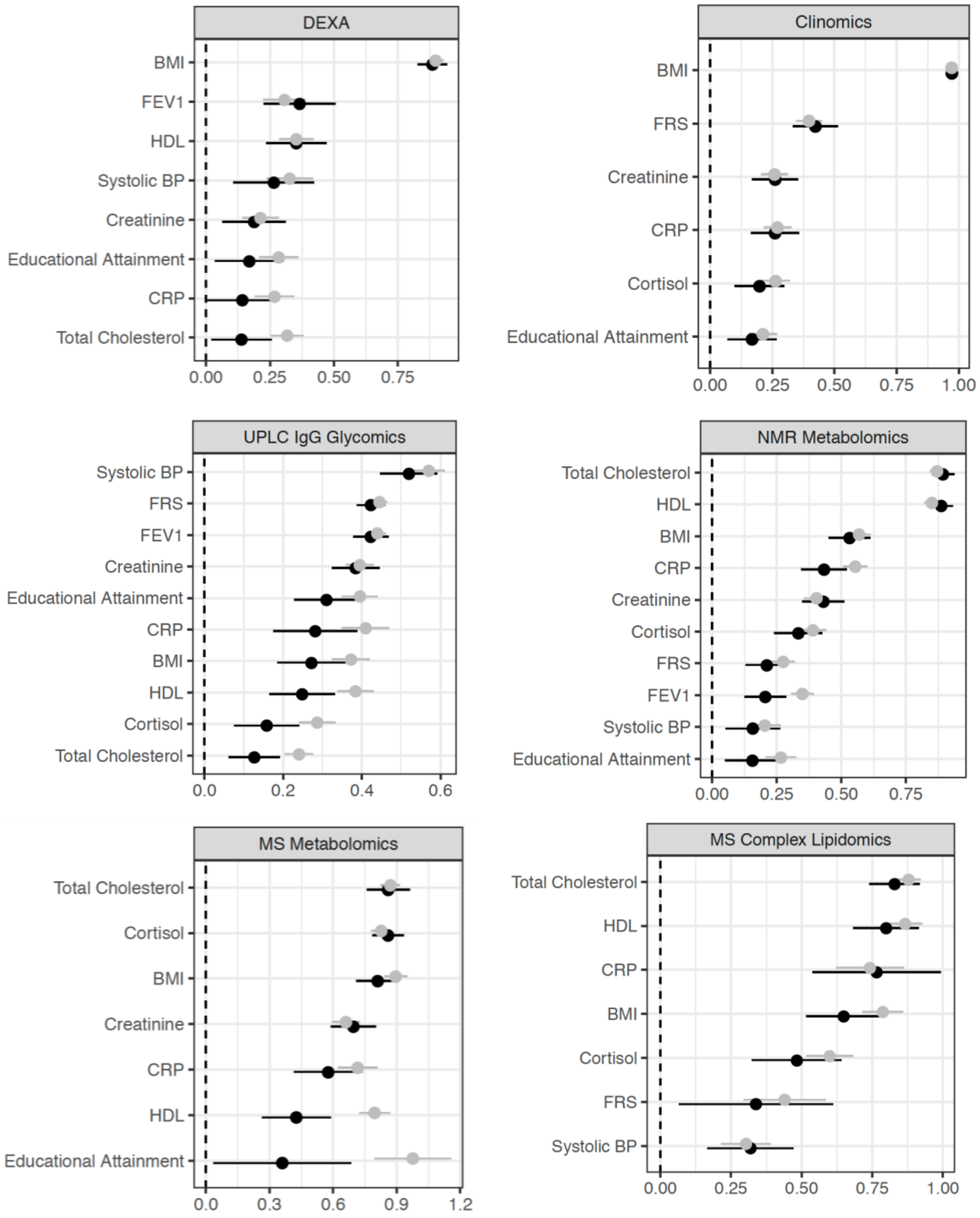


Figure 16. Significant Omics Scores Predict Incident Disease in Training and Testing Samples. *Showing the standardised effect size (both score and outcome in units of phenotypic standard deviation) of the omics scores (panels) on the outcomes on the y axis. Only results that passed 5% FDR significance threshold are shown. E10.E14: Diabetes mellitus. E: All block E metabolism related disorders. I10.I15: Hypertensive diseases. E65.E68: Obesity and other hyperalimentation. I20.I25: Ischaemic heart diseases.*


DEXA
BMI
FEV1
HDL
Systolic BP
Creatinine
Educational Attainment
CRP
Total Cholesterol
0.00
0.25
0.50
0.75
Clinomics
BMI
FRS
Creatinine
CRP
Cortisol
Educational Attainment
0.00
0.25
0.50
0.75
1.00
UPLC IgG Glycomics
Systolic BP
FRS
FEV1
Creatinine
Educational Attainment
CRP
BMI
HDL
Cortisol
Total Cholesterol
0.0
0.2
0.4
0.6
NMR Metabolomics
Total Cholesterol
HDL
BMI
CRP
Creatinine
Cortisol
FRS
FEV1
Systolic BP
Educational Attainment
0.00
0.25
0.50
0.75
MS Metabolomics
Total Cholesterol
Cortisol
BMI
Creatinine
CRP
HDL
Educational Attainment
0.0
0.3
0.6
0.9
1.2
MS Complex Lipidomics
Total Cholesterol
HDL
CRP
BMI
Cortisol
FRS
Systolic BP
0.00
0.25
0.50
0.75
1.00

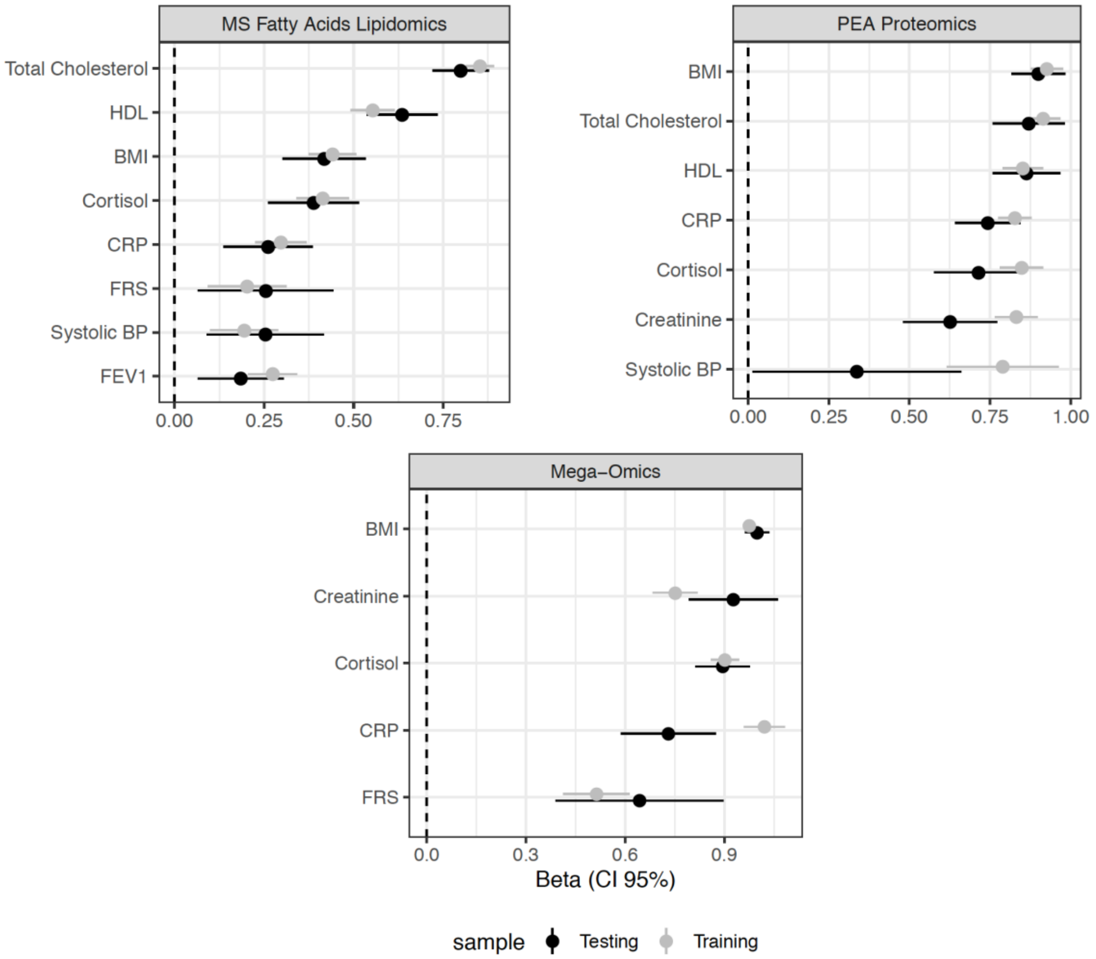


Figure 17. Significant Omics Scores Predict Risk Factors in Training and Testing Samples. *Showing the standardised effect size (both score and outcome in units of phenotypic standard deviation) of the omics scores (panels) on the outcomes on the y axis. Only results that passed 5% FDR significance threshold are shown. FRS: Framingham Risk Score. E10.E14: Diabetes mellitus. E: All block E metabolism related disorders. I10.I15: Hypertensive diseases. E65.E68: Obesity and other hyperalimentation. I20.I25: Ischaemic heart diseases.*

There is a distinct L-shape pattern of disease block results (**Figure 15**): Clinomics scores show significant associations with the most (5) different disease blocks and Diabetes Mellitus (E10-E14) was associated with 6 different omics scores. The reasonable sample size for diabetes (N Average cases across omics = 36.9), including both types I and II, compared to other disease blocks could contribute to the performance of the score. The Mega-omics score for Diabetes mellitus had the largest effect size 0.43 (SE: 0.12), with PEA Proteomics and Clinomics scores also with standardised effect sizes >0.4.

In general, omics scores for health-related risk factors were more effective (based on standardised effect sizes) than those trained on incident disease blocks. The lower sample size and in particular low number of incident cases for disease blocks could contribute to the difference in performance observed.

To investigate patterns of biomarker specificity, I looked at the number of different outcomes for which each biomarker was selected for score inclusion. 52.7% were outcome specific, with 83.5% selected for inclusion in scores for <4 outcomes (**Figure 18**). In contrast, glucose and albumin were selected for inclusion in scores for 19 different outcomes, with glucose having a positive coefficient in scores for 16 different outcomes and 3 negative, with albumin being included in 11 scores with a positive effect and 8 negative. Three biomarkers were selected for inclusion in scores for 18 different outcomes: height (11+ve and 7-ve), the essential amino acid histidine (6+ve and 12-ve) and the ratio of monounsaturated fatty acids to total fatty acids (MUFA/FA) (15+ve and 3-ve).

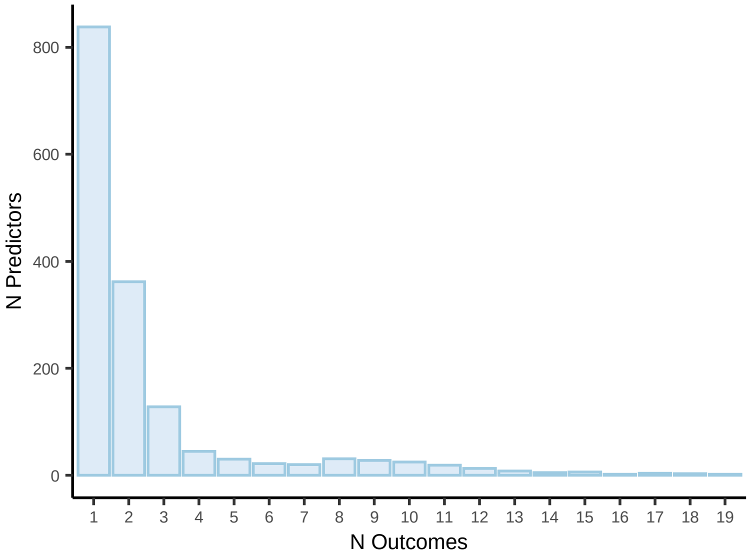


Figure 18. Omics Predictors Selected for Inclusion in Scores for Multiple Outcomes. *The frequency of predictors (y axis) selected for inclusion in scores for the number of outcomes indicated on the x axis.*

4.3.4 Score Profiling

For the omics scores that were most prognostic of incident disease, I wanted to investigate which of the included biomarkers were contributing most to the performance of the score. I limited this analysis to the L-shape of significant omics scores highlighted in the previous section (**Figure 15**). That is the 5 Clinomics scores and 4 scores for E10-E14 Diabetes mellitus. For each score I partitioned the variance explained in the score amongst the included biomarkers, creating score profiles.

I used two different methods for variance partitioning: hierarchical partitioning and iterative ranking. As the hierarchical partitioning approach implemented by the hier.part R package is designed for models with 10 predictors or less, an additional method termed iterative ranking (See Methods for Details) was used, allowing comparison between methods for scores including <10 biomarkers.

I was also interested to see how the profile of my Clinomics score for FRS compared to that of the Framingham risk score itself (FRS Clin). As the FRS Clin is not a linear combination of its components (**Equation 1**), I constructed a score that was a linear combination of the predictors in my Clinomics assay that overlapped with those in FRS Clin. However, as the multivariate R^2 from fitting this score on FRS Clin is less than 1, it is this total that I then partitioned, meaning that in **Figure 19** segments are the proportions of this multivariate R^2 that were additionally explained by the addition of each predictor in the score, rather than the absolute additional R^2, as is the case with the other Clinomics scores, as their total multivariate R^2 was 1. The profile for my Clinomics score for FRS is almost identical to the profile of FRS Clin, with systolic blood pressure dominating and with total and HDL cholesterol also contributing to the variance explained in the scores.

Encouragingly, profiles created using the two methods were comparable for Clinomics scores across outcomes (**Figure 19**). The scores for four outcomes were dominated by one predictor. The Clinomics score for Diabetes mellitus (E10-E14) was 96.2% explained by glucose levels. Weight is contributing 95% of the Clinomics score for obesity (E65-E68). Glucose levels accounted for 98.4% of the Clinomics score for Ischaemic heart disease (I20-I25) and SBP accounted for 79.5% of the Clinomics prediction of FRS. Glucose also appears to be contributing the majority of the variance explained in the score for metabolic disorders (E). This is likely due to E being predominantly cases of diabetes and therefore glucose driving the score. Unsurprisingly, weight is contributing 95% of the Clinomics score for obesity (E65-E68), with height and glucose levels making small contributions. The Clinomics score for hypertensive diseases (I10-I15) was distributed amongst more biomarkers than other Clinomics scores, systolic blood pressure having the largest contribution (53.7%) however, weight, glucose, uric acid and LDL cholesterol also contributed.

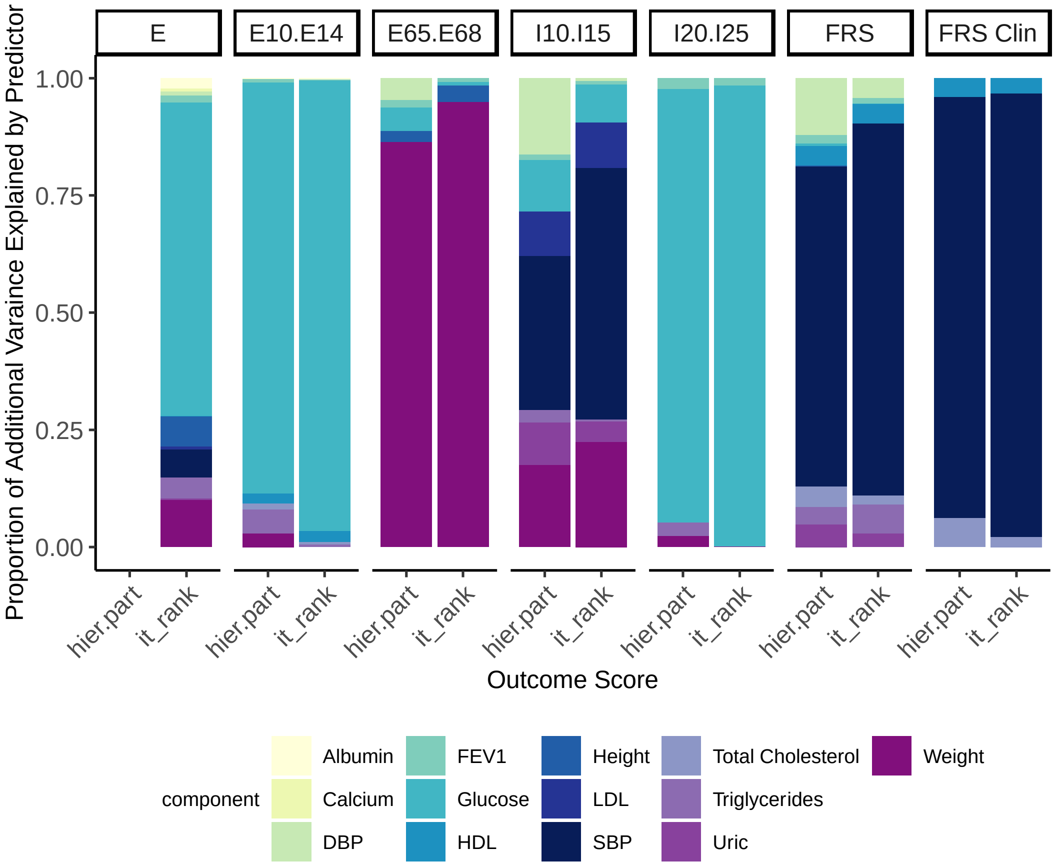


Figure 19. Clinomics Score Profiles. *Indicating the proportion of variance explained in the score by each of the predictors in the score. Only Clinomics scores for outcomes that passed 5% FDR significance threshold are shown. For each outcome (panel), the estimates of variance explained by each component from two different methods are shown (x axis) hier.part: hierarchical partitioning and it_rank: iterative ranking (See Methods for details). Only results from the iterative ranking method are shown for the disease block E score, as it is a linear combination of >10 predictors. FRS: Framingham Risk Score. FRS Clin: a score that is a linear combination of the components of the clinically used Framingham risk score (**Equation 1**) that overlap with the*

Clinomics set of predictors. E10.E14: Diabetes mellitus. E: All block E metabolism related disorders. I10.I15: Hypertensive diseases. E65.E68: Obesity and other hyperalimentation. I20.I25: Ischaemic heart diseases.

Interestingly, we found a stark contrast between the profiles for the Clinomics score for subsequent ischaemic heart disease (I20-I25) and FRS, given that FRS is prognostic of incident cardiac outcomes, one may therefore expect these scores to have similar profiles. This could be due in part to the differences in sample size, there were N=97 incident cases of I20-I25 with the Clinomics assay and 1669 individuals with FRS values. However, both the Clinomics score for I20-25 and the FRS in clinical use had comparable standardised effects on subsequent incident ischaemic heart disease (**Figure 20**).

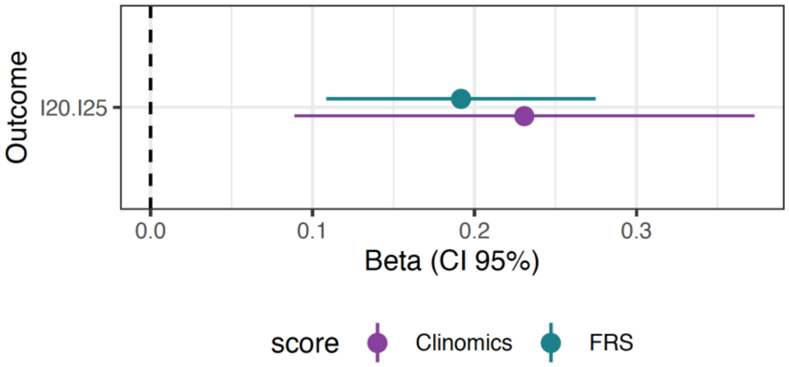
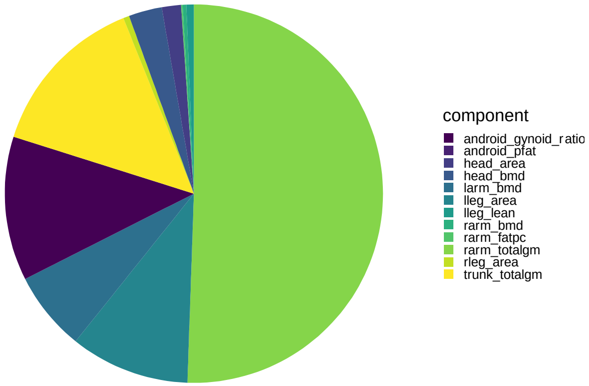


Figure 20. Clinomics score and FRS show comparable standardised effect on Ischaemic Heart Disease. *Beta: standardised effect size (both score and outcome in standard deviation units) and 95% confidence intervals of the Clinomics score for Ischaemic heart disease (I20-I25) and FRS on Ischaemic heart disease.*

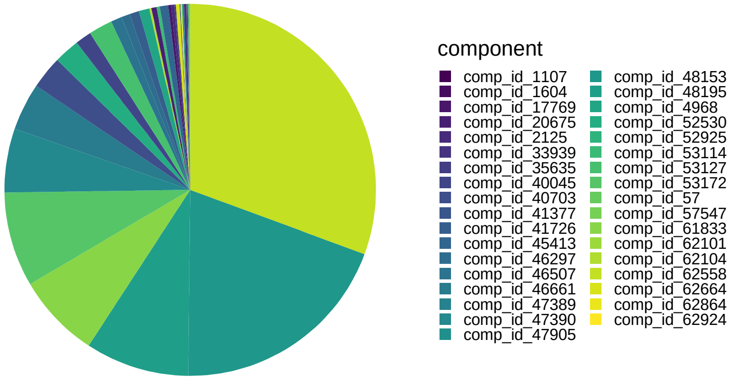
A distinct pattern was found across the 4 significant omics score profiles for Diabetes Mellitus (**Figure 21**). Namely, that 4 or fewer biomarkers were contributing greater than two thirds of the variance explained in the scores, independent of the number of included biomarkers. This suggests that scores containing only a few biomarkers could be prognostic of incident disease. The circulating plasma protein levels of Mevalonate kinase (MVK), carboxylesterase 1 (CES1) and adhesion G protein-coupled receptor 1 (ADGRG1) are shown to be contributing large portions of the variance explained in the PEA Proteomics score. Interestingly, these three protein levels are amongst the 5 biomarkers contributing the greatest proportion of variance explained in the Mega omics score alongside the levels of Matrilin 3 (MATN3) protein and a metabolite level from the MS Metabolomics platform. These proteins are involved in sterol synthesis (MVK), drug metabolism (CES1), brain cortical patterning (ADGRG1) and development and homeostasis of bone and cartilage (MATN3).

As glucose contributed 96.2% of the Clinomics score for Diabetes mellitus (E10-E14), it was notable by its absence from the profile of the Mega-omics for diabetes. However, this was due to circulating glucose levels being absent from the final Mega-omics dataset for diabetes, having been removed during the quality control process of creating a complete non-missing dataset.

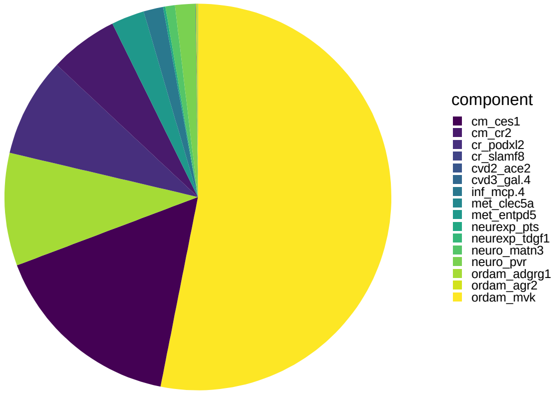
DEXA
(total mass right arm | total mass trunk | ratio of android-gynoid fat percentage)



MS Metabolomics
(CompID 62558 | CompID 48153 | CompID 48195)



PEA Proteomics
(MVK | CES1 | ADGRG1)



Mega Omics
(ADGRG1 | CompID 48195 | CES1)

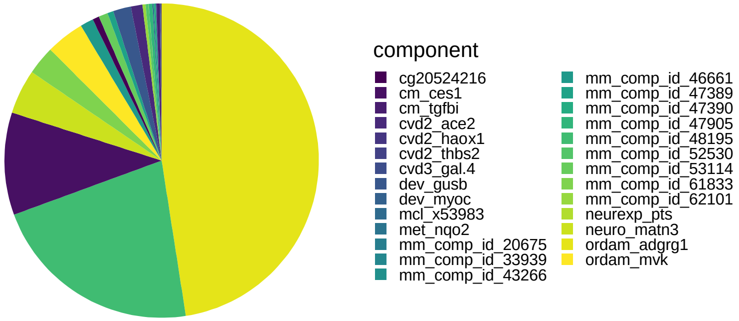


Figure 21. Score Profiles for Diabetes Mellitus. *Indicating the proportion of variance explained in the score by each of the predictors in the score. Only scores for Diabetes Mellitus that passed 5% FDR significance threshold are shown. For each omics score, the estimates of variance explained by each component from the it_rank: iterative ranking (See Methods for details) as all 4 scores are linear combinations of >10 predictors. The three predictors with the largest contribution to the variance explained in the score are indicated under the assay name.*

4.4 Discussion

I found that 2,686 single omics biomarkers were significantly associated with 10 health-related risk factors and 44 incident disease blocks. 23.4% of significant biomarkers had outcome specific associations, with 76.6% associated with multiple outcomes. I showed that multivariable Clinomics scores were prognostic of incident Diabetes mellitus, obesity, hypertensive disorders and ischaemic heart disease and that DEXA, MS Metabolomics, PEA Proteomics and Mega omics scores were prognostic of subsequent incident Diabetes mellitus. 9 omics scores were significantly associated with BMI, and Clinomics and UPLC IgG Glycomics scores were significantly associated with the most health-related risk factors. By creating profiles for the significant omics scores, I highlight that only a handful of biomarkers are contributing the majority of the variance explained in the score, suggesting that it would be possible in the future to construct scores with even fewer biomarkers. This analysis also indicated MVK, CES1, ADGRG1 and MATN3 plasma protein levels as potential biomarkers for diabetes mellitus.

Pietzner *et al.*'s finding that 65.6% of significant metabolites associated with at least 2 different diseases is comparable to the 67.9% (375/552) I reported. My results suggest that this pattern, of more biomarkers being associated with multiple outcomes rather than being outcome specific, holds across omics assays with 76.6% of significant biomarkers associated with more than one outcome. Pietzner *et al*. reported that the outcome with the most metabolite associations was all-cause mortality, out of all-cause mortality and 27 incident noncommunicable diseases. Due to the low number of deaths in the ORCADES cohort, I was unable to assess mortality as an outcome but found that Diabetes mellitus was the disease block with the most significant biomarker associations (140), with Total cholesterol, HDL cholesterol and BMI being the health-related risk factors with the most associations (**Supplementary Table 24**).

A consistent pattern was observed across results from both single omics biomarker-outcome associations and multivariable omics score-outcome associations: that superior performance, either in terms of number of significant outcome associations or larger standardised effect size on outcome, were observed for more specific outcomes, compared to more broad outcomes. For single biomarker-outcome associations, I found that all cancers (C) had fewest associations but disease blocks that contained specific cancer subtypes had >3 significant (5% FDR) associations. Conceptually this makes sense, as the heterogeneity of phenotypes and therefore biological pathways involved across all types of cancer will limit the number of single omics biomarker associations. Whereas specific types of cancer will have a more homogeneous omics profile, allowing significant associations to be discovered. Similarly, the Clinomics score for diabetes had a larger effect than the Clinomics score for all metabolic/endocrine disorders (E). This mirrors the results found in **Supplementary Table 24**, and again suggests that a score trained on an outcome that is capturing a broad range of heterogeneous phenotypes, such as all of block E disorders, despite its larger number of cases, will be less effective than one trained on a more specific outcome (e.g. diabetes).

In order to limit multiple testing of omics score-outcome associations, I estimated that 82 risk factor- and 144 disease block-omics score associations were likely to pass a 5% FDR significance threshold (See Methods for Details). For risk factors, 84.1% of those estimated passed the significance threshold in the testing sample however, for disease blocks, only 8.3% of the expected associations were significant in the testing sample. The conservative use of two-sided p-values given our issue with power, with disease blocks having considerably lower sample sizes (**Figure 14**, **Supplementary Table 16**), could have contributed to this considerable difference between risk factors and disease blocks. Additionally, the bimodal distribution of Martingale residuals compared to the risk factors, which more closely approximated normal distributions could have had an impact however, the comparison of effect sizes between training and testing samples were predominantly comparable (**Supplementary Figure 44**).

I observed that Clinomics, both in regard to single biomarkers and multivariable scores, showed the most significant outcome associations. This mirrors the results in **Chapter 3: Biological Ageing Clocks**, where Clinomics OCAA showed most significant health outcome associations compared to those built from other omics assays. Again, this is unsurprising given that the predictors that comprise the Clinomics assay were chosen for their common clinical use as prognostic indicators of adverse health outcomes. The effectiveness of these common traditional risk factors is why they are often the basis of risk scores in clinical use[101,230], therefore my results support their utility.

Interestingly, we found that very few biomarkers contribute the majority of the predictive power of our significant omics risk scores. For, DEXA, MS Metabolomics, PEA Proteomics and Mega-omics scores for Diabetes mellitus $\leq$4 biomarkers contributed >66% of the variance explained in the score. For the Clinomics scores for obesity, diabetes mellitus, FRS, hypertensive diseases and ischaemic heart disease only a single predictor dominated the scores, providing >79.5% of the variance. This is a major finding as it suggests that effective risk scores prognostic of subsequent incident disease can be made using only a few predictors. Not only is this desirable due to being less invasive for patients, but it is also likely to be more cost effective, to be able to provide personalised advice to patients based on less than four or even one measure.

I showed that 3 protein levels, MVK, CES1 and ADGRG1, dominated the PEA Proteomics score for diabetes, and with the addition of MATN3, also contributed 63.4% of the variance explained in the Mega-omics score for diabetes. This domination of Mega-omics by proteins was also observed in the composition of the Mega-omics ageing clocks in **Chapter 3: Biological Ageing Clocks**, with 26.6% of omics predictors selected for model inclusion coming from the PEA Proteomics assay.

These four protein levels are potential novel biomarkers of subsequent incident diabetes mellitus. While each has been previously implicated as associated with related traits, through significant SNPs from GWAS being mapped to their encoding genes, none has been reported associated with either type 1 or 2 diabetes. Mevalonate kinase (MVK) is involved in cholesterol synthesis[236] and has been previously associated with weight[237], height[238], BMI[239], HDL cholesterol[236,240], various fat mass percentages[237], SBP and coronary artery disease (CAD)[241]. Carboxylesterase 1 (CES1) is involved in drug metabolism and has been reported associated with LDL cholesterol[242] and DBP[243]. Adhesion G Protein-Coupled Receptor G1 (ADGRG1), involved in brain cortical patterning, has been significantly associated with HDL cholesterol[240], systemic lupus erythematosus (SLE)[244] and prostate cancer[245]. Matrilin 3 (MATN3), contributing to the extracellular matrix, is involved in development and homeostasis of bone and cartilage has been associated with height[238], LDL cholesterol[242], HDL cholesterol[240], triglyceride levels[240], DBP[243], CAD[246], MI, WHR and BMI. The fact that the levels of these proteins dominate the variance explained in risk scores for diabetes mellitus suggests that they should be further investigated as potential novel biomarkers for diabetes.

I found that 18 single UPLC IgG Glycomics measures were significantly associated with FRS, 6 of these glycans (GP6, GP14, GP6n, FBS1/FS1, FBStotal/FStotal and FBS1/(FS1+FBS1)) have been shown to be associated with ASCVD, the 10-year atherosclerotic cardiovascular risk score[230], by Menni *et al.*[108]. All six had consistent effect directions between associations with FRS and ASCVD. Additionally, it is this UPLC IgG Glycomics data from ORCADES that provided the replication for Menni *et al.*'s associations with ASCVD, so by definition the 10 glycans that they reported associated with ASCVD in TwinsUK replicated in ORCADES.

A key finding was that glucose explained 98.4% of the variance in the Clinomics score for subsequent incident ischaemic heart disease. This result could be spurious, as despite the score being significantly associated with incident ischaemic heart disease it only explained 2.31% of variance in the outcome (**Supplementary Figure 45**). This is considerably lower than the R^2 for the other significant Clinomics scores but does not appear to be due to lack of power as there were more cases of ischaemic heart disease (N=97 with Clinomics) than of Diabetes mellitus (N=49 with Clinomics). Moreover, glucose was not a large contributor to the profile of FRS, given that these two scores would be expected to capture similar underlying biology, as are prognostic of overlapping outcomes, this was surprising. However, the consistent effect size of the Clinomics score on incident ischaemic heart disease between training and testing samples (**Figure 16**) indicates that overfitting was not the issue, and that the association is likely to be valid. I also took the additional step of comparing the standardised effect sizes of the Clinomics score for ischaemic heart disease and the FRS in clinical use on subsequent incident ischaemic heart disease, demonstrating their comparable predictive ability (**Figure 20**). Overall my results point to glucose being a significant biomarker for ischaemic heart disease. This is further supported by the literature as glucose levels and glucose metabolism have been shown to be risk factors for cardiovascular disease even below diabetic levels[247], and implicated in tissue remodelling in the heart in ischaemic heart disease patients[248] respectively. Together this suggests that glucose has the potential to be a significant biomarker of ischaemic heart disease and future work should seek to investigate the integration of glucose into clinical cardiovascular risk scores.

A strength of this analysis was the sheer number of biomarkers assessed across a wide range of different omics assays, previous studies investigating omics biomarkers of disease have been limited to one high-dimensional assay or platform[108–111]. However, this aspect of the study, when combined with the small sample size and in particular, low numbers of cases for some disease blocks, limited the power to detect associations and to retain those we did find due to correction for multiple testing. For example, our finding that 616 single omics biomarkers showed no significant association with any of the 54 different health outcomes assessed, could be in part due to this lack of power. Being aware that this would be a limitation when designing the analysis, I took steps to minimise the number of formal statistical tests by only performing those powered to produce significant associations.

In addition to constructing omics scores for health outcomes, I took a step further in creating score profiles and investigated which biomarkers included were contributing most to the scores. As mentioned in the Methods section, I was unable to use square standardised effect size estimates to calculate the variance explained in the score by each included biomarker, due to the measures within omics assays not being orthogonal. To circumvent this issue, I used an established hierarchical partitioning approach[235], unfortunately I was only able to apply this method to scores that contained <10 biomarkers and used an iterative ranking approach on scores with 10 or more biomarkers. This approach is not being presented as the definitive partition of variance explained amongst predictors, as by only fitting the predictors in one order to calculate the additional multivariate R^2 this is not possible. However, as an important aim was to highlight predictors which are driving the predictive power of the score, by producing the largest contribution to the variance explained. Selecting an order to fit the predictors based on an iterative ranking of their univariate R^2, will produce an estimate of the partition of variance in which the largest contributors will predominate. This was sufficient given our interest in the few biomarkers with the largest predictive contribution however, to comment on the underlying biology or causation a more balanced approach such as hierarchical partitioning would have been more suitable. It is therefore with this caveat that I present profiles calculated using this method.

For the analysis is this chapter I restricted the omics measures to those included in the biological ageing clocks chapter, meaning that for DNA methylation, only CpG sites that were included in the Hannum and Horvath epigenetic ageing clocks were assessed[55,66]. While for the analysis in **Chapter 3: Biological Ageing Clocks**, as extensive work had been done previously to find CpG sites that track biological and chronological age, it was deemed unnecessary to seek to repeat this process when the likelihood of improving on the results was poor considering my much smaller sample size. As the aim for this chapter was to use the methods and pipelines already in place in an exploratory analysis to investigate whether this range of omics measures were biomarkers of, or could be used to create scores for, disease, I did not include all 850,000 CpG sites for two reasons. First, it would have required fundamental structural changes to the pipeline and its implementation to allow for 850,000 predictors. Second, for the single omics biomarker-outcome associations, it would have massively exacerbated the issue of multiple testing and would have reduced the power to detect associations that were not false positives. However, future research dedicated to unravelling the relationship between DNA methylation and incident disease would be extremely worthwhile. Particularly given the known link between DNA methylation pattern and environmental factors that are themselves risk factors for disease such as smoking[32,249,250].

The use of first hospital admission for a disease block defined by multiple ICD10 codes over a 10-year follow up period as a proxy for subsequent incident disease was a limitation of this analysis and is discussed in detail in **Chapter 6: Discussion**.

Unlike several previous studies that identified omics biomarkers of incident disease[108–111,114], I used Martingale residuals for incident disease blocks as outcomes and performed linear associations, rather than fitting full Cox proportional hazard models. This approach has been used in GWAS for time-to-event outcomes such as parental lifespan[232,251,252]. Cox proportional hazard models with fixed effects covariates were fitted, Martingale residuals calculated and then used as the quantitative trait for linear SNP associations. Additionally, the existing pipeline for constructing omics scores using penalised regression techniques was optimised for continuous variables as outcomes. I tested the suitability of this approach for this analysis, and showed that effect size estimates of OCAA on incident disease were consistent using Martingale residuals and fitting Cox proportional hazard models (**Supplementary Figure 41** and **Table 13**). A consequence of choosing this approach, however, was applying penalised regression techniques that assume normality of both predictors and outcome to bimodally distributed Martingale residuals. As Martingale residuals vary between 1 and $-\infty$ (individuals with an event in the study period) and 0 and $-\infty$ (individuals without an event in the study period), they are therefore bimodally rather than normally distributed. The central limit theorem states that, with increasing sample size any variable will tend towards a normal distribution, but the small sample sizes in this analysis means that this is unlikely to apply. This issue could have contributed towards: the failure of elastic net regression to produce consistent estimated effect directions between training and testing samples, several omics assay-outcome pairs LASSO being unable to construct models and in general for omics scores explaining small proportions of variance in outcomes. But the declared results are valid due to the rigorous training testing split.

A potential additional verification step to increase confidence in the score-outcome association results, would be to cross-check if fitting Cox proportional hazard models produce the same estimated effect sizes as those obtained using linear models and Martingale residuals.

An alternative approach would be to repeat the analysis using regularised Cox regression to build multivariable omics scores[201].

4.5 Conclusion

I have shown with exploratory analysis that these omics assays are useful sources of potential biomarkers for numerous outcomes and highlighted that <4 biomarkers, in some cases one biomarker, can dominate the predictive power of multivariable omics risk scores for incident disease. This analysis reported significant biomarker-outcome associations, omics scores prognostic of incident diabetes, obesity, hypertensive disorders and ischaemic heart disease and the importance of biomarkers selected for score inclusion. I also highlighted glucose as a biomarker of incident ischaemic heart disease and the levels of MVK, CES1, ADGRG1 and MATN3 as novel biomarkers of subsequent incident diabetes mellitus.

Despite these discoveries, this analysis has been held back due to sample size. This was a limitation on multiple levels: i) the limited number of samples with omics assays ii) the low number of cases in several disease blocks (<5) iii) that these extremely low case numbers were further reduced splitting the data into training and testing (3:1) and iv) that the number of cases were reduced again during 10-fold cross-validation in the training sample. I suggest that future work should seek to replicate my findings and repeat this analysis with a larger sample size, both in terms of individuals with omics biomarkers and cases of incident disease. Future work should also investigate if there are potential advantages in using penalised Cox proportional hazard models given the Martingale residual distribution issue discussed above. Overall, this analysis has demonstrated the potential of these omics assays as disease biomarkers.

Chapter 5: Genome-wide Association Meta-analysis of 184 Plasma Protein Levels

5.1 Introduction

5.1.1 Context

So far, this thesis has focussed on investigating how a broad range of omics measures from multiple assays have performed as potential biomarkers of biological ageing, health-related risk factors and subsequent incident disease. However, in this chapter I take a narrower view, focussing on only one of the omics assays available in ORCADES, specifically proteomics, to take a more detailed look at the biology underpinning protein measures and how they relate to health and disease.

The decision was made to focus on proteomics as proteins that circulate in the plasma are potential druggable targets. Many approved drugs target circulating proteins[37–39,128–130] and recently genetic studies, using Mendelian randomisation, have inferred causal relationships between protein levels and disease[116,149,153,178] and prioritised several proteins as potential novel drug targets[116,149].

The techniques that leverage genetic information such as genome-wide association, colocalisation, genetic correlation and Mendelian randomisation that facilitate investigations of this type, however, rely on sample size to power discoveries[157,160,163,184,253]. If limited to only ORCADES and Croatia-Vis, where I had access to both individual level genetic and proteomics data, the combined sample size would have been ~2,000 and I would have most likely only been able to detect extremely strong association signals in close proximity to the gene encoding the protein (*cis*-signals) and little else. I therefore sought to increase the sample size, as previous genetic studies of plasma protein levels have shown that increased sample size increased: the total number of protein quantitative trait loci (pQTL) discovered, the number of *trans*-pQTL detected as well as providing more power for downstream analyses.

As ORCADES had already contributed to a previous genome-wide association meta-analysis (GWAMA) lead by the SCALLOP Consortium[149], this existing collaboration provided a framework to organise a GWAMA of the 184 plasma proteins whose levels were measured on the cardiovascular II and cardiovascular III Olink panels. These two panels were chosen as, at time of data collection, these were the panels available in the most collaborating cohorts, unsurprising given the importance of cardiovascular disease to global health[11].

With 16 cohorts from the SCALLOP Consortium contributing GWAS summary statistics, in addition to ORCADES and Croatia-Vis, I had a maximum sample size of N=26,494. This is larger than those used in previously published GWAMAs of plasma protein levels[115,116,120,149,153,254,255], giving my analysis the potential to build on previous discoveries.

The aims of the analysis presented in this chapter were as follows:

- ⇒ To find genetic loci associated with the variation in plasma protein levels
- ⇒ Elucidate potential mechanisms of action of these associated loci
- ⇒ Investigate the relationships between protein levels and disease
- ⇒ Highlight any potential therapeutic targets

5.1.2 Contributions

The SCALLOP Consortium and Jim Wilson conceived the idea for this project, based on a previous similar SCALLOP Consortium project that focussed on a different set of proteins[149]. The principal investigators of the 18 participating cohorts (**Supplementary Table 25**) gathered the data and organised the collection of the Olink Proteomics assay for their respective cohorts.

QC of the genotype data and its subsequent imputation for the ORCADES and Croatia-Vis cohorts was carried out as described in **Chapter 2: Data & Methods** by other analysts and I was given access to the QC'd HRC imputed data. I received the proteomics data for both ORCDAES and Croatia-Vis from Olink having already undergone normalisation and in NPX units (**2.5 Genome Wide Association Studies**). I then performed the removal of outliers, trait transformation and genome-wide association study for the levels of the 184 plasma protein levels in these two cohorts.

I used an existing pipeline to perform the GWA studies created by Peter Joshi, David Clark, Paul Timmers and Andrew Bretherick.

Analysts from each of the remaining 16 cohorts contributing data carried out genome-wide association analyses (GWAS) of as many of the 184 proteins in our set as they had available. These GWAS followed the analysis plan drawn up by Jim Wilson, Peter Joshi and me. GWAS summary statistics were collected on the University of Edinburgh secure server.

I performed the harmonisation of the summary statistics from all contributing cohorts, using scripts that I adapted from Lasse Folkersen and performed basic quality control. I also performed the 184 GWAMAs and the following downstream analysis: conditional analysis, definition of significant loci, heritability, colocalisation of discovered pQTL with publicly available eQTL datasets, genetic correlations and the Mendelian randomisation analysis.

Lucija Klarić prepared the list of significant pQTL from 22 previously published GWAS on plasma proteins levels and prepared the list of drugs and drug targets from the drugbank database[256] that were used to derive the results in the relevant sections. Further I adapted an existing pipeline written by Paul Timmers to perform the SMR-HEIDI analysis.

I compiled all of the results, created the figures and wrote the first draft of the manuscript. Jim Wilson and Peter Joshi contributed to the redrafting of the manuscript and all co-authors commented on the manuscript prior to submission to medRxiv.

The following manuscript has been placed on medRxiv https://doi.org/10.1101/2021.08.03.21261494 and at the time of writing is being revised prior to submission to a peer-reviewed journal.

5.2 Manuscript Pre-print

Mapping genetic determinants of 184 circulating proteins in 26,494 individuals to connect proteins and diseases

Erin Macdonald-Dunlop[1], Lucija Klarić[2], Lasse Folkersen[3], Paul R.H.J. Timmers[1,2], Stefan Gustafsson[4], Jing Hua Zhao[7], Niclas Eriksson[8], Anne Richmond[2], Stefan Enroth[9], Niklas Mattsson-Carlgren[10,11,12], Daria V. Zhernakova[13,38], Anette Kalnapenkis[15], Martin Magnusson[10,16,17,44], Eleanor Wheeler[18], Shih-Jen Hwang[19,20], Yan Chen[21,22], Andrew P Morris[23], Bram Prins[7], Urmo Võsa[15], Nicholas J. Wareham[6,18], John Danesh[6,7,24], Johan Sundstrom[4,25], Bruna Gigante[27], Damiano Baldassarre[43], Rona J. Strawbridge[6,27,28], Harry Campbell[1], Ulf Gyllensten[9], Chen Yao[19,20], Daniela Zanetti[29], Themistocles L. Assimes[29,30], Per Eriksson[27,31], Daniel Levy[19,20], Claudia Langenberg[6,18,32], J. Gustav Smith[10,33,34,35,36,37], Tõnu Esko[15], Jingyuan Fu[13,14], Oskar Hansson[11,39], Åsa Johansson[9], Caroline Hayward[2], Lars Wallentin[4,26], Agneta Siegbahn[4], Lars Lind[4], Adam S. Butterworth[6,7,40], Karl Michaëlsson[41], James E. Peters[5,6,7], Anders Mälarstig[22,42], Peter K. Joshi[1]*, James F. Wilson[1,2]*

1 Centre for Global Health Research, Usher Institute, University of Edinburgh, Teviot Place, Edinburgh, UK
2 MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, UK
3 Danish National Genome Center, Copenhagen, Denmark
4 Department of Medical Sciences, Uppsala University, Uppsala, Sweden
5 Department of Immunology and Inflammation, Faculty of Medicine, Imperial College London, London, UK
6 Health Data Research UK, Wellcome Genome Campus and University of Cambridge, Cambridge, UK
7 British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
8 Uppsala Clinical Research Center (UCR), Uppsala University, Uppsala, Sweden
9 Department of Immunology, Genetics and Pathology, Uppsala University, Sweden
10 Wallenberg Center for Molecular Medicine, Lund University, Sweden
11 Clinical Memory Research Unit, Faculty of Medicine, Lund University, Lund, Sweden

12 Department of Neurology, Skåne University Hospital, Lund University, Lund, Sweden
13 Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
14 Department of Pediatrics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
15 Estonian Genome Centre, Institute of Genomics, University of Tartu, Riia 23b, 51010 Tartu, Estonia
16 Department of Clinical Sciences, Lund University, Malmö, Sweden
17 Hypertension in Africa Research Team (HART), North West University, Potchefstroom, South Africa
18 MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge, UK
19 Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA
20 Framingham Heart Study, Framingham, MA, USA.
21 Department of Medicine, Karolinska Institute, Stockholm, Sweden
22 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
23 Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, UK
24 Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK
25 The George Institute for Global Health, University of New South Wales, Sydney, Australia
26 Uppsala Clinical Research Center, Uppsala University, Uppsala, Sweden
27 Division of Cardiovascular Medicine, Department of Medicine, Karolinska Institutet, Stockholm, Sweden
28 Institute of Health and Wellbeing, College of Medicine, Veterinary and Life Sciences, University of Glasgow, UK
29 Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA
30 Palo Alto VA Healthcare System, Palo Alto, CA, USA
31 Karolinska University Hospital, Stockholm, Sweden
32 Computational Medicine, Berlin Institute of Health (BIH) at Charité – Universitäts Medizin Berlin, Germany
33 Department of Cardiology, Clinical Sciences, Lund University
34 Skåne University Hospital, Lund, Sweden
35 Lund University Diabetes Center, Lund University, Lund, Sweden
36 The Wallenberg Laboratory/Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University
37 Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden
38 Laboratory of Genomic Diversity, Center for Computer Technologies, ITMO University, St. Petersburg, Russia
39 Memory Clinic, Skåne University Hospital, Malmö, Sweden
40 National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, United Kingdom
41 Department of Surgical Sciences, Unit of Medical Epidemiology, Uppsala University, Uppsala, Sweden
42 Pfizer Worldwide Research, Development and Medical, Sweden
43 Department of Medical Biotechnology and Translational Medicine, Università di Milano, Milan, Italy; Centro Cardiologico Monzino, IRCCS, Milan, Italy
44 Department of Cardiology, Skane University Hospital, Malmoe, Sweden

*Authors contributed equally. Correspondence to J.F.W. (jim.wilson@ed.ac.uk).

Abstract

We performed the largest genome-wide meta-analysis (GWAMA) (Max N=26,494) of the levels of 184 cardiovascular-related plasma protein levels to date and reported 592 independent loci (pQTL) associated with the level of at least one protein (1308 significant associations, median 6 per protein). We estimated that only between 8-37% of testable pQTL overlap with established expression quantitative trait loci (eQTL) using multiple methods, while 132 out of 1064 lead variants show evidence for transcription factor binding and found that 75% of our pQTL are known DNA methylation QTL. We highlight the variation in genetic architecture between proteins and that proteins share genetic architecture with cardiometabolic complex traits. Using *cis*-instrument Mendelian randomisation (MR), we infer causal relationships for 11 proteins, recapitulating the previously reported relationship between PCSK9 and LDL cholesterol, replicating previous pQTL MR findings and discovering 16 causal relationships between protein levels and disease. Our MR results highlight IL2-RA as a candidate for drug repurposing for Crohn's Disease as well as 2 novel therapeutic targets: IL-27 (Crohn's disease) and TNFRSF14 (Inflammatory bowel disease, Multiple sclerosis and Ulcerative colitis). We have demonstrated the discoveries possible using our pQTL and highlight the potential of this work as a resource for genetic epidemiology.

Introduction

Proteins are the key functional elements in the body and are instrumental in most biological processes including, growth, repair, transport and signalling. Dysregulation of proteins circulating in the blood is often observed in disease and, moreover, is often part of the causal pathway, making them ideal candidate drug targets. The plasma proteome, which consists of both proteins which are actively secreted and passively leaked from cells, is an attractive and accessible system to study. As samples are easy to store, collection is minimally invasive for study participants, and hundreds to thousands of different proteins can be measured, plasma proteins have been investigated as biomarkers for numerous diseases[109]. The recent advances in targeted proteomics technologies have allowed thousands of circulating plasma protein levels to be measured simultaneously, even in large sample sizes[115,116,153,254,257–260]. Uncovering relationships between protein biomarkers and disease has the potential to aid in prediction of risk, diagnosis and development of new therapies for disease[261]. Cardiovascular disease (CVD)-related proteins are of particular interest as CVD was the leading cause of morbidity and mortality globally in 2019[11], being responsible for 18.6 million deaths and 393 million disability adjusted life years (DALYs).

As circulating plasma protein levels are partly heritable[262], genome-wide association studies (GWAS) have been used to discover genetic loci that are associated with regulation of protein levels; protein quantitative trait loci (pQTL)[115–121]. Previous pQTL studies have uncovered potential mechanisms of action for how common genetic variants affect circulating protein levels[116,149].

Using Mendelian Randomisation (MR) to assess potential causal relationships between biomarkers and disease phenotypes has become an increasingly utilised approach for drug target discovery and validation and has also successfully predicted outcomes of randomised controlled trials (RCTs)[263]. Despite many associations between levels of circulating biomarkers and various diseases in the literature, positing causal roles for these biomarkers has only been possible through the application of methods such as MR. The study of the genetics of circulating biomarkers such as plasma protein levels therefore has the capacity to uncover pathways, disease aetiology, therapeutic targets and biomarkers to aid detection and diagnosis of disease[261]. Unlike GWAS of complex traits, targets highlighted studying the plasma proteome are themselves directly actionable.

Previous large GWAS of plasma protein levels have discovered hundreds of associated loci, uncovered mechanisms for pQTL, causal relationships between proteins and diseases and posited how plasma protein levels may act to influence disease risk[115,116,120,149,153,254,255]. In order to maximise the potential for pQTL discovery and MR to find causal associations with disease and build on previous work, we performed genome-wide meta-analysis with the largest sample size for 184 cardiovascular-related plasma proteins. We uncovered insights into the genetic architecture of these proteins, indicating potential mechanisms for pQTL from altering gene expression and beyond. Using a broad exploratory analysis, we demonstrate the power of pQTL as genetic instruments in MR and highlight potential causal relationships between proteins and disease. These results suggest putative drug targets and repurposing opportunities. With this work we have created a resource of pQTL data that will aid the field of genetic epidemiology and provide tools for targeted experiments in the future.

Results

Discovery of Protein Quantitative Trait Loci

We performed genome-wide association meta-analyses (GWAMA) of the levels of 184 cardiovascular-related plasma protein levels measured by the Olink proximity extension assay in a maximum of 26,494 individuals of European ancestry from 18 cohorts. We identified 1,073 significant protein-locus associations (*cis*: $p<1.18 \times 10^{-7}$, *trans*: $p<5.55 \times 10^{-10}$, where *cis* was defined as ±1 Mb flanking the protein-coding region). After performing conditional analysis, we report a further 235 conditionally-independent protein-variant associations. In total we found 1,308 significant lead variant-protein associations, 288 *cis*-associations and 1,020 in *trans* (**Figure 22a**, **Supplementary Table 26**). This equates to the discovery of 592 independent loci significantly associated with the levels of at least one protein (**Supplementary Table 27**).

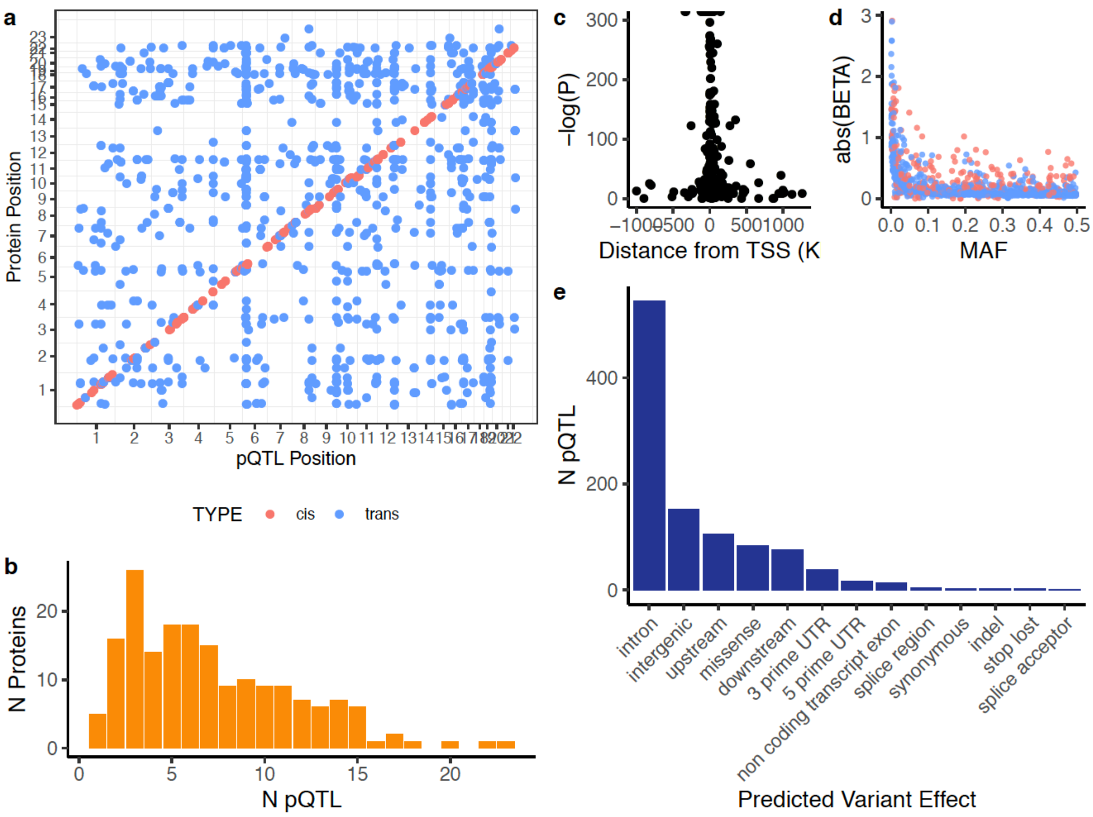


Figure 22. Significantly associated loci from GWAMA of 184 proteins. *a) Points indicate the genomic position of 1,308 significant pQTL against the genomic position of the transcriptional start site (TSS) of the gene encoding the protein that the variant is associated with. Colour indicates if the variant is a cis- or trans-pQTL. Cis- is defined as any variant within ±1 Mb of the coding region of the gene encoding the protein, trans- defined as any variant outside that region. b) Histogram of the counts of significant pQTL per protein. c) Relationship between -log P-value and distance of each cis-pQTL from the TSS of the gene. d) Magnitude of effect size (absolute beta) shows a typical L-shaped relationship with minor allele frequency (MAF) of our pQTL (cis in blue, trans in pink). e) The frequency of predicted effect of the sentinel variants.*

Only two proteins, growth hormone 1 (GH) and Inhibitor of nuclear factor kappa B kinase regulatory subunit (NEMO), had no significant pQTL. For an additional 16 proteins we found no significant *cis*-signals (ACE2, CCL22, CD40-L, CD93, Ep-CAM, GDF-2, HAOX1, ICAM-2, IL-6, ITGB1BP2, MB, PDGF subunit B, PECAM-1, SRC, t-PA, VEGF-D). For five of the above eighteen proteins, ACE2, CD40-L, NEMO, ITGB1BP2 and VEGF-D, this is expected as they are encoded on the X chromosome, which was not studied here. For the remaining 13 proteins, the minimum p-value and *cis* regions are shown in **Supplementary Figure 46**. Altogether, we report significant *cis*-pQTL for 92.7% of the plasma proteins (where *cis*-regions were tested). 182 out of the 184 proteins analysed had a pQTL, 155 proteins had both *cis*- and *trans*-pQTL, 11 *cis*-only and 16 *trans*-only.

The majority of proteins had 6 or fewer significant pQTL (Median N pQTL = 6), with 3 proteins (CD163, CTSL1 and RAGE) having more than 20 (**Figure 22b**). In general proteins with multiple significant pQTL had 1-3 *cis*-associated pQTL with any additional associated loci being in *trans* (**Supplementary Figure 47**). Interestingly, 216 loci contained lead variants for pQTL for more than one protein, with the *HLA* and *ABO* regions being associated with the most proteins (**Supplementary Table 28**). We saw a distinct pattern with our significant *cis*-associated variants, such that variants nearest the transcriptional start site (TSS) of the relevant gene displayed the strongest associations (**Figure 22c**). As seen for most complex traits, the magnitude of effect size increased with decreasing minor allele frequency (**Figure 22d**). Using single variant annotation from Ensembl variant effect predictor[264] we found that 70% of our lead variants are either intronic or intergenic (**Figure 22e**).

Six hundred and twenty-one (47.5%) of our significant lead variants (or variants in LD, r^2>0.5, with our lead variants) have previously been reported as genome-wide significantly associated with the relevant protein in previous GWAS of plasma protein levels (Details of previous studies summarised in **Supplementary Table 29**). Thus, 687 (52.5%) of our significant protein-variant associations are novel. We also reported 20 novel proteins with significant pQTL.

Genetic Architecture of Plasma Protein Levels

Unlike traditional complex polygenic traits, many of the proteins have an extremely strong *cis*-association signal and have individual variants that explain relatively large proportions of variance in the phenotype. Having very few (or one) strong signals is rare outside of molecular phenotypes, as many weak signals with small effects are common in complex traits. We found that 75 of our 1,308 lead variants have estimated phenotypic variance explained (R^2) of >5% (51 *cis* and 24 *trans*), with the highest being rs12141375:A, estimated to explain 32.7% of the variance in plasma CHIT1 levels (**Supplementary Figure 48**).

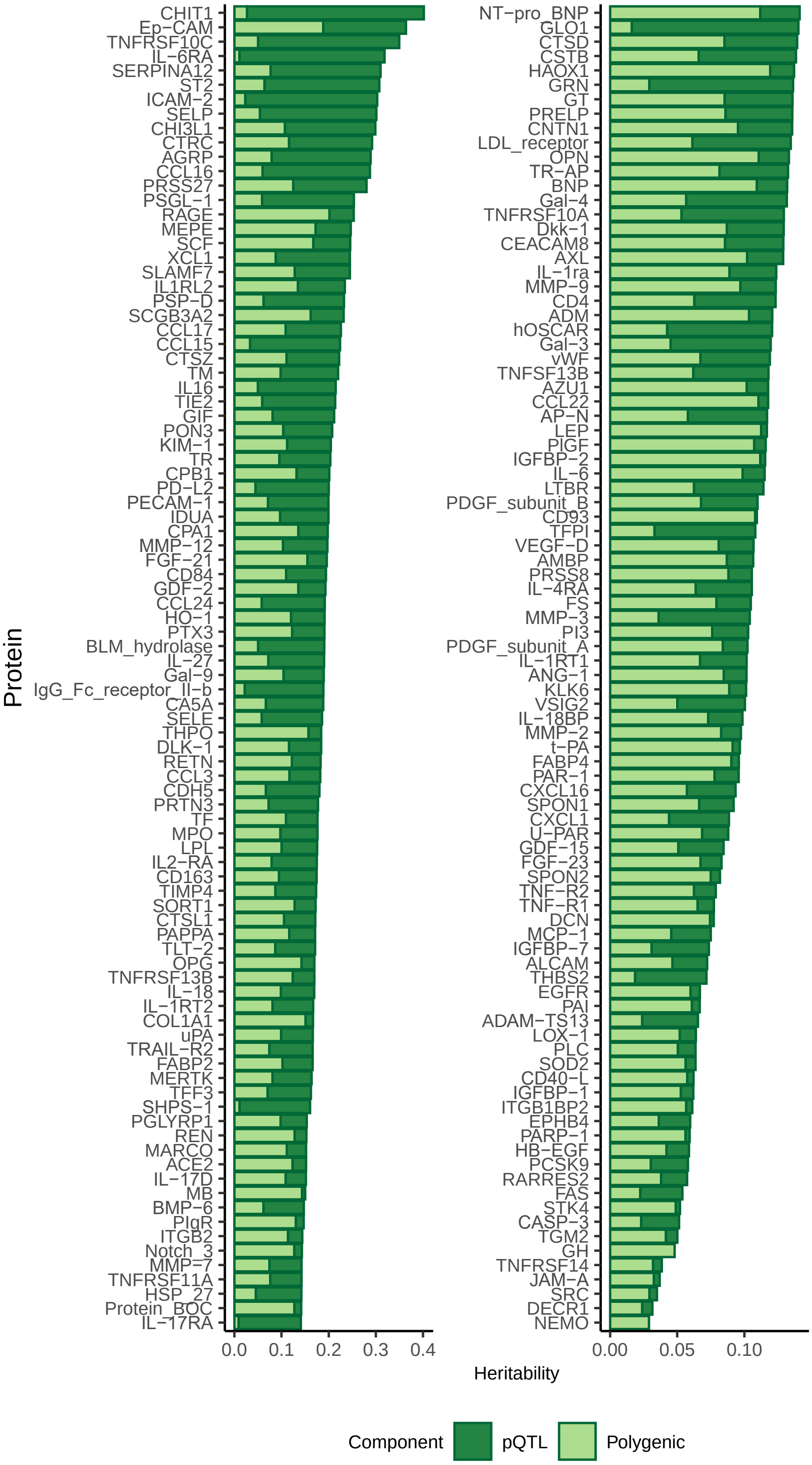


Figure 23. pQTL vs Polygenic Contribution to SNP heritability. *The SNP heritability estimated for each protein, stratified by contributions from significant pQTL and polygenic effects. Polygenic: LDSC-estimated SNP heritability excluding variants indexed by the lead variants, pQTL: sum of the estimated variance in protein level explained by the lead variants (See Methods for details).*

Standard methods for estimating single nucleotide polymorphism (SNP) heritability from association summary statistics assume a polygenic model that is unlikely to hold for proteins. We therefore calculated the heritability contributed by significant pQTL (pQTL component) and the remaining genome-wide SNP heritability (polygenic component), separately for each protein (**Figure 23**). The pQTL component was calculated as the sum of the estimated variance in protein level explained by the lead SNPs and the polygenic component estimated using LD-score regression[157,160] (see methods for details). Estimates of total genetic component ranged from 2.9% for NEMO protein levels to 40.2% for CHIT1. Genetic architecture, however, varied across proteins with IL-6RA and CHIT1 protein levels having identified pQTL accounting for 96.5% and 93% of their SNP heritabilities, respectively. Conversely, the genetic components of NEMO and GH protein levels appear entirely polygenic, having no significant pQTL in this analysis.

We observed that there is a relationship between the number of significant pQTL we found and the estimated SNP heritability, with increasing heritability estimates with increasing number of pQTL (**Supplementary Figure 49**).

Colocalisation of pQTL & eQTL

We sought to uncover potential mechanisms by which our pQTL might act to influence the level of proteins circulating in plasma. Biologically, the most direct route, would be for the significantly associated variants to affect protein levels by altering gene expression. 36.5% of the lead *cis*-variants have been previously reported as *cis*-expression QTL (eQTL) for the gene encoding the protein of interest (eQTLGen[180], at 5% FDR (permutation-based)). However, for each of our pQTL, the lead variant (strongest association based on p-value) is not necessarily the causal variant. The lead variant commonly tags the signal for multiple variants in high LD, any of which could be the true causal variant.

To further define whether the signals were shared we used two different approaches. We first looked for evidence of gene expression mediating the effect of our pQTL on plasma protein abundances using summary based Mendelian Randomisation (SMR) and tested that these estimates were not due to linkage using the heterogeneity in dependent instruments (HEIDI) test[183]. We found associations between 1,371 transcripts and 168 proteins (P_{SMR}<1.68 x 10^{-7}, $P_{HEIDI}\geq$0.01) in at least one of four eQTL datasets (eQTLGen[180],GTEx v7, Westra *et al.*[181] and Cage[182]). The number of significant transcript-protein associations across different eQTL resources are shown in **Supplementary Table 30**, as well as how many of our proteins are associated with the expression level of the transcript encoding the protein, compared to other transcripts.

Secondly, to formally test whether the association signals with gene expression and protein level share the same causal variant or are driven by different variants, we looked for evidence of colocalisation. The Bayesian framework implemented by coloc[184] assesses several hypotheses simultaneously by estimating separate posterior probabilities (PP) of the eQTL and pQTL i) sharing a single causal variant or ii) being caused by two independent variants. We found that 18 out of 220 testable *cis*-pQTL showed strong evidence of colocalisation (PP>0.8) with the *cis*-eQTL in whole blood using the eQTLgen summary statistics, with an additional 2 being likely to share a causal variant (PP>0.5). Using eQTL data for 48 different tissues from GTEx v7, we found that 40 out of 277 testable *cis*-pQTL colocalise (PP>0.8) with the *cis*-eQTL in at least one tissue, with 12 more being likely to share a causal variant (PP>0.5). The majority of pQTL which colocalise with eQTL do so in <6 tissues, however there are several that colocalise with eQTL across >20 tissues (**Supplementary Figure 50**). Interestingly, there are very few that appear to be tissue specific.

Both of these approaches share the caveat that they are unable to distinguish causality from pleiotropy. However, given that we are assessing the effect of genetic variants on gene expression and protein levels, the central dogma suggests these relationships are likely to be causal, but a definitive statement of causality for individual associations cannot be made using current methods.

Other Potential Mechanisms

As not all pQTL appear to act by altering gene expression, we looked for other potential mechanisms of action. For some proteins we found *trans*-pQTL that map to that protein's receptor or vice versa. For example, despite having no significant *cis*-pQTL, IL-6 has an extremely strong association in the IL-6 receptor (IL6-RA) region.

Given that 70% of our lead variants are intronic or intergenic, we next looked for existing annotation of regulatory function using RegulomeDB[265] (**Supplementary Figure 51a**). Of the 1064 of our 1093 lead variants that have an entry in RegulomeDB, 50 (8 *cis* and 42 *trans*) were placed in category 1, meaning they are known eQTL with varying additional levels of support (e.g. transcription factor (TF) binding, TF motif, DNase footprint), for the variant being located in a functional region. 82 of our lead variants (25 *cis*, 56 *trans* & 1 both *cis* and *trans*, but for different proteins) were scored in category 2, meaning that despite not being known eQTL, variants have direct evidence of binding from ChIP-seq and DNase footprinting. These results suggest that a substantial minority of pQTL that are not yet reported as being significant eQTL influence TF-binding.

To uncover potential mechanisms for our *trans*-pQTL, we used annotation databases to see if *trans*-genes share pathways or are known to interact with the protein of interest. We defined *trans*-genes as all genes whose coding regions overlapped with a 1 Mb window centred on the lead variant of the *trans*-pQTL. We found that 85 of our *trans*-pQTL have a *trans*-gene with a known interaction with the protein of interest using STRINGdb[266]. Similarly, 37 *trans*-pQTL have a *trans*-gene that shares a common KEGG[267] pathway with the gene encoding the protein of interest, 158 share common gene ontology (GO) terms and 816 have a *trans*-gene that is mentioned in a publication together with the protein of interest (**Supplementary Table 31**).

pQTL Associated with Metabolites, DNA methylation levels & Complex Traits

Aside from affecting gene expression and plasma protein levels, our pQTL have also been previously associated with the levels of metabolites circulating in the plasma, with methylation of CpG dinucleotides and with complex traits. Using Phenoscanner[268] we established that, of our 1093 unique lead variants: 96 have been reported as significantly ($p<5 \times 10^{-8}$) associated with circulating metabolite levels, 816 with DNA methylation and 547 with complex traits in GWAS (**Supplementary Figure 51b**).

The 547 lead variants reported in previous GWA studies, have been significantly associated with a broad range of phenotypes, from cardiovascular-related phenotypes to immune and inflammatory diseases (**Supplementary Table 32**). Lead variants were also associated with anthropometric and adiposity-related traits, which are themselves risk factors for cardiovascular health; several causes of death in the UK Biobank (e.g. heart failure, vascular disease); and, unsurprisingly, blood protein, lipid and metabolite levels, as well as various red blood cell and immune cell counts.

As these results are association-based, they do not confirm the causal direction of the relationship between protein level and disease phenotype. Similarly, the observation that the lead variant at a pQTL is associated with another trait provides no evidence that the same variant is causal for both traits. An alternative approach to detect evidence of shared genetic risk variants, rather than these single-SNP associations, is to look systematically across the whole genome to see if alleles that increase plasma protein levels also increase disease risk.

Genetic Correlations

To investigate if our proteins share genetic architecture with complex traits or cardiometabolic risk factors, we used High definition likelihood[161] to estimate genetic correlations of our proteins with 14 important risk factors or outcomes (**Supplementary Table 33**, full results are in **Supplementary Figure 52**). Genetic correlations that remained statistically significant after Bonferroni correction for multiple testing ($p<1.95 \times 10^{-5}$) are shown in **Figure 24**. Interestingly, the traits with the most significant correlations with protein levels were BMI, WHR, creatinine and T2D; for BMI, WHR and T2D the majority were positive correlations, whereas all significant correlations between protein levels and creatinine were negative. IGFBP-2 levels are significantly genetically correlated with the most (8) traits including: lower BMI, WHR, total triglyceride levels, Type II diabetes and creatinine, but with increased HDL levels.

Several genetic correlations recapitulated known relationships. For example, we find that leptin levels (LEP) are genetically correlated with increased BMI, WHR, type II diabetes, coronary artery disease (CAD), risk of myocardial infarction (MI), and lower HDL levels. This finding recapitulates known biology as leptin is involved in the regulation of energy homeostasis and is linked to type II diabetes and cardiovascular phenotypes[269].

We also discovered novel correlations. The levels of LTBR (lymphotoxin beta receptor) circulating in the plasma were genetically correlated with increased BMI while BOC (Brother of CDO) and VSIG2 (V-set and immunoglobulin domain containing 2) levels correlated with WHR. None of these three proteins has been previously associated with adiposity-related traits.

In summary, our plasma protein levels share risk variants across the genome with health-related risk factors and disease outcomes, although an important caveat is that we are unable to distinguish the direction of these relationships from this analysis.

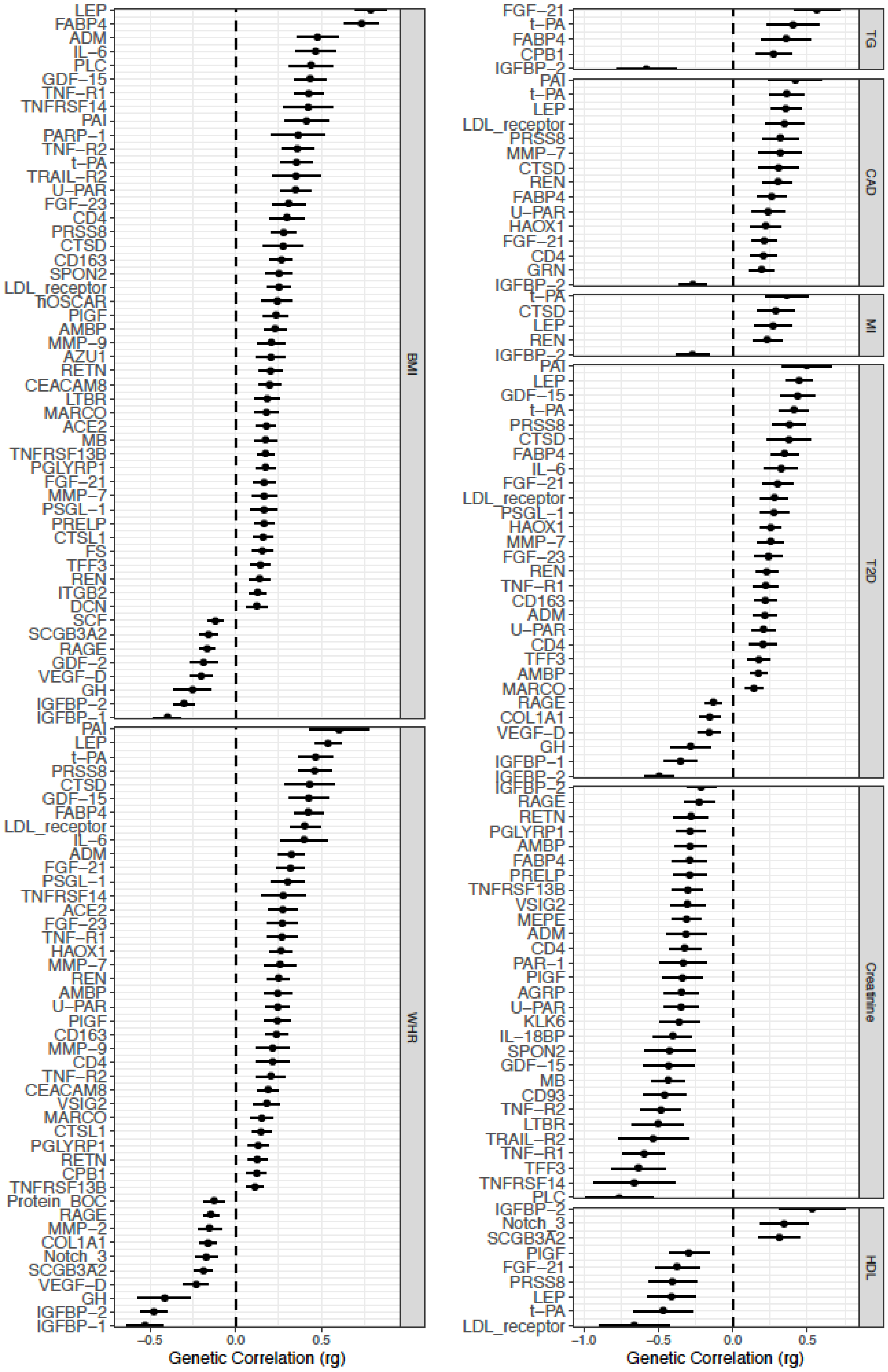


Figure 24. Genetic correlations show shared architecture between plasma protein levels and complex traits. *Genetic correlation coefficients (r_g) calculated using High definition likelihood for protein levels and complex traits. Only traits passing Bonferroni significance ($p<1.95 \times 10^{-5}$) are included (full results in **Supplementary Figure 52**). Error bars indicate 95% confidence intervals of r_g estimation. BMI: Body mass index, WHR: waist-to-hip ratio, TG: triglyceride level, CAD: coronary artery disease, MI: myocardial infarction, T2D: type II diabetes mellitus, HDL: high density lipoprotein.*

Causal inference using Mendelian Randomisation

To identify potential causal relationships between plasma protein levels and disease we used Mendelian Randomisation (MR). We limited our analysis to only using *cis*-associated variants as instrumental variables (IVs) to reduce the influence of pleiotropy on our results. We also excluded any variants in the highly pleiotropic *HLA* and *ABO* regions. An LD threshold of $r^2>0.001$ was used to remove correlated variants. We tested the association of the 169 proteins that had IVs meeting these criteria with 121 outcomes available from MR-Base[170] (outcomes listed in **Supplementary Table 34**). 96 protein-outcome causal effect estimates passed a 1% FDR (Benjamini-Hochberg method[233]) significance threshold.

MR relies on assumptions that are difficult to test empirically. To increase confidence in our results, we therefore performed additional sensitivity analysis (**Supplementary Table 35**). To test the consistency of the causal estimates across IVs, we excluded any protein-outcome pairs if there was evidence of significant heterogeneity using Cochran's Q test (q-value <0.05)[270]. To limit the chance of reverse causality influencing our results we performed bidirectional MR[271] and excluded protein-outcome pairs that had significant causal effect estimates of outcome on protein level ($p<3.62 \times 10^{-6}$). For proteins with multiple *cis* IVs, we performed the pleiotropy-robust method MR-Egger[272]. An MR-Egger intercept estimate that is significantly different from zero can be interpreted as indicative of horizontal pleiotropy[272,273]. We therefore excluded protein-outcome MR estimates that had MR-Egger intercept p-values <0.05, leaving 59 significant protein-outcome causal estimates. Finally, to distinguish causal relationships from confounding due to LD we performed colocalisation analysis to look for evidence of a shared causal variant underpinning the genetic associations with protein and outcome. We used coloc[184] and only considered protein-outcome pairs with a posterior probability (PP) of >0.8 of the hypothesis of a shared causal variant. We report 20 protein-outcome causal effect estimates that meet all of these criteria, involving 11 proteins associated with 16 different outcomes (**Figure 25**).

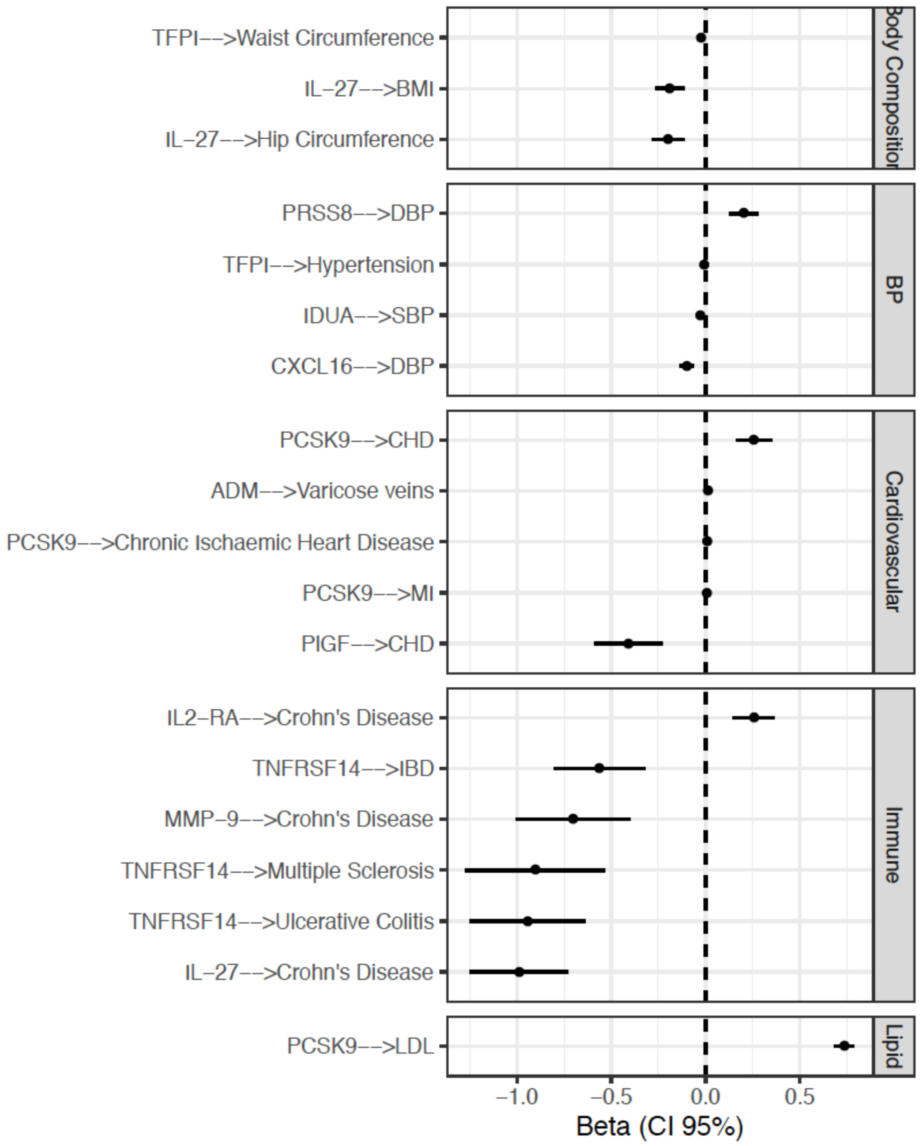


Figure 25. Cis-Instrument Mendelian Randomisation of plasma protein levels on complex diseases and health-related risk factors. *MR causal effect estimates and 95% confidence intervals of the effect of plasma protein levels on outcome. Results from the fixed effects Inverse variance-weighted (IVW) method that passed 1% FDR, had a heterogeneity Q-value >0.05, an MR-Egger intercept p-value of >0.05, as well as evidence of a shared causal variant from colocalisation analysis (posterior probability of a shared causal variant >0.8) are shown. Associations are grouped by type of outcome. Causal effect estimates from additional MR methods that are robust to horizontal pleiotropy and relax the assumption of IVW allowing correlations between genetic associations with the exposure and outcome are in* ***Supplementary Figure 53****, as further sensitivity analyses. BMI: body mass index, DBP: diastolic blood pressure, SBP: systolic blood pressure, CHD: coronary heart disease, MI: myocardial infarction, IBD: inflammatory bowel disease, LDL: low density lipoprotein cholesterol levels.*

The significant MR causal effect estimate of increased PCSK9 levels increasing LDL cholesterol levels (beta:0.74, SE:0.026) provides validation of the approach since the causal relationship of PCSK9 levels and LDL and total cholesterol levels is firmly established[274]; pharmacological inhibition of PSCK9 results in dramatic reductions in LDL cholesterol. In addition, our MR analysis confirmed previous reports indicating that PCSK9 increases risk of cardiovascular disease. This result is consistent with the findings of reduction in cardiovascular events in randomised clinical trials of PCSK9 inhibitors[37]. We also replicated other results from previous MR studies examining the role of circulating proteins in cardiovascular diseases and traits. This included the finding that a genetic tendency to higher placenta growth factor (PlGF) protein levels decreases the risk of CHD[149], and that a genetic tendency to higher C-X-C Motif Chemokine Ligand 16 (CXCL16) protein levels decreases diastolic blood pressure[178].

Variation in the genes encoding several of the proteins examined in our study have been associated with particular phenotypes. For example, SNPs mapped to Serine Protease 8 (*PRSS8)*, Interleukin 2 Receptor Subunit Alpha (*IL2RA)* and Tissue Factor Pathway Inhibitor (*TFPI)* have been associated with DBP[237], Crohn's disease (CD)[207,262,275] and waist circumference[237], respectively. Here, we advance these associations by demonstrating likely causal relationships between the circulating protein and the corresponding phenotypes through MR for the first time.

Our MR analysis provides novel insight into the pathogenesis of inflammatory bowel disease (IBD), which encompasses Crohn's disease (CD) and ulcerative colitis (UC). IL27 is a heterodimeric cytokine that has complex biological functions including both pro- and anti-inflammatory effects in the intestine. IL27 can inhibit differentiation of Th17 cells, an important cell type in the pathogenesis of IBD. There are conflicting data on IL27's role in IBD. In most[276,277], although not all[278], murine models of gut inflammation, IL27 is protective: IL27R genetic knockout worsens colitis while exogenous administration of IL27 ameliorates disease. In patients with IBD, *IL27* gene expression is elevated compared to controls[279]. Here we show that, in contrast to the observational human data, a genetic tendency to higher circulating IL27 is associated with lower risk of CD. This raises the possibility that the IL27 elevation in IBD patients results from reverse causation, perhaps as a response to dysregulated gut inflammation. Our data is in keeping with the observation that individuals with the risk allele for CD have lower *IL27* gene expression[280]. Together this supports the concept that IL27 acts to protect the gut from aberrant inflammatory responses and raises the possibility that IL27 might be of therapeutic benefit in IBD.

By evaluating whether proteins play a causal role in disease aetiology, MR provides a valuable tool to identify and validate potential drug targets before embarking on costly clinical trials. We therefore examined whether any of the 11 proteins with inferred causal relationships in our MR analysis (**Figure 25**), were already current targets, using the DrugBank database[256]. In addition to PCSK9, which, as described previously, is a target of existing drugs used successfully in the treatment of hypercholesterolaemia and cardiovascular disease[37,38,281,282], we found that 5 other proteins: PlGF, PRSS8, IL2-RA, MMP-9 (Matrix Metallopeptidase 9) and TFPI are also targets for drugs in various stages of development (**Supplementary Table 36**).

Our results highlighted IL2-RA as a potential candidate for drug repurposing. IL2-RA is the target for three approved drugs, two of these: Denileukin diftitox and Basiliximab, inhibit IL2-RA and are used for cutaneous T-cell Lymphoma (CTCL)[39] and to prevent kidney transplant rejection[129], respectively. The third, Aldesleukin, is an agonist and increases IL2-RA activity, inducing the adaptive immune response in the treatment of renal cell carcinoma[40,128]. Basiliximab has been piloted for use in IBD (UC) patients with apparent success in an uncontrolled open-label study[283], however no benefit was found in an RCT[284]. Our finding that genetically increased levels of IL2-RA protein increase risk of CD (Beta: 0.26, SE: 0.06) suggest that further investigation is warranted into whether the suitability of Basiliximab (given the previous contradictory findings) may have a role in the management of CD.

Our inference that genetic predisposition to elevated MMP-9 decreases the risk of CD (Beta: -0.7, SE: 0.15) aligns with previous GWAS results: SNPs mapped to the *MMP9* gene have been associated ($p<5 \times 10^{-8}$) with lower risk of CD[262,275]. These genetics results are contrary to previous observational findings that increased serum MMP-9 levels were prognostic of clinical flare ups in CD patients[285]. Since MR is less prone to confounding and reverse causation than observational studies, we hypothesise that raised MMP-9 levels during flares of CD are likely to arise from reverse causation, perhaps reflecting an injury response. In keeping with this the MMP-9 inhibitor, Andecaliximab, was ineffective in phase 2 trials[286], as would have been predicted by MR. This example highlights how integrating genetics and proteomics can be useful in deprioritising therapeutic targets.

We demonstrate the novel finding that MR identifies TNFRSF14 (HVEM) as protective against multiple immune-mediated diseases (IBD and MS). Notably, MS is also associated with polymorphisms in the TNFSF14 gene region, which encodes LIGHT, the ligand for TNFRSF14. The MS risk allele at TNFSF14 (LIGHT) is associated with lower serum levels of this protein[287]. This, together with our data, demonstrate that higher levels of both the ligand and its receptor are protective against MS, clearly indicating a causal role for this pathway in the maintenance of immune tolerance and raising the possibility that it could be manipulated for therapeutic benefit.

Discussion

We have performed the largest pQTL study (Max N=26,494) on 184 plasma protein levels to date and report 592 independent loci significantly associated with the levels of at least one protein (1,308 protein-lead variant associations), with 687 lead variant-protein associations being novel. We found that estimates of the proportion of pQTL that overlap with eQTL ranges from 8.2-36.5% using multiple publicly available eQTL datasets and methods. Our results highlight that the majority of pQTL do not appear to be explained by eQTL. Given this finding, we highlight other potential mechanisms of action such as regulation of ligand-receptor pairs and transcription factor binding. The genetic architecture of plasma protein levels varies across proteins, from entirely polygenic (NEMO, GH) to single loci explaining almost all of the estimated genetic component (IL6-RA). Plasma protein levels also share genetic architecture with health-related risk factors and complex traits, with 52 protein levels being genetically correlated with BMI and 21 sharing heritability with CAD and MI. We also performed an extensive exploratory MR analysis using *cis*-pQTL as instruments, and found significant causal effect estimates for the levels of 11 proteins on 16 different outcomes. Our MR analysis highlighted plasma proteins that are candidate novel therapeutic targets and a candidate for drug repurposing.

In line with the larger size of our study, the discovery of a significant *cis*-pQTL for 92.7% of the plasma protein levels (where we tested the *cis*-regions) surpasses previous GWAS of plasma protein levels (18.5% Sun *et al.* N~5,000, 86% Folkersen *et al.* N~15,000)[116,149]. Additionally, CD93, ICAM-2, IL-6, PECAM-1 and t-PA levels had variants with p-values passing the genome-wide significance threshold for *cis*-signals ($p<1 \times 10^{-5}$) but were lost after correction for multiple testing, suggesting that our analyses were still underpowered and further *cis*-pQTL could be found in larger studies. Other than an issue of power, it is possible that our definition of *cis* (±1 Mb surrounding the coding region of the gene encoding the protein) is not capturing all signals however, no additional signals were found when the *cis*-region was widened to ±2 Mb. Ep-CAM, CD93, HAOX1, ICAM-2, MB, PECAM-1 and SRC proteins are intracellular[288], which could contribute to significant signals not being found in samples from plasma. Expanding on previous studies, 78% of our significant pQTL were *trans*-associated compared to 68% in Folkersen *et al.* and 72% Sun *et al.* This is most likely due to our increased sample size, as like the aforementioned studies we found that proteins tended to have at most about 3 *cis*-pQTL, with any additional pQTL being *trans*-associated (**Supplementary Figure 47**), indicating that the increase in power allows the discovery of further *trans*-signals, which are likely to have smaller effect sizes (Welch T Test Two-sided p-value=1.48×10^{-9}).

Akin to findings by Folkersen *et al.*[149], we found that proteins varied in terms of their genetic architecture, with some proteins almost monogenic while others have polygenic architecture.

In terms of eQTL/pQTL overlap, our results based on direct lookup of lead variants found that 36.5% of our *cis*-pQTL had been previously reported as significant *cis*-eQTL (5% FDR). This is comparable to the 26% overlap based on variant lookup reported by Folkersen *et al.* and the 40% (including proxies LD $r^2 \geq 0.8$) by Sun *et al*. However, only 8.2% and 14.4% of our *cis*-pQTL showed strong evidence of colocalisation (PP>0.8) with the eQTL for the corresponding gene in eQTLgen and GTEx (at least one tissue), respectively. Additionally, coloc assumes that a single causal variant, included in the analysis, is driving the association signal in the region being considered. Given the strength of some of our *cis*-pQTL in particular, it is possible that the assumption of only one independent association signal could be violated. We limited the pQTL regions to ± 200 kb flanking the lead variant in our analysis, to minimise the chance of including multiple association signals. However, our findings are considerably lower than the reported 78.5% of 228 testable pQTL that showed evidence of colocalisation with eQTL in at least one tissue (PP>0.8) by Sun *et al.* One reason for this could be due to the difference in study design. Coloc assumes that the populations used to derive association statistics for the two traits have the same underlying pattern of LD. By meta-analysing multiple different populations, the LD structure in our sample will be different from those used to generate the eQTL datasets, whereas *Sun et al.* used only the INTERVAL cohort of English blood donors which may be a closer match to the GTEx population which was the source of their eQTL comparison. Recent methods that allow for multiple causal variants could overcome some of these issues. For example, the sum of single effects (SuSiE) regression framework coloc method[289], however, this approach does require LD matrices for both populations and the use of a reference such as UK Biobank or 1000 Genomes will still not completely capture the LD structure in a multi-cohort GWAMA sample.

More generally, there are several reasons why colocalisation approaches might fail to indicate eQTL/pQTL overlap other than eQTL and pQTL having two independent causal variants: namely, differences in: sample size, assay technology or tissues between the two traits. The issue of tissue of origin is of particular concern here as, despite plasma having benefits as a medium, it does not accurately capture the levels of proteins in the tissues or cell types in which they are expressed and subsequently secreted, an inherent limitation when drawing conclusions about mechanisms. It is likely that higher eQTL/pQTL overlap would be observed if high-powered eQTL or pQTL datasets were available for the tissues from which the genes encoding these proteins are expressed. Despite GTEx having multiple different tissues, the small sample size means that its power is limited. This could also contribute to the low number of apparent tissue-specific overlapping eQTL/pQTL found using GTEx, as only those strong and robust *cis*-eQTL that are shared between tissues were able to be detected.

Our finding that 74.7% of our pQTL have been previously reported as DNA methylation QTL (meQTL) mirrors previous findings that 82% of *cis*-pQTL[120,255] are also *cis*-meQTL[290]. These results highlight the link between DNA methylation and regulation of protein expression and exploring the interaction between plasma proteins and the epigenome would be an interesting avenue for further study, as would whether pQTL act by influencing mRNA splicing.

We restricted our MR analysis to *cis* IVs only, in contrast to previous studies[153,178,291]. This decision was made to fully take advantage of the direct biological link between *cis*-pQTL and protein level and to prevent highly pleiotropic *trans*-pQTL influencing our results by breaking the assumptions underlying MR. A systematic assessment of *cis* vs *trans* Ivs would require the use of all of the most recent MR methods[292,293] and meaningful results would be lost due to multiple testing. Additionally, we performed sensitivity analysis in line with the procedure set out by Zheng *et al.*[178], for using pQTL as Ivs and showed that our causal effect estimates were consistent across multiple MR methods with varying assumptions (**Supplementary Figure 53**), therefore increasing confidence in the robustness of our results[291].

Our exploratory MR analysis for plasma protein levels with a broad range of outcomes using the *cis* Ivs was able to recapitulate the well-documented causal associations (PCSK9 with LDL and total cholesterol levels) and replicate findings reported by previous pQTL MR studies: genetically increased levels of CXCL16 and PlGF decreasing DBP[178] and the risk of CHD respectively[149]. We also found evidence of novel causal associations between circulating protein levels (PRSS8, IL2-RA, TFPI and IL-27) and traits where the corresponding gene was already known to be associated, and novel causal protein-outcome relationships for ADM and IDUA. Using pQTL as Ivs also highlighted IL2-RA as a potential candidate for drug repurposing and TNFRSF14 (IBD and MS) and IL-27 as novel therapeutic targets. Together these findings demonstrate the strength of our *cis*-pQTL as Ivs and the potential for future discoveries by disease-specific analyses using this resource[291].

Our increased sample size compared with previous pQTL studies[115,116,149,153], is a particular strength of this work as it allowed us to discover novel pQTL for use as instruments. Additionally, the breadth of our approach exhibits the range of possible downstream uses of GWAS of circulating plasma protein levels. However, this breadth is also a limitation, as our work has uncovered numerous findings that inspire further research. For example, we found a significant proportion of pQTL did not overlap with the corresponding eQTL. This could be due in part to pQTL acting to influence the protein levels via other mechanisms such as influencing translation, clearing of the protein, export or expression of the protein’s receptor. However, this could also be due to the predominant use of whole blood eQTL datasets and the limited power of the multi-tissue dataset (GTEx), given that our proteins are also secreted in several tissues other than blood. Further analyses using high-powered eQTL datasets from the relevant tissues would be required to untangle the mechanisms of action of these pQTL. Similarly, we emphasise that the potential therapeutic targets identified by MR are preliminary, and extensive investigations into other factors (e.g. druggability, safety) will also play a key role in determining the suitability of therapeutic intervention. As novel targets, TNFRSF14, ADM and IL-27, are either secreted into blood or retained membrane-bound or intracellular, dependent on isoform, further research into the specific functions of different isoforms is needed to validate their candidacy.

Our work builds on previous pQTL studies using a larger sample size for more proteins allowing the discovery of 1,308 significant protein-locus associations. By studying the genetic architecture of plasma protein levels, we have provided insight into the genetic regulation of protein levels, disease aetiology and casual relationships between circulating protein levels and cardiovascular disease phenotypes. In highlighting the power of our pQTL as IV to uncover candidate novel therapeutic targets in a broad exploratory analysis, we showcase the potential of this study as a resource to drive highly targeted research questions in the future.

Author Contributions

E.M.D contributed to the meta-analysis, visualisation of results. E.M.D, L.F, S.G, J.Z, N.E, A.R, S.E, N.M.C, D.V.Z, A.K., M.M, E.W, S.J.H, Y.C, A.P.M, B.P, U.V, N.J.W, J.D, J.S, B.G, D.B, R.J.S, H.C, U.G, C.Y, D.Z, T.L.A, P.E, D.L, C.L, J.G.S, T.E, J.F, O.H, Å.J, C.H, L.W, A.S, L.L, A.S.B, K.M, J.E.P and A.M, contributed to the cohort level analysis. E.M.D, L.K and P.R.H.J.T: contributed to other downstream analysis. E.M.D, P.K.J, J.F.W and J.E.P, contributed to manuscript drafting and editing. E.M.D, J.F.W and A.M contributed to project conception. P.K.J and J.F.W contributed to project supervision. All other authors commented and approved the manuscript prior to submission.

Acknowledgements

The proteomic work was carried out under the aegis of the SCALLOP consortium. We thank Aikaterini Siopi for administrative assistance. We thank the UK Biobank resource, approved under application 19655. Cohort-specific acknowledgements can be found in **Supplementary Table 25**. J.F.W. and C.H, acknowledge support from the MRC Human Genetics Unit programme grant “Quantitative traits in health and disease” (U.MC_UU_00007/10). The work of L.K. was supported by an RCUK Innovation Fellowship from the National Productivity Investment Fund (MR/R026408/1). E.M.D. acknowledges the MRC Doctoral Training Programme in

Precision Medicine (MR/N013166/1). RJS is supported by a UKRI Innovation- HDR-UK Fellowship (MR/S003061/1). We acknowledge the Swedish national research infrastructure SIMPLER (https://simpler4health.se/) for provisioning support in the analyses of COSM-C and SMCC. SIMPLER receives funding through the Swedish Research Council under the grant no 2017-00644.

Disclaimer: The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Competing Interests

Outside the submitted work J.D reports grants, personal fees and non-financial support from Merck Sharp & Dohme (MSD), grants, personal fees and non-financial support from Novartis, grants from Pfizer and grants from AstraZeneca outside the submitted work. J.D sits on the International Cardiovascular and Metabolic Advisory Board for Novartis (since 2010); the Steering Committee of UK Biobank (since 2011); the MRC International Advisory Group (ING) member, London (since 2013); the MRC High Throughput Science 'Omics Panel Member, London (since 2013); the Scientific Advisory Committee for Sanofi (since 2013); the International Cardiovascular and Metabolism Research and Development Portfolio Committee for Novartis; and the Astra Zeneca Genomics Advisory Board (2018). O.H has acquired research support (for the institution) from AVID Radiopharmaceuticals, Biogen, Eli Lilly, Eisai, GE Healthcare, Pfizer, and Roche. In the past 2 years, he has received consultancy/speaker fees from AC Immune, Alzpath, Biogen, Cerveau and Roche. P.K.J is a consultant to Humanity, Inc., a company developing direct-to-consumer measures of biological ageing and an advisor to Global Gene Corp, a company developing direct-to-consumer and business-to-business genomic solutions. A.S.B reports grants outside of this work from AstraZeneca, Biogen, BioMarin, Bioverativ, Merck, Novartis, Pfizer and Sanofi and personal fees from Novartis. J.E.P has received travel and accommodation expenses and hospitality from Olink to speak at Olink-sponsored academic meetings. E.M.D has received travel and accommodation expenses and hospitality from Olink to speak at Olink-sponsored academic meetings.

Data Availability

Meta-analyses summary statistics will be made publicly available upon publication.

Code Availability

METAL software for meta-analysis is available from http://csg.sph.umich.edu/abecasis/metal/download/. SMR-HEIDI is available from https://cnsgenomics.com/software/smr/#Download. The Coloc R package is available from https://github.com/chr1swallace/coloc.

Methods

Proteomics Assay

Participating cohorts performed protein measurement using an antibody-based proximity extension assay (Olink Bioscience, Uppsala, Sweden)[224] from EDTA plasma in 2 x 92-protein panels: ‘cvd2’ and ‘cvd3’. These targeted assays contained promising cardiovascular related proteins that also had two specific antibodies available for different epitopes. Analysis of all cohorts were conducted at one of two core laboratories with Olink Bioscience of SciLifeLab in Uppsala, Sweden.

Genome-wide Association

Summary statistics were obtained from 18 cohorts of European ancestry. Details of which cohorts contributed data for each protein are in **Supplementary Table 37**.

The maximum sample size across all proteins was 26,494 however, average per-protein maximum and mean sample sizes were 23,981 and 18,141 respectively. It is worth noting that the CCL22 GWAS had a considerably smaller sample size (Max N=7460) than the other proteins as it was removed from the CVDIII panel by Olink during the data collection phase of this study, meaning only a subset of contributing cohorts returned CCL22 summary statistics.

The majority of cohorts provided data imputed with 1000 Genomes Project phase 3 or higher or to the Haplotype Reference Consortium (HRC) reference (**Supplementary Table 25**). Cohorts applied quality control filters for call rates, gender mismatch, cryptic relatedness and ancestry outliers. Cohorts performed genome-wide association studies on the inverse rank normalised NPX values. Below lower-limit-of-detection values (<LOD) were included in the analysis. Cohorts ran linear models adjusting for study-specific covariates such as batch or genotyping array as well as: age, sex, first 10 principal components of the genotypes to account for population structure, plate number, plate row, plate column, sample time in storage (days) and season of venepuncture. Studies containing related individuals corrected for kinship.

Meta-analysis

METAL[253] software was used to perform inverse-variance-weighted meta-analysis (STDERR scheme) with the additional filters that only variants with an imputation quality score >0.4 and that were assessed in three or more cohorts were included. Heterogeneity of variant effect estimates between cohorts were also calculated using METAL.

Locus definition

In order to prevent heterogeneity influencing our results, only variants that had an I^2<30% or have both: i) effect direction consistent with the meta in at least 3 individual cohorts and ii) be nominally significant (p<0.05) in at least 3 individual cohorts, were eligible to be considered genome-wide significant. Separate significance thresholds pre-correction for multiple testing were used for *cis*- (1×10^{-5}) and *trans*-variants (5×10^{-8}). A more liberal threshold was used for *cis*-signals as by only testing variants in the *cis*-region rather than genome-wide, fewer tests were performed. As the protein levels are correlated, rather than correcting the significance threshold for 184 traits, we calculated the number of PCs required to explain 95% of the variance in the 184 protein levels and took this value as the number of independent traits tested, as done previously by Kettunen *et al.*[122]. We found that 85 PCs explained 95% of the variance in the levels of 184 protein in ORCADES (using the "prcomp" function in R), we repeated the analysis in CROATIA-Vis and again found that 85 PCs explained 95% of the variance. Our thresholds for significance were therefore 1.18×10^{-7} (Bonferroni $1 \times 10^{-5}/85$) for *cis*- and 5.9×10^{-10} for *trans*-associated variants.

In order to identify non-overlapping loci associated with a given protein, 1 Mb windows were created around every significant variant for that protein. Starting with the region with the lowest p-value, any overlapping windows were then merged, this was repeated until no more 1 Mb windows remained. To refine a list of non-overlapping loci that are associated with at least one of our 184 proteins we repeated this process of merging overlapping 1 Mb windows on the list of significant protein-locus associations.

Conditional Analysis

Conditional analysis was performed per protein using the --cojo-slct method from GCTA-cojo[150]. A minor allele frequency (MAF) filter of 1% and a p-value threshold of 1×10^{-5} were used. A random 10,000 unrelated genetically genomically British individuals from the UK Biobank were used as linkage disequilibrium (LD) reference.

Due to the particularly strong *cis*- signals we further filtered conditional variants, retaining per protein those with $r^2<0.001$. The criteria to limit heterogeneity for our primary variants were also applied to conditionally associated variants, retaining those with $I^2<30$ or if $I^2>30$ then at least 3 cohort level results that have consistent effect direction with the meta-analysis and nominally significant at the cohort level ($p<0.05$). As with primary associated variants, the threshold for significance corrected for multiple testing was 5 x 10^{-8}/85 for *trans* variants and 1 x 10^{-5}/85 for *cis* variants. Finally, akin to the primary variants, for each protein 1 Mb windows were created around each significant conditionally independent variant, with overlapping windows being merged, starting with the lowest p-value, until none are remaining.

Novelty of pQTL

To establish novelty of pQTL, we tested whether our 1,308 lead variants (or variants in LD, $r^2>0.5$, with our lead variants) had been previously associated with the relevant protein in 22 published GWAS or plasma protein levels (**Supplementary Table 29**).

Heritability

Estimates of total SNP heritability for each circulating plasma protein level were calculated as the sum of the contributions from two independent partitions of the SNPs: pQTL and the polygenic component. The pQTL component was calculated as the sum of the estimated variance explained (VE) in protein level by the lead variants of the primary pQTL. VE for each lead variant was estimated as $2pq\beta^2$ where β is the meta-analysis effect size, p is the effect allele frequency and $q = 1 - p$. The polygenic component was estimated using linkage disequilibrium-score regression (LDSC)[157] using variants present in the European 1000 Genomes Phase 3 Reference sample[294]. To ensure that variance explained by SNPs in LD with lead variants was not counted twice, variants within ±10 Mb of lead variants were excluded from calculations of the polygenic component.

Annotation of Significant Loci

Previously reported associations of all 1,093 of our significant variants and their proxies with an r^2>0.8 based on a 1000 Genomes Phase 3 European reference with GWAS traits, eQTL, proteins, metabolites and methylation QTLs were extracted from Phenoscanner v2[268,295], with a p-value threshold of 5 x 10^{-8}. Lead variants were also queried for evidence of being in a regulatory region using RegulomeDB[265].

For each of the *trans*-associated variants we defined a set of *trans* genes. These *trans* genes were any genes whose coding regions overlapped with a ± 500 kb window surrounding our significant variant using the Homo.sapiens[296] annotation package in R. For each of the *trans* genes we looked to see if the protein they encode have any known interactions with the protein we found it associated with using the STRINGdb[266] R package (database version 10). Similarly, for each *trans* gene we looked to see if they had any known pathways, gene ontology (GO) terms or publications in common with the gene encoding the protein we found them associated with. This was done using the KEGGREST[267] and org.Hs.eg.db[297] R packages.

Colocalisation of pQTL and eQTL

We looked up whether any of our significant variants had been previously reported as a significant eQTL (5% FDR (permutation-based)) in whole blood expression data from eQTLgen[180] and from 48 different tissues using the Genotype-Tissue Expression project (GTEx) v7.

SMR-HEIDI[183] was used to test whether a single causal variant is influencing gene expression and protein level due to either causality or pleiotropy, however it cannot distinguish between the two. We tested if ±500 kb regions flanking all 1,308 of our significant lead variants were associated with gene expression using four publicly available eQTL datasets: 48 GTEx tissues, both *cis* and *trans* eQTLs from eQTLgen[180], Westra *et al.*[181] and Cage[182]. Correction for multiple testing was carried out per eQTL dataset, with results with P_{SMR} passing Bonferroni correction for number of proteins vs probes and $P_{HEIDI} \geq 0.01$ considered significant.

In order to distinguish between causality and pleiotropy, we performed colocalisation using the "coloc.abf" function from the "coloc"[184] R package, with default priors. This approach simultaneously calculates posterior probabilities (PP) of eQTL and pQTL i) sharing a single causal variant and ii) being driven by two independent variants. For each of our *cis*-pQTL, the region within ±200 kb of the lead variant was tested for colocalisation with the gene encoding the protein in both the eQTLgen and 48 tissues from GTEx v7. We considered a PP>0.8 for the hypothesis that eQTL and pQTL share a causal variant as strong evidence of colocalisation and a PP>0.5 as likely to colocalise[116].

Genetic Correlations

The High definition likelihood[161] R package was used to estimate genetic correlations between the levels of our 184 proteins and the following cardiovascular-related traits using publicly available summary statistics (Download URLs in **Supplementary Table 33**): body mass index (BMI), coronary artery disease (CAD), chronic obstructive pulmonary disease (COPD), creatinine levels, Crohn's Disease, high density lipoprotein cholesterol (HDL), low density lipoprotein cholesterol (LDL), myocardial infarction (MI), Rheumatoid arthritis (RA), type II diabetes (T2D), total cholesterol, triglyceride levels and waist-hip ratio (WHR). To aid visualisation, proteins and complex traits were ordered using Euclidean distance-based hierarchical clustering with the hclust function in R.

Mendelian Randomisation

Instrument selection: For each protein, instruments were selected from genome-wide significant variants that passed the additional criteria of i) having a meta-analysis heterogeneity I^2<30 or if I^2>30, then ii) must have effect direction consistent with the meta-analysis in at least 3 cohorts and iii) be nominally significant (p<0.05) in at least three cohorts. These variants were then clumped for LD using an r^2 filter of 0.001 with the "TwoSampleMR"[170] R package. For each protein, MR was run using *cis* variants, with any variants within the HLA (chr6:29645000-6:33365000, build 37) and ABO (chr9:136131052-9:136150605, build 37) regions excluded from selection as instruments.

Primary MR Analysis: "TwoSampleMR" was used to perform Mendelian randomisation (MR) analysis. Protein level exposures were tested against 121 outcomes available in the MR-Base database[170] (the full list of outcomes tested is in **Supplementary Table 34**) using the fixed effects inverse variance-weighted (IVW) method. Outcomes were selected due to their relation to cardiovascular disease risk or immune-related disorders, given the proportion of immune system-related proteins in our set. For each outcome, summary statistics with the largest sample size and closest ancestry match with our GWAMA population were chosen.

Sensitivity analyses: To minimise the risk of heterogeneity between IVs influencing our results, only those without evidence of significant heterogeneity, using Cochran's Q test (q-value>0.05)[270], were considered. Additionally, to limit the effect of horizontal pleiotropy, we excluded protein-outcome MR estimates that had MR-Egger intercept significantly deviating from zero (P<0.05)[272,273]. We also performed MR analysis using MR-Egger, weighted median and weighted mode methods, which are more robust to horizontal pleiotropy[173,292] (**Supplementary Figure 53**). We also used the maximum likelihood (ML) method[177] which relaxes the assumption used by the IVW method, allowing both: uncertainty in the effect size of the IVs with the exposure and correlations between the genetic associations with the exposure and outcome. Consistency in causal estimates across MR methods with varying assumptions increases the chance of robust results.

Colocalisation: To distinguish causal relationships from confounding due to LD, we tested for evidence of a shared causal variant between each protein-exposure outcome pair using colocalisation. Variants within ±200 kb of each IV were tested for colocalisation with the overlapping variants in the outcome GWAS (extracted from MR-Base using the "associations" function from the ieugwasr R package). Only those with a posterior probability estimate of >0.8 for hypothesis 4 were considered further. Sample sizes for the 26 outcomes in the 59 protein-outcome associations passing 1% FDR heterogeneity and pleiotropy filters ranged from N=462,116 to N=173,082, for quantitative traits (N=119,731 to N=7,735 cases, for binary traits).

Bi-directional analyses: We tested for evidence of causal associations of the 121 outcomes on proteins using the IVW method. Protein-outcome pairs that had a causal effect estimate with $p<3.62 \times 10^{-6}$ (Bonferroni 0.05/13,810) were not considered further due to the potential for the estimate for the effect of protein on outcome to be influenced by reverse causality.

Drug Targets

The DrugBank Release Version 5.1.7[256] was used to see if the 11 proteins that had evidence of significant causal associations (P_{MR} passed 1% FDR & additional criteria described above) in our MR analysis are current drug targets.

5.3 Conclusion

The large sample size (Max N=26,494) of our genome-wide association meta-analysis (GWAMA) of the levels of 184 circulating plasma proteins, surpassing previous pQTL studies[115,116,120,149,153,254,255], facilitated the discovery of 1,308 protein-lead variant associations, with 687 being novel. We showed that between 8-37% of *cis*-pQTL overlap with published eQTL and therefore may act to influence the level of proteins circulating in the plasma by affecting the transcript of the encoding gene. We showed that 66 protein levels share genetic architecture with health-related risk factors such as BMI, WHR, Creatinine levels and HDL cholesterol levels and cardiometabolic diseases such as type 2 Diabetes mellitus, coronary artery disease (CAD) and history of myocardial infarction (MI). Finally, we inferred 20 causal protein-outcome relationships: replicating RCTs[178,274], replicating previous MR findings from pQTL studies[149,178] and novel causal relationships. These results allowed us to highlight IL2-RA as a potential candidate for drug repurposing to combat Crohn's Disease (CD) and IL-27 and TNFRSF14 as novel therapeutic targets.

The analysis presented in this chapter highlighted avenues for further research, for example, our finding that, of our lead variants, 8.8% had been previously reported as associated with circulating metabolite and 74.6% with DNA methylation levels. This emphasises the need for future research focussed on untangling the nature of the relationship between protein levels and the regulation of these other omics layers, particularly DNA methylation.

An issue emphasised by this study is the growing need for large eQTL and indeed pQTL datasets derived from tissues other than blood. We were limited in our ability to detect colocalisation of pQTL with eQTL due to the fact than many of the proteins of interest are expressed and secreted from other tissues, not blood where both our protein measures and the well powered publicly available eQTL dataset from eQTLGen[180], were derived from. Despite GTEx having data for multiple tissues, it is limited by its small sample size. Together this means that there may be pQTL that act by affecting the transcript level of the encoding gene, but only in the relevant tissue, meaning we were unable to detect them in our blood focussed analysis. The availability of high-powered eQTL and pQTL datasets (as this limitation applies to the protein levels measured as well) from multiple tissues would substantially increase the field's ability to form a more comprehensive understanding of where proteins are expressed, how their levels are altered by disease and the tissues or pathways that could be targeted by interventions.

This work added to the field with our GWAMA discovering of 687 novel protein-lead variant associations and despite the broad exploratory Mendelian randomisation analysis, that primarily aimed to demonstrate the power of our pQTL as instrumental variables, we were able to infer causal relationships for 20 protein level-health outcome pairs. Further we reported a candidate for drug repurposing (mentioned above) due to the inferred causal relationships and highlighted 2 potential novel therapeutic targets for Crohn's disease, Inflammatory bowel disease, Multiple sclerosis and Ulcerative colitis.

Obviously, further rigorous assessment is required to determine if these proteins would make suitable candidates for therapy, however our ability to suggest them in the first place demonstrates the power of our pQTL as IVs. The resource of GWAMA summary statistics produced in this chapter have the potential to aid future researchers to answer highly targeted research questions.

Chapter 6: Discussion

In this thesis, I sought to take advantage of the extremely broad range of omics assays available in ORCADES to improve our understanding of the underlying biology of ageing and disease. Specifically, I aimed to investigate the relationships between these omics measures and ageing, health-related risk factors and future health outcomes.

6.1 Summary of Findings

6.1.1 Biological Ageing Clocks

I took advantage of the ORCADES cohort, unique in terms of its broad annotation, comprising 9 different omics assays: DNA methylation, PEA Proteomics, UPLC IgG Glycomics, NMR Metabolomics, MS Metabolomics, MS Complex Lipidomics, MS Fatty Acid Lipidomics, DEXA scans and a collection of common clinical measures which I termed Clinomics. Using this resource, I was able to perform the most comprehensive comparison of nine different omics assays as potential sources of biomarkers of biological age (BA). Previous comparisons of multiple omics ageing clocks had focussed on epigenetic clocks or those built from traditional risk factors and frailty[93–95,98,99]. While there are numerous publications detailing multiple epigenetic clocks, whose effects have replicated across studies, few have investigated how multiple other omics assays compare. I showed that is possible to construct an accurate chronological age (chronAge) predictor with 8 of the 9 omics assays tested (correlation of OCA with chronAge ranged 0.66-0.97). Moreover, despite the small sample size, the DNA methylation, PEA Proteomics, Clinomics and UPLC IgG Glycomics clocks trained in the isolated population ORCADES, replicated in independent populations, indicating their validity.

I replicated inter-clock omics clock age acceleration (OCAA) correlation patterns reported by previous studies[93–95,97]. I also investigated further and for the first time showed that, as well as overlapping 94% in the variance they explain in chronAge, OCAAs built using different omics assays overlap more than would be expected if they were independently sampling from a complete set of latent predictors.

In order to determine if OCAAs were capturing something biological, I tested whether they were tracking health-related risk factors, that are themselves biomarkers for disease. In addition to finding that OCAAs were significantly associated with the health-related risk factors: total cholesterol, Framingham Risk Score, C-reactive protein and systolic blood pressure, I found that OCAAs were also prognostic of incident disease, using hospital admission as a proxy. Clinomics OCAA was prognostic of diabetes mellitus, group E metabolic disorders, hypertensive diseases and chronic lower respiratory diseases, while DNAme Horvath CpGs and NMR Metabolomics OCAAs were prognostic of acute lower respiratory infections and other respiratory diseases principally affecting the interstitium respectively. As well as these 6 statistically significant (FDR<10%) OCAA-disease block associations, there was also strong evidence of enrichment of association of OCAA with incident disease collectively across all tests (20% were nominally significant p<0.05), suggesting that I was underpowered and that with a larger sample size, may have found more significant signals.

A key finding was our estimate that one year of OCAA has an effect of 0.46/0.45 years of chronAge on risk factors/disease incidence. No previous studies have quantified the proportion of their OCAA that may be capturing noise compared to potentially true underlying BA. In fact, our finding highlights how serious an issue noise is in omics ageing clocks trained on chronAge and suggests different approaches may produce more effective biomarkers of BA.

My findings that the PEA Proteomics OCAA appears to track specific risk factors while UPLC IgG Glycomics and DNA methylation OCAAs appear to capture generalised ageing, indicate that there may be multiple distinct (sometimes organ-specific) BAs as well as one underlying measure that encapsulates overall body ageing. The observation that 8 out of 11 omics ageing clocks had hazard ratios <1 for risk of Melanoma and other malignant neoplasms of the skin, in contrast with the overwhelming trend of OCAAs being associated with increased risk of incident disease, further supports the hypothesis that there may be multiple potentially organ-specific BAs.

I showed that it is possible to reduce the dimensionality of the omics predictors presented to the clock construction algorithm and produce OCAAs that achieve the same performance as those presented with the full set of available predictors per assay. First, I showed that simply reducing the number of biomarkers available for model inclusion to a core set, produced comparable correlations between omics clock ages (OCAs) and chronAge to those achieved with my standard clocks. While, as mentioned previously, this has been shown to be possible for PEA Proteomics by Enroth *et al.*[54], this has not been shown to be the case systematically across 9 different omics assays. Second, I showed for the first time that OCAA derived from clocks built using a few principal components of omics assays were as prognostic as those presented with all available features. These are key findings if the purpose of building omics ageing clocks are for them to be clinically useful.

Given the generally modest effect sizes of OCAA on incident disease found in both my analysis and in the literature[93–95], I conclude that the shift of focus of ageing clocks from chronAge predictors to those trained on mortality and or morbidity is the best approach to derive OCAAs that capture underlying BA and are likely to be more prognostic of health outcomes.

6.1.2 Omics Biomarkers of Incident Disease

In addition to assessing OCAA measures as potential biomarkers for health-related risk factors and incident disease, we wanted to take advantage of the extensive assays available in ORCADES and investigate whether these omics measures are themselves biomarkers of diseases and risk factors directly.

Even with our limited power due to the small sample size, we found 8,526 significant (5% FDR) biomarker-outcome associations between 2,686 single omics biomarkers and 54 outcomes (incident disease or risk factor). We also found evidence of enrichment of associations (12.8% of tests had a p-value less than 5%). The majority of significant biomarkers were associated with more than one outcome, with only 23.4% being outcome specific (**Supplementary Figure 43**). This mirrors the finding in Pietzner *et al.*'s, study of MS Metabolomics biomarkers of disease, that 65.6% of significant metabolites were associated with multiple outcomes.

Starting with risk factors, by combining biomarkers into multivariable omics scores we found 69 omics scores significantly (5% FDR) associated with 10 health-related risk factors (**Figure 15** & **Figure 17**). Clinomics and UPLC IgG Glycomics scores were associated with the most risk factors, although the UPLC IgG Glycomics with low effect sizes. This perhaps is unsurprising for Clinomics, as in the biological ageing clocks analyses, it outperformed other omics assays however, as discussed this could be due to its constituent predictors being highly correlated with the risk factors being considered.

Turning now to disease blocks, we found fewer omics scores prognostic of subsequent incident disease blocks, with 12 significant (5% FDR) omics score-disease associations (**Figure 15** & **Figure 16**). Again, Clinomics scores were prognostic of the most incident disease blocks: Metabolic disorders, Diabetes Mellitus, Obesity, hypertensive disorders and ischaemic heart disease. Conversely, we found 5 different omics scores associated with incident diabetes mellitus.

A key finding upon investigating the relative importance of the biomarkers included in the omics scores, was that only a handful of biomarkers were actually driving each score. Four Clinomics scores were dominated by one biomarker, with between 98.3-79.5% of variance in Diabetes Mellitus (glucose), obesity (weight), ischaemic heart disease (glucose) and FRS (systolic blood pressure) scores being explained. Similarly, 4 or fewer biomarkers contributed greater than two thirds of the variance explained in four significant omics scores (DEXA, MS Metabolomics, PEA Proteomics and Mega-omics) for Diabetes mellitus. This pattern, suggesting a few biomarkers are necessary for an effective score, mirrors the results in **Chapter 3: Biological Ageing Clocks**, where we showed that substantial sub-setting of biomarkers created OCAs with comparable correlations with chronAge to models presented with all available biomarkers (**Figure 7**). This is of course ideal for scores if the aim is for them to potentially be clinically useful.

Similar to our findings in **Chapter 3: Biological Ageing Clocks**, where the Mega omics clock is predominantly (26.6%) composed of PEA Proteomics biomarkers, we found that proteins dominated the Mega omics score for Diabetes mellitus (**Figure 21**). With the 3 proteins that together contribute 78.7% of the PEA Proteomics score for Diabetes mellitus contributing 63.4% of the variance explained in the Mega omics score. The levels of these three proteins: MVK, CES1, and ADGRG1 along with the levels of MATN3 are four out of the top 5 largest contributors of variance explained in the Mega omics diabetes mellitus score. Our analysis therefore highlights these proteins as potential biomarkers of diabetes mellitus.

Overall, despite our limited power, we demonstrated that both individually and combined in multivariable models, omics measures are effective biomarkers of health-related risk factors and subsequent incident disease. Further, we observed that Clinomics scores appear to be more effective than those derived from high throughput high dimensional platforms, at least for scores trained in the modest sample size we had available.

6.1.3 GWAMA of Plasma Protein Levels

By combining the proteomics data available in ORCADES with that of 17 other cohorts, I performed the largest genome-wide association meta-analysis of the levels of 184 plasma proteins to date. With the maximum sample size of N=26,494, we reported 1,308 significant protein-variant associations, 687 of them novel, reporting the highest proportion of proteins tested (92.7%) with a significant *cis*-pQTL, compared to previous pQTL studies[116,149].

Mirroring results from Folkersen *et al.*[149], we showed that genetic architecture varies across proteins, with some such as NEMO and GH, being extremely polygenic and others such as IL6-RA and CHIT1 being almost monogenic (**Figure 23**). For those proteins whose genetic architecture is characterised by a small number of loci, it is often due to extremely strong *cis*-signals, for example, CHIT1 where the lead variant rs12141375:A is estimated to explain 32.7% of the variation in the protein level. This is in stark contrast to most polygenic traits where individual variants explain low proportions of trait variation.

In the search for potential mechanisms of action for our pQTL on circulating plasma protein levels, we showed that: between 8.2-36.5% (method dependent) of pQTL overlap with eQTL, suggesting these variants influence the transcript level of the encoding gene thus affecting circulating protein level. The fact that this proportion of pQTL-eQTL overlap is lower than reported in previous pQTL studies is discussed in detail in **Chapter 5: Genome-wide Association Meta-analysis of 184 Plasma Protein Levels**.

However, we also found evidence that our pQTL are involved in other types of regulation: 132 out of our 1064 lead variants show evidence of transcription factor binding (**Supplementary Figure 51**) and 74.7% of our pQTL have been previously reported as significantly ($p<5 \times 10^{-8}$) associated with DNA methylation levels mirroring findings by Huan *et al.*[290]. These results highlight the variety of mechanisms by which genetic variants may act to influence the levels of protein circulating in the plasma and emphasise the interconnected nature of different omics layers with respect to function.

We also demonstrated links between plasma protein levels and disease: 547 of our lead variants have been reported as significantly ($p<5 \times 10^{-8}$) associated in complex trait GWAS (**Supplementary Figure 51b**) and we found that 66 proteins share genetic architecture with cardiovascular outcomes and risk factors such as BMI, Type 2 diabetes mellitus, total triglyceride levels, coronary artery disease (CAD) and myocardial infarction (MI) (**Figure 24**). These findings highlight the interplay between plasma protein levels and complex traits and prioritises protein-disease relationships for future highly targeted research.

Similar to previous pQTL studies, we used Mendelian randomisation (MR) to investigate potential causal relationships between the levels of proteins circulating in the plasma and diseases[116,149,153,178]. Using our *cis*-pQTL as instruments we first, recapitulated the well documented association between increased levels of PCSK9 and increased levels of total and LDL cholesterol[178,274], replicating findings from randomised control trials (RCTs). Second, we replicated MR findings from previous pQTL studies, namely the negative MR effects of PlGF on coronary heart disease risk[149] and CXCL16 on diastolic blood pressure[178]. Third, for genes encoding 4 of our proteins (*IL2RA*, *PRSS8*, *TFPI* and *IL27*) that have been previously associated with specific disease outcomes in GWA studies, we inferred causal relationships between the levels of proteins circulating in the plasma with these phenotypes for the first time. Finally, we inferred 10 novel causal protein-disease associations (**Figure 25**), demonstrating the discoveries possible with our well powered pQTL instruments.

Our MR analysis inferred causal relationships for 11 of our proteins, 6 of which are already current drug targets (**Supplementary Table 36**). We highlight a drug that targets one of our proteins as a potential candidate for repurposing in light of our inferred causal effect of genetically increased levels of IL2-RA on Crohn's disease (CD). Finally, we highlight two proteins as potential novel therapeutic targets: IL-27 for CD and TNFRSF14 for multiple sclerosis, ulcerative colitis and inflammatory bowel disease. Together these findings exemplar the discoveries possible and the potential of this study as a resource to drive highly targeted future research.

6.2 Strengths

The two strengths of the research presented in this thesis are the range of omics assays available in the ORCADES cohort and the combined sample size possible due to collaboration with the SCALLOP Consortium.

First, given the range of data available for the ORCADES cohort, both in terms of the range of omics assays and the 10-year follow in the form of electronic health records (EHR). I was able to take advantage of the breadth of omics assays in both the search for biomarkers of biological age, subsequent incident disease and health-related risk factors. It provided the unique opportunity to test 3,302 different biomarkers in the same set of individuals and allowed me to investigate biomarkers capturing more areas of biology than previous studies.

Second, the total sample size for the GWAMA of plasma protein levels and the 17 SCALLOP Consortium cohorts contributing summary statistics, was a major strength. Without this collaboration I would not have had sufficient power to discover novel: pQTL, causal relationships with disease and therapeutic targets. I reported a significant *cis*-pQTL in a higher percentage of proteins tested than previous plasma protein GWAMAs[116,149], most likely possible due to larger sample size. Similarly, the increase in proportion of *trans*-pQTL discovered compared with previous studies[116,149] can be explained by my increased power to detect *trans*-signals, as was the case with studies by Folkersen *et al.* when progressing from a single cohort to a GWAMA of the same set of 90 proteins[115,149]

6.3 Limitations & Future Work

6.3.1 Sample Size

Despite being a strength in the plasma protein level GWAMA, sample size in general was a limitation in the research presented in this thesis. The fact that only approximately 1,000 individuals in ORCADES were annotated with all 9 omics assays and that this number was further reduced by creating a complete non-missing sample, limited power in the biological ageing and biomarkers of incident disease analyses.

This small sample size increased the risk of multivariable models being overfit however, the majority of omics ageing clocks showed consistent OCA-chronAge correlations between training and testing samples. PEA Proteomics, DNA methylation, UPLC IgG Glycomics and Clinomics clocks also replicated in independent populations. Together these results suggest that overfitting was successfully avoided in the omics ageing clocks analyses. There was some evidence of potential overfitting in the biomarkers of incident disease analysis however, LASSO regression was chosen over elastic net for construction of multivariable disease scores due to more consistent effect estimates of score on outcome between training and testing samples, to limit this issue influencing the results.

In an attempt to limit the power lost while removing missing values, I used an approach that maximised both the number of samples and omics predictors available for selection for model inclusion.

The low numbers of cases for incident disease blocks, as mentioned previously, limited power to build omics disease scores as well as the ability to assess how prognostic OCAAs were of incident health outcomes. Additionally, due to the low number of deaths recorded amongst the individuals in ORCADES with omics measures (as yet), I was unable to investigate these omics as potential biomarkers of mortality or assess whether OCAAs were prognostic of mortality.

In order to overcome the sample size limitation in the future, these analyses should be repeated in samples with either: increased numbers of individuals with omics assays, incident disease cases or recorded deaths. Maximising the number of complete non-missing samples with omics measures for multivariable model construction, both clock and score, by imputing missing omics values using methods such as predictive mean matching[298] or k-nearest neighbours[299], would increase the sample size. Additionally, collaborating with cohorts with overlapping omics assays and meta-analysing would increase the power of future analyses.

6.3.2 Multiple Testing

The decision to capitalise on the breadth of omics assays also creates the inherent limitation of multiple testing. This meant that while finding numerous suggestive results, I was able to draw few conclusions based on formal statistical significance.

I took active steps to limit the impact of multiple testing when assessing the effect of omics scores on incident diseases and health-related risk factors. I used association results in the training sample to estimate which omics score-outcome pairs had sufficient power to detect associations in the testing sample passing a 5% FDR significance threshold.

Any of the techniques to overcome limitations of sample size mentioned above will also aid to circumvent issues created by multiple testing. An alternative would be to select which tests to perform based on prior evidence possibly from pilot studies, the literature or based on biological function of the biomarkers.

6.3.3 Sex Differences

All of the analyses presented in this thesis considered both sexes together in the study populations. Sex was explicitly accounted for in all of the analysis by being included as a fixed effect covariate when correcting raw omics measures. For omics ageing clocks and omics scores for risk factors and incident disease, this is described in **2.3.1 QC of ORCADES Omics Data** and for GWAS of plasma protein levels, this is described in **2.5 Genome Wide Association Studies**. However, we did not investigate sex-specific effects.

For the multi-omics analyses in ORCADES, both ageing clocks and biomarkers of disease, the decision to not perform sex stratified analysis was taken due to the small sample size, both in terms of individuals with omics and incident disease cases, meaning doing so would most likely be underpowered. For the GWAS of plasma protein levels this was not investigated simply to limit the scope given the number of other analyses planned.

Sex differences have been demonstrated in several areas, including GWA studies of a diverse range of phenotypes[300–302] and are worth investigating. Future work should seek to repeat the association analysis of OCAA and omics scores of health outcomes with risk factors and incident disease, stratifying by sex and assess whether the estimated effect directions are consistent between sexes. Similarly, GWAS of plasma protein levels should be repeated separately for each sex and results compared to the combined analysis to see if there any pQTL that have sex specific effects. This would massively improve our understanding of underlying biology, as sex is a factor that is too often not investigated[303]. If conclusions are drawn based on combined data with the assumption that the conclusion holds true for both sexes, when in fact this may not be the case[304], this could have dire consequences, particularly if these conclusions inform widespread medical treatment.

6.3.4 Replication

While I replicated 5 omics ageing clocks trained in ORCADES in independent populations, future work should seek to replicate the remaining: NMR Metabolomics, MS Metabolomics, DEXA, MS Complex Lipidomics and Fatty Acid Lipidomics clocks. Similarly, maximising sample size in the protein level GWAMAs rather than using a discovery and replication approach[305], meant I was able to report 687 novel protein-variant associations, however I was unable to validate them in an independent replication sample and thus future analysis should seek to do so. Making the GWAMA summary statistics publicly available upon publication, as planned, will facilitate other studies to replicate the novel pQTL.

Both DEXA and NMR Metabolomics clocks trained in ORCADES showed considerably lower correlations between OCA and chronAge in the UK Biobank and the Estonian Biobank respectively than in the ORCADES training and testing samples (**Supplementary Figure 28**). This could be due in part to differences in underlying characteristics, environment or patterns of behaviour between the isolated population of the Orkney Islands and general population samples from the UK and Estonia. For example, the range of occupations of participants and climate between ORCADES and the UK Biobank differ drastically. For omics ageing clocks or disease risk scores to be generalisable they must be effective across populations, therefore training models using a meta-analysis containing individuals from a diverse range of populations would be ideal for such analyses.

Further, all of the study populations used for analysis in this thesis are predominantly of European ancestry. This is an issue more generally in research, with an overwhelming proportion of human studies using participants of European ancestry. This homogeneity hinders research as it leads to bias and population specific results that will not hold true for millions of individuals. However, recently there has been a drive to use study populations from more diverse ancestries.

This is a particularly well-documented issue in terms of genetic association studies (GWAS). Differing allele frequencies between populations can result in genetic variants that are significantly associated with a trait of interest in one population not being found in another[306]. The finding that polygenic risk scores are considerably more predictive in individuals from the population that the scores were trained in compared to their performance across ancestries, unless the training population contained a mix of ancestries, also illustrates this issue[307]. Further, LD patterns differ across ancestries meaning that, when considered together, they increase our ability to determine the causal variants in an LD block and therefore allow more informed estimates of the functional consequences of these causal variants[308]. As directions of effects of trait-associated genetic variants tend to be consistent across ancestries[309], methods of trans-ethnic GWAS that stratify by ancestry or adjust for admixture[310], increase power to detect associated genetic variants. Heterogeneity in estimates of effect sizes for genetic variants across populations may arise due to differences in: disease prevalence, disease treatment, environmental exposures, lifestyle and diet[311]. However, methods to perform trans-ethnic GWAS that account for these forms of heterogeneity, minimising power loss given large enough sample sizes, have been developed[312,313].

Future research should seek to replicate our reported pQTL in populations with non-European ancestries or use our summary statistics to perform trans-ethnic GWAS. This is particularly important if the causal relationships between plasma protein levels and disease, inferred using pQTL instruments, are used to inform therapies. Then it is essential that these relationships hold across different populations, otherwise conclusions could be drawn that lead to the misuse of therapies or incorrect diagnosis that may have negative consequences for millions of patients.

6.3.5 Rare Variants

Like the majority of genome-wide association meta-analyses, the GWAMA of plasma protein levels focussed on common variants (MAF>1%)[314]. Despite not setting a MAF filter for the meta-analysis itself, we applied additional criteria that variants had to meet – i) being measured in ii) having effect directions consistent with the meta-analysis and iii) having a nominal p-value – in at least three of the contributing cohorts in order to be considered significant, that effectively ruled out rare variants. Further MAF thresholds of 1% were used for downstream analysis.

Historically GWAS have focussed on common variants due to the common disease common variant hypothesis[315–318] and on a more practical note, due to the fact that the majority of cohorts rely on genotype or imputation data which are poor at capturing rare variants[319,320]. However, given the large amount of missing heritability not accounted for by GWAS of common variants it was suggested that rare variants with large effects could be a source of this heritability, which has been shown to be the case for a number of polygenic traits[321]. As any variant with a large deleterious effect on fitness will be eliminated from the population by mutation-selection balance, those with large effects are likely to be rare or recessive[322]. Additional issues arise in the use of rare variants: the rigorous quality control necessary[323,324], the fact that statistical approximations used for common variant analysis assume large sample sizes which may not hold for rare variants[325] and exacerbate the issue of multiple testing[326]. Despite these issues, approaches have been developed that consider multiple variants simultaneously to overcome some of these limitations and facilitate association testing using rare variants[327]. These variants are potentially valuable sources of information and should be used to uncover missing heritability.

The study, by Gilly *et al.*, used whole genome sequence (WGS) data to find rare variants associated with circulating plasma protein levels[153], however with a sample size of N=1,328 the power is limited. Increasing the number of cohorts with both proteomics and WGS or exome-wide sequence (EWS) data could be the way forward. Future work should seek to perform dedicated analysis on plasma proteomics using WGS or EWS in larger sample sizes, possibly by meta-analysing to increase the power to detect rare variants with large effects. Given that rare variants are rarely considered in large scale studies, addressing this gap would allow us to form a more comprehensive picture of the genetic regulation of plasma protein levels.

6.3.6 Publicly Available Ageing Clocks

By constructing ageing clocks from scratch using each of the 9 omics assays available, we were only able to compare effect size estimates with previously published studies. Rather than directly compare the performance of published clocks in our sample.

For DNA methylation, our clocks were based on the Hannum[55] and Horvath[66] published epigenetic clocks. Given how established these clocks are and the number of studies validating their ability to predict chronAge in the literature, there would have been little point in training rival DNA methylation clocks in our much smaller sample. However, due to data security protocols prior to GDPR[328], the subset of the CpG sites available in ORCADES were used to construct our DNA methylation clocks rather than uploading raw individual level data to Horvath's online calculator (http://dnamage.genetics.ucla.edu/).

Due to the updated regulations outlined in GDPR increasing the security around individual level data, future work should seek to use available online calculators for the Hannum and Horvath epigenetic clocks and van den Akker *et al.*'s for their NMR Metabolomics clock[86] (https://metaboage.researchlumc.nl/) as well as calculating GlycanAge from Krištić *et al.*'s study[87] in ORCADES. This would provide the opportunity to assess how these published BA measures replicate in an additional population and compare their performance to the naïve clocks trained in ORCADES as well as with each other in the same set of individuals.

We also limited our analysis to clocks trained on chronAge, as the aim of our study was to characterise the properties of these OCAAs, which have been understudied across multiple omics due to the rise of second-generation clocks. However, future work could compare OCAAs trained on chronAge with second-generation ageing clocks such as GrimAge[97] and DNAm PhenoAge[96] in the same sample (again using http://dnamage.genetics.ucla.edu/), in order to systematically assess the differences in properties of first and second generation clocks.

6.3.7 BA vs Risk Factors

When assessing if my OCAA measures were prognostic of incident disease, I reported effect sizes scaled by the effect of chronAge on disease (**Figure 10**), to illustrate the effect of OCAA beyond chronAge. A potential further step would have been to investigate how OCAA compare with commonly used clinical risk factors of incident disease, as carried out by previous studies[95,203].

This was not an area explored in our analysis explicitly, we did however include the Clinomics OCAA in our comparison of the prognostic ability of OCAAs, which is a multivariable model entirely built of common clinical risk factors. As our primary aim was to characterise OCAAs derived from multiple omics clocks trained on chronAge and within this, establish if they were prognostic of future health outcomes, we focussed on their comparison with chronAge.

However, testing whether models fitting both OCAA and clinical risk factors together, outperform models fitting clinical risk factors alone, in predicting incident disease is an exciting avenue for further research. On the one hand, if OCAA were found to add predictive value beyond clinical risk factors, this would evidence the practical utility of OCAAs trained on chronAge as biomarkers of ageing. On the other, if OCAA were found to not contribute additional predictive ability or indeed contribute a small amount, this would concur with the suggestion that clocks trained on outcomes other than chronAge are the way forward[95–97].

6.3.8 Tissue- or Organ-Specific BA

As mentioned in **Chapter 3: Biological Ageing Clocks**, both the work presented in this thesis and the majority of omics ageing clock studies in the literature assume that there is one single underlying BA that captures an individual's ageing, risk of incident disease and mortality[54,55,66,87,96,97]. However, this may not be the only hypothesis to consider, what if different organ systems or tissues age at different rates? Is it possible that an individual could have a cardiovascular age that is different from their immune system age or their musculoskeletal age?

My analysis of multiple omics ageing clocks found that certain clocks, UPLC IgG Glycomics and DNA methylation, appear to track generalised ageing as they were prognostic (positive association) of multiple incident diseases, whereas others such as PEA proteomics appeared to track specific risk factors (**Supplementary Figure 34**). Together with the observation that multiple OCAAs were negatively associated with incident malignant neoplasms of the skin, in contrast to the trend of positive OCAA-disease block associations, these results suggest that there be more than one type of BA.

The only publication to touch on this, Lehallier *et al.*, constructed multiple proteomics clocks based on subsets of proteins determined on their: functions as described in the literature, previously published associations with age and pathway enrichment scores[91]. However, as the clocks based on pathway enrichment showed lower correlation with chronAge than models fitting more proteins, their properties were not further investigated.

Future work should seek to further investigate the potential of organ system- or pathway-specific ageing clocks to provide a more comprehensive picture of how the body ages. Additionally, the use of omics measures derived from multiple tissues, rather than blood, which has been predominantly used by the field to date, would facilitate a more accurate understanding of the underlying biology of ageing.

6.3.9 Risk-based Age

Given the known limitations of BA clocks that are trained on chronAge, an alternative approach has been suggested (Fischer *et al.* Unpublished). Fischer *et al.* found that estimated effect directions of associations between OCAA's trained on chronAge on risk factors, often oppose the expected (and observed) hazard ratios of those same risk factors on mortality. For example, previous cancer diagnosis and previous MI were associated with lower OCAA despite their hazard ratio of >1 for risk of all-cause mortality. These findings suggest the presence of confounding, possibly by chronAge, for risk factors such as smoking behaviour or risk factors having an effect on mortality that is not captured by omics biomarkers.

In order to overcome these limitations, Fischer *et al.* considered the definition of BA in terms of risk. As discussed previously the BA of an individual is defined as the chronAge of an average individual in the study population that has the same risk of future health outcomes and functional capacity as the individual of interest has at their current chronAge. However, if put in terms of "risk" meaning risk of death as captured in all-cause mortality then this risk-based BA could be calculated using survival functions. The risk-based BA for an individual i, is the chronAge of the average individual in the cohort when their risk of mortality equals individual i's current risk of death.

Preliminary results show that risk-based BA outperforms omics clocks trained on chronAge. Framing BA in terms of risk provides a clearer interpretation than OCAA and the direct hazard ratios of risk factors on mortality. This is an extremely promising direction for future research and has the potential to change the way the field thinks about BA.

6.3.10 Hospital Admissions

Using first hospital admission for groups of ICD10 codes as proxies for incident disease has several limitations.

First, using hospital admissions limits the scope of our investigation to only those conditions that are severe enough to require admission to hospital. On the one hand, this effective filtering by severity ensures that effect estimates or hazard ratios are less likely to be confounded by severity. On the other hand, it will not be an accurate reflection of all incident cases of disease. This approach will also fail to capture diseases that tend to be treated in the community such as dementia and multiple sclerosis. The use of GP records in conjunction with hospital admission data in future studies would help overcome this limitation.

Additionally, hospital admission itself for any purpose is confounded due to differences in lifestyle, behaviour and socio-economic status[329]. Similarly, variation or changes in admission and screening policies across hospitals over time will influence our results using this approach, as will coding inaccuracies often due to manual data entry.

Second, limiting the analysis to the first admission for a particular disease block loses information about whether the disease recurs. Again, the use of GP records would help to overcome this issue, as being able to predict recurrence or relapses is of great value in terms of precision medicine.

Third, considering groups of related ICD10 codes together as disease blocks rather than assessing individual diagnoses is a further limitation. This approach had practical advantages in that it limited the multiple testing burden and pooling multiple diseases increased the number of cases per disease block available with our limited numbers. By doing this however, we excluded subsequent diagnosis for any of the other diseases in a disease block after the first diagnosis, therefore losing disease cases.

Similarly, we were unable to assess multi-morbidity within disease blocks for this reason. Finally, this meant that we were not able to assess individual diseases, this is an issue as despite ICD10 chapters containing broadly similar disorders there is still heterogeneity. The disease block E10-E14 is an ideal example of this, as it contains both type 1 and type 2 diabetes mellitus. Where type 1 is a chronic autoimmune disease, the bodies inability to produce insulin, is often diagnosed young, requires lifelong management with unknown cause[330]. In contrast, type 2 is much more common, develops in later life, results in reduced insulin levels or receptors no longer recognising insulin with obesity and low physical activity as known risk factors[331]. These are extremely different conditions, likely to have differing disease aetiologies and considering them in the same disease block limits our ability to draw conclusions. Well powered future work should seek to consider diseases individually, this would increase our understanding of the underlying biology of these conditions in a way that considering chapters based on ICD10 codes does not.

6.3.11 Assessment of Prediction Accuracy

Despite creating models in both chapters 3 and 4 that predict subsequent incident disease, using omics clock age acceleration (OCAA) measures and omics biomarkers respectively, I did not formally assess their prediction accuracy.

As in the omics ageing clocks analysis, the aim was to determine if the OCAAs were associated with incident disease and risk factors, and if so, how this compared to chronological age. This was done by assessing to what extent OCAAs were associated with these health outcomes over and above chronological age. As one of the main aims of the omics ageing clocks chapter was to characterise what OCAAs are measuring, are they just capturing chronological age? Or are they capturing aspects of an underlying biological age?

An additional reason for this focus on the comparison with chronAge, was due to the limited power in ORCADES. Both the limited number of samples with omics assays and the low number of cases in several disease blocks, means that the likelihood of producing an OCAA that would be prognostic of subsequent incident disease, with a high prediction accuracy in independent cohorts, is low.

Similarly for the analysis in chapter 4, where I constructed omics scores predicting subsequent incident disease, the low numbers of disease cases (further reduced by the training-testing split and further still during 10-fold cross-validation in the training sample), limited my ability to make prediction models that would be suitable for use in independent populations. For this reason, I restricted my analyses to those scores that were significantly (5% FDR) associated with health outcomes, highlighting biomarkers that were selected for model inclusion and investigating which of these biomarkers are contributing most to the scores.

For both of these analyses (to a greater extent for the analysis in chapter 4), there were insufficient case numbers in the testing sample to determine a meaningful estimate of prediction accuracy that would be a genuine reflection of the model's performance in unseen data.

Future studies with sufficient power, should assess the classification accuracy of such models using methods such as receiver operating curves (ROC) and area under the curve (AUC), to establish the ability of their models to distinguish between cases and controls[332]. Further, to account for the prevalence of subsequent incident cases for each disease, future studies should calculate the negative predictive values (NPV) and positive predictive values (PPV) to indicate the proportions of predictions that are true negatives and true positives respectively, for each omics score or OCAA on each outcome[333]. Calculation of measures of prediction accuracy would therefore allow the comparison of novel prediction models to those published.

6.3.12 Colocalisation & Mendelian Randomisation

Limitations of the colocalisation and Mendelian randomisation analyses performed were mentioned in the discussion section of **Chapter 5: Genome-wide Association Meta-analysis of 184 Plasma Protein Levels** however, I will highlight three opportunities for future work worthy of further discussion.

First, that the colocalisation of discovered pQTL with eQTL should be investigated using the recently proposed sum of single effects (SuSiE) regression method[289]. As this approach is not constrained by the assumption that there is only one causal variant in the region being considered, which as mentioned previously is potentially violated by broad peaks observed in *cis*-regions that contain multiple genes.

Second, in addition to sample size used for pQTL discovery, power for the colocalisation analysis also depends on the sample size of the eQTL study. As mentioned previously, this is small for most publicly available multi-tissue eQTL datasets such as the GTEx data used in my analysis. The sizes of whole blood eQTL datasets are progressively increasing, with the eQTLgen data used reaching an N of ~30,000[180]. However, these sample sizes are being achieved only for *cis*-eQTL, it is still extremely difficult to identify *trans*-eQTL due to the increased multiple testing burden[334]. Until approaches to overcome this issue are developed studies will have to remain focussed on colocalisation with *cis*-eQTL.

Third, we restricted the Mendelian randomisation (MR) analysis to *cis*-IVs only, as they are biologically more directly linked with the protein level and less likely to be affected by horizontal pleiotropy, therefore minimising the risk of violating the first and second assumptions of MR. We performed additional sensitivity analyses (described previously) in order to maximise the stringency of our results. For inferred causal relationships meeting these criteria, we assessed what difference repeating the MR analysis using both *cis* and *trans* IVs had on results (**Supplementary Figure 53**). We observed that estimated MR effect directions were not always consistent between *cis* and *trans* (pan) and *cis* only, suggesting that the effect of including *trans* IVs should be further investigated. Future studies dedicated to untangling the consequences of including *trans* IVs are required to determine if this is a robust approach for MR analysis at all, given the difficulty of horizontal pleiotropy in this scenario. A systematic investigation into this issue was beyond the scope of our exploratory analysis, however, it would be extremely valuable to the field given the rate of new MR studies.

6.3.13 Potential Data Leakage

In machine learning, data leakage is the phenomenon where information from outside the training dataset is used to create the model. This is an issue as this additional information can give the model an unrealistic advantage to make better predictions, leading to overestimation of the performance of the model when making predictions in unseen data[335].

The quality control (QC) pipeline used for the preparation of the final multiple omics datasets used in the analyses presented in chapters 3 and 4 is a limitation. This is due to the order in which the steps were carried out, meaning that models constructed in these chapters were susceptible to potential indirect data leakage[336].

Specifically, raw omics measures were pre-corrected for covariates (using linear regression), scaled and centred prior to splitting the sample into training and testing datasets. This means that information from the testing sample was implicitly involved in creating the final qc'd training dataset that was used for model construction therefore informing the final model. This also means that the testing set is not completely independent.

Future work should swap the order of QC steps so that pre-correction for covariates, scaling and centring of the omics measures occurs post training-testing split. Ideally these steps would occur within each fold of the 10-fold cross validation in the training sample and in the testing sample separately.

6.4 Conclusion

In this thesis I took advantage of the range of omics assays available in ORCADES and investigated their relationships with ageing, health-related risk factors and subsequent incident disease.

First, I showed that omics biomarkers can be used to build measures of biological age that contain more predictive information of risk factors and incident disease than chronological age alone. This exhaustive comparison of multiple omics ageing clocks produced several novel findings that further our understanding of the properties of these models. Namely, that these ageing clocks built from multiple different omics overlap more than would be expected by chance, that clocks built using a substantial subset or a few principal components of omics biomarkers produce models that are just as effective and quantified the proportion of OCAA that is capturing noise rather than true underlying biological age by clocks trained on chronological age.

Second, I demonstrated that these omics biomarkers, both individually and combined in multivariable models, are associated with health-related risk factors and are prognostic of subsequent incident diseases. I found that the majority of single omics biomarkers were associated with multiple health outcomes rather than being outcome specific, emphasising how interconnected these omics layers are. While determining the relative importance of biomarkers included in omics scores, I found that only a handful of biomarkers, only one in extreme cases, were driving the effectiveness of the score. These omics biomarkers are therefore extremely promising candidates for risk scores that could be used to prevent disease.

Third, I used methods that leverage genetic data to investigate how genetic variation affects the levels of cardiovascular-related proteins circulating in the plasma and how these protein levels affect disease risk. By performing the largest genome-wide association meta-analysis on the levels of 184 proteins, I: discovered 592 associated regions of the genome, unravelled potential mechanisms and pathways by which these regions may act to influence the levels of the proteins we find circulating. I further inferred causal relationships between protein levels and diseases and by doing so identified novel therapeutic targets and an opportunity for drug repurposing. This analysis also created a resource of association summary statistics and protein instruments for causal inference that will continue to benefit the field.

I conclude that with statistical techniques such as machine learning and large sample sizes, omics assays have the potential to deliver answers to questions regarding the mechanisms and pathways that underly ageing and disease that genomics alone has failed to answer. I would also like to stress how important it is to integrate multiple omics if we want to have a chance of filling the gaps in our understanding of why we are the way we are.

Bibliography

1. Polack, F. P. *et al.* Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N. Engl. J. Med.* **383**, 2603–2615 (2020).
2. Voysey, M. *et al.* Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet* **397**, 99–111 (2021).
3. Baden, L. R. *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N. Engl. J. Med.* **384**, 403–416 (2021).
4. Coronavirus (COVID-19) Vaccinations - Statistics and Research. *Our World in Data* https://ourworldindata.org/covid-vaccinations.
5. *A Dictionary of Public Health*. *A Dictionary of Public Health* (Oxford University Press, 2007).
6. Eijk, P. J. van der. *Hippocrates in Context: Papers Read at the XIth International Hippocrates Colloquium (University of Newcastle upon Tyne, 27-31 August 2002)*. (BRILL, 2018).
7. Halliday, S. Death and miasma in Victorian London: an obstinate belief. *BMJ* **323**, 1469–1471 (2001).
8. Sterner, C. S. A Brief History of Miasmic Theory. *Bull. Hist. Med.* **22**, 747 (1948).
9. D'Agramont, J., Winslow, C.-E. A. & Duran-Reynals, M. L. Regiment de preservacio a epidimia o pestilencia e mortaldats: epistola de Maestre Jacme d'Agramont als honrats e discrets seynnors pahers e conseyll de la Ciutat de leyda 1348 = regimen of protection against epidemics or pestilence and mortality. *Bull. Hist. Med.* **23**, (1949).
10. *Methods in Enzymology, Volume 13: Citric Acid Cycle*. (Academic Press, 1969).
11. Vos, T. *et al.* Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* **396**, 1204–1222 (2020).
12. Abbott, S. & Fairbanks, D. J. Experiments on Plant Hybrids by Gregor Mendel. *Genetics* **204**, 407–422 (2016).
13. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
14. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74**, 5463–5467 (1977).
15. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
16. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).
17. Wang, D. G. *et al.* Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* **280**, 1077–1082 (1998).
18. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

19. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
20. 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).
21. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
22. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, (2020).
23. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
24. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease. *Science* **322**, 881–888 (2008).
25. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
26. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: Illuminating the Dark Road from Association to Function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).
27. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
28. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
29. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci.* **103**, 1412–1417 (2006).
30. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
31. Mohn, F. *et al.* Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. *Mol. Cell* **30**, 755–766 (2008).
32. Breitling, L. P., Yang, R., Korn, B., Burwinkel, B. & Brenner, H. Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. *Am. J. Hum. Genet.* **88**, 450–457 (2011).
33. Dhingra, R. *et al.* Evaluating DNA methylation age on the Illumina MethylationEPIC Bead Chip. *PLOS ONE* **14**, e0207834–e0207834 (2019).
34. Lundberg, M., Eriksson, A., Tran, B., Assarsson, E. & Fredriksson, S. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Res.* **39**, e102–e102 (2011).
35. Assarsson, E. *et al.* Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. *PLOS ONE* **9**, e95192 (2014).
36. Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLOS ONE* **5**, e15004 (2010).

37. Oyama, K. *et al.* Effect of Evolocumab on Complex Coronary Disease Requiring Revascularization. *J. Am. Coll. Cardiol.* **77**, 259–267 (2021).
38. Devito, F. *et al.* Focus on alirocumab: A PCSK9 antibody to treat hypercholesterolemia. *Pharmacol. Res.* **102**, 168–175 (2015).
39. Turturro, F. Denileukin diftitox: a biotherapeutic paradigm shift in the treatment of lymphoid-derived disorders. *Expert Rev. Anticancer Ther.* **7**, 11–17 (2007).
40. Waldmann, T. A. Anti-Tac (daclizumab, Zenapax) in the Treatment of Leukemia, Autoimmune Diseases, and in the Prevention of Allograft Rejection: A 25-Year Personal Odyssey. *J. Clin. Immunol.* **27**, 1–18 (2007).
41. Varki, A. Biological roles of glycans. *Glycobiology* **27**, 3–49 (2017).
42. Lauc, G. *et al.* Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers. *PLOS Genet.* **9**, e1003225 (2013).
43. Gudelj, I., Lauc, G. & Pezer, M. Immunoglobulin G glycosylation in aging and diseases. *Cell. Immunol.* **333**, 65–79 (2018).
44. Nicholson, J. K. *et al.* Metabolic phenotyping in clinical and surgical environments. *Nature* **491**, 384–392 (2012).
45. Population estimates for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2019estimates.
46. English Life Tables No.17 - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/englishlifetablesno17/2015-09-01.
47. National life tables – life expectancy in the UK - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/2017to2019.
48. Guzman-Castillo, M. *et al.* Forecasted trends in disability and life expectancy in England and Wales up to 2025: a modelling study. *Lancet Public Health* **2**, e307–e313 (2017).
49. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194–1217 (2013).
50. Lindsey, J., McGill, N. I., Lindsey, L. A., Green, D. K. & Cooke, H. J. In vivo loss of telomeric repeats with age in humans. *Mutat. Res.* **256**, 45–48 (1991).
51. Besingi, W. & Johansson, Å. Smoke-related DNA methylation changes in the etiology of human disease. *Hum. Mol. Genet.* **23**, 2290–2297 (2014).
52. Epel, E. S. *et al.* Accelerated telomere shortening in response to life stress. *Proc. Natl. Acad. Sci.* **101**, 17312–17315 (2004).
53. Blair, S. N. *et al.* Physical Fitness and All-Cause Mortality: A Prospective Study of Healthy Men and Women. *JAMA* **262**, 2395–2401 (1989).
54. Enroth, S., Enroth, S. B., Johansson, Å. & Gyllensten, U. Protein profiling reveals consequences of lifestyle choices on predicted biological aging. *Sci. Rep.* **5**, 1–10 (2015).

55. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cell* **49**, 359–367 (2013).
56. Baker, G. T. & Sprott, R. L. Biomarkers of aging. *Exp. Gerontol.* **23**, 223–239 (1988).
57. Fraga, M. F. & Esteller, M. Epigenetics and aging: the targets and the marks. *Trends Genet.* **23**, 413–418 (2007).
58. Christensen, B. C. *et al.* Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *PLOS Genet.* **5**, e1000602 (2009).
59. Bollati, V. *et al.* Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mech. Ageing Dev.* **130**, 234–239 (2009).
60. Teschendorff, A. E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**, 440–446 (2010).
61. Mugatroyd, C., Wu, Y., Bockmühl, Y. & Spengler, D. The Janus face of DNA methylation in aging. *Aging* **2**, 107–110 (2010).
62. Rodríguez-Rodero, S., Fernández-Morera, J. L., Fernandez, A. F., Menendez-Torre, E. & Fraga, M. F. Epigenetic Regulation of Aging. *Discov. Med.* **10**, 225–233 (2010).
63. Bell, J. T. *et al.* Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population. *PLOS Genet.* **8**, e1002629 (2012).
64. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **13**, R97 (2012).
65. Bocklandt, S. *et al.* Epigenetic Predictor of Age. *PLOS ONE* **6**, e14821 (2011).
66. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115–R115 (2013).
67. Marioni, R. E. *et al.* The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *Int. J. Epidemiol.* **44**, 1388–1396 (2015).
68. Horvath, S. *et al.* Accelerated epigenetic aging in Down syndrome. *Aging Cell* **14**, 491–495 (2015).
69. Horvath, S. & Ritz, B. R. Increased epigenetic age and granulocyte counts in the blood of Parkinson’s disease patients. *Aging* vol. 7 1130–1142 https://www.aging-us.com/article/100859/text (2015).
70. Maierhofer, A. *et al.* Accelerated epigenetic aging in Werner syndrome. *Aging* vol. 9 1143–1152 https://www.aging-us.com/article/101217/text (2017).
71. Horvath, S. *et al.* Huntington’s disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels. *Aging* **8**, 1485–1504 (2016).
72. Levine, M. E., Lu, A. T., Bennett, D. A. & Horvath, S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer’s disease related cognitive functioning. *Aging* vol. 7 1198–1211 https://www.aging-us.com/article/100864/text (2015).
73. Breitling, L. P. *et al.* Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. *Clin. Epigenetics* **8**, 21 (2016).

74. Walker, R. F. *et al.* Epigenetic age analysis of children who seem to evade aging. *Aging* vol. 7 334–339 https://www.aging-us.com/article/100744/text (2015).
75. Horvath, S. & Levine, A. J. HIV-1 Infection Accelerates Age According to the Epigenetic Clock. *J. Infect. Dis.* **212**, 1563–1573 (2015).
76. Vidal, L. *et al.* Specific increase of methylation age in osteoarthritis cartilage. *Osteoarthritis Cartilage* **24**, S63 (2016).
77. Levine, M. E. *et al.* Menopause accelerates biological aging. *Proc. Natl. Acad. Sci.* **113**, 9327–9332 (2016).
78. Horvath, S. *et al.* Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring. *Aging* **7**, 1159–1170 (2015).
79. Chen, B. H. *et al.* DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging* **8**, 1844–1865 (2016).
80. Zhang, W. G. *et al.* Select aging biomarkers based on telomere length and chronological age to build a biological age equation. *Age* **36**, 1201–1211 (2014).
81. Chen, W. *et al.* Three-dimensional human facial morphologies as robust aging markers. *Cell Res.* **25**, 574–587 (2015).
82. Cole, J. H. *et al.* Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage* **163**, 115–124 (2017).
83. Cole, J. H. *et al.* Brain age predicts mortality. *Mol. Psychiatry* **23**, 1385–1392 (2018).
84. Goyal, M. S. *et al.* Persistent metabolic youth in the aging female brain. *Proc. Natl. Acad. Sci.* **116**, 3251 (2019).
85. Sokolova, K., Barker, G. J. & Montana, G. Convolutional neural-network-based ordinal regression for brain age prediction from MRI scans. in *Medical Imaging 2020: Image Processing* vol. 11313 113132B (International Society for Optics and Photonics, 2020).
86. van den Akker, E. B. *et al.* Metabolic Age Based on the BBMRI-NL 1H-NMR Metabolomics Repository as Biomarker of Age-related Disease. *Circ. Genomic Precis. Med.* **13**, 541–547 (2020).
87. Krištić, J. *et al.* Glycans Are a Novel Biomarker of Chronological and Biological Ages. *J. Gerontol. Ser. A* **69**, 779–789 (2014).
88. Lehallier, B. *et al.* Undulating changes in human plasma proteome across lifespan are linked to disease. *bioRxiv* 751115–751115 (2019) doi:10.1101/751115.
89. Lehallier, B., Shokhirev, M. N., Wyss-Coray, T. & Johnson, A. A. Data mining of human plasma proteins generates a multitude of highly predictive aging clocks that reflect different aspects of aging. *Aging Cell* **n/a**, e13256.
90. Alpert, A. *et al.* A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nat. Med.* **25**, (2019).
91. Lehallier, B., Shokhirev, M. N., Wyss-Coray, T. & Johnson, A. A. Data mining of human plasma proteins generates a multitude of highly predictive aging clocks that reflect different aspects of aging. *Aging Cell* **19**, e13256 (2020).
92. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* **11**, 54–54 (2019).

93. Belsky, D. W. *et al.* Eleven Telomere, Epigenetic Clock, and Biomarker-Composite Quantifications of Biological Aging: Do They Measure the Same Thing? *Am. J. Epidemiol.* **187**, 1220–1230 (2017).

94. Maddock, J. *et al.* DNA methylation age and physical and cognitive ageing. *J. Gerontol. Ser. A* (2019) doi:10.1093/gerona/glz246.

95. Li, X. *et al.* Longitudinal trajectories, correlations and mortality associations of nine biological ages across 20-years follow-up. *eLife* **9**, e51507 (2020).

96. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).

97. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, (2019).

98. Jansen, R. *et al.* An integrative study of five biological clocks in somatic and mental health. *eLife* **10**, e59479 (2021).

99. Liu, Z., Leung, D. & Levine, M. Comparative analysis of epigenetic aging clocks from CpG characteristics to functional associations. *bioRxiv* 512483 (2019) doi:10.1101/512483.

100. Fahy, G. M. *et al.* Reversal of epigenetic aging and immunosenescent trends in humans. *Aging Cell* **18**, (2019).

101. D'Agostino Ralph B. *et al.* General Cardiovascular Risk Profile for Use in Primary Care. *Circulation* **117**, 743–753 (2008).

102. Ferrario, M. *et al.* Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation. *Int. J. Epidemiol.* **34**, 413–421 (2005).

103. Zhang, X.-F., Attia, J., D'Este, C., Yu, X.-H. & Wu, X.-G. A risk score predicted coronary heart disease and stroke in a Chinese cohort. *J. Clin. Epidemiol.* **58**, 951–958 (2005).

104. Dorresteijn, J. A. N. *et al.* Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. *Heart* **99**, 866–872 (2013).

105. Biancari, F. *et al.* Prediction of severe bleeding after coronary surgery: the WILL-BLEED Risk Score. *Thromb. Haemost.* **117**, 445–456 (2017).

106. Gratwohl, A. The EBMT risk score. *Bone Marrow Transplant.* **47**, 749–756 (2012).

107. Idzerda, N. M. A., Tye, S. C., Zeeuw, D. de & Heerspink, H. J. L. A novel drug response score more accurately predicts renoprotective drug effects than existing renal risk scores. *Ther. Adv. Endocrinol. Metab.* **12**, 2042018820974191 (2021).

108. Menni Cristina *et al.* Glycosylation Profile of Immunoglobulin G Is Cross-Sectionally Associated With Cardiovascular Disease Risk Score and Subclinical Atherosclerosis in Two Independent Cohorts. *Circ. Res.* **122**, 1555–1564 (2018).

109. Enroth, S. *et al.* High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. *Commun. Biol.* **2**, 221–221 (2019).

110. Gisby, J. *et al.* Longitudinal proteomic profiling of dialysis patients with COVID-19 reveals markers of severity and predictors of death. *eLife* **10**, e64827 (2021).

111. Pietzner, M. *et al.* Plasma metabolites to profile pathways in noncommunicable disease multimorbidity. *Nat. Med.* **27**, 471–479 (2021).

112. Li, H., Habes, M., Wolk, D. A. & Fan, Y. A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimers Dement.* **15**, 1059–1070 (2019).

113. Ambler, G., Seaman, S. & Omar, R. Z. An evaluation of penalised survival methods for developing prognostic models with rare events. *Stat. Med.* **31**, 1150–1161 (2012).

114. Berggrund, M. *et al.* Identification of Candidate Plasma Protein Biomarkers for Cervical Cancer Using the Multiplex Proximity Extension Assay *[S]. *Mol. Cell. Proteomics* **18**, 735–743 (2019).

115. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLOS Genet.* **13**, e1006706 (2017).

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

117. Enroth, S., Johansson, Å., Enroth, S. B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nat. Commun.* **5**, 4684 (2014).

118. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).

119. Melzer, D. *et al.* A Genome-Wide Association Study Identifies Protein Quantitative Trait Loci (pQTLs). *PLOS Genet.* **4**, e1000072 (2008).

120. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018).

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

122. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7**, 11122–11122 (2016).

123. Tabassum, R. *et al.* Genetic architecture of human plasma lipidome and its link to cardiovascular disease. *Nat. Commun.* **10**, 4329 (2019).

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

125. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).

126. Abreu, R. de S., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* **5**, 1512–1526 (2009).

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

128. Waldmann, T. A. Daclizumab (anti-Tac, Zenapax) in the treatment of leukemia/lymphoma. *Oncogene* **26**, 3699–3703 (2007).

129. Choy, B. Y. *et al.* IL2-receptor antagonist (basiliximab) induction therapy is associated with lower morbidity and mortality in renal transplant recipients. *Transplant. Proc.* **35**, 195 (2003).

130. Siegal, D. M. *et al.* Andexanet Alfa for the Reversal of Factor Xa Inhibitor Activity. *N. Engl. J. Med.* **373**, 2413–2424 (2015).

131. Pulst, S. M. Genetic linkage analysis. *Arch. Neurol.* **56**, 667–672 (1999).

132. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nat. Methods* **9**, 459–462 (2012).

133. Consortium, the H. R. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

134. LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* **37**, 4181–4193 (2009).

135. Thomas, S. C. & Hill, W. G. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* **155**, 1961–1972 (2000).

136. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits*. (OUP USA, 1998).

137. Belonogova, N. M., Svishcheva, G. R., Duijn, C. M. van, Aulchenko, Y. S. & Axenovich, T. I. Region-Based Association Analysis of Human Quantitative Traits in Related Individuals. *PLOS ONE* **8**, e65395 (2013).

138. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components–based method for whole-genome association analysis. *Nat. Genet.* **44**, 1166–1170 (2012).

139. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

140. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749–1755 (2019).

141. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

142. Lewontin, R. C. & Kojima, K. The Evolutionary Dynamics of Complex Polymorphisms. *Evolution* **14**, 458–472 (1960).

143. Aulchenko, Y. S., Struchalin, M. V. & van Duijn, C. M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11**, 134 (2010).

144. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

145. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

146. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Optimal designs for two-stage genome-wide association studies. *Genet. Epidemiol.* **31**, 776–788 (2007).

147. de Bakker, P. I. W. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).

148. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

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

150. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).

151. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).

152. Sanna, S. *et al.* Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLOS Genet.* **7**, e1002198 (2011).

153. Gilly, A. *et al.* Whole-genome sequencing analysis of the cardiometabolic proteome. *Nat. Commun.* **11**, 6336 (2020).

154. Cutting, G. R. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat. Rev. Genet.* **16**, 45–56 (2015).

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

156. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).

157. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

158. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

159. Falconer, D. S. *Introduction to Quantitative Genetics*. (Longman, 1995).

160. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

161. Ning, Z., Pawitan, Y. & Shen, X. High-definition likelihood inference of genetic correlations across human complex traits. *Nat. Genet.* **52**, 859–864 (2020).

162. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).

163. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *Int. J. Epidemiol.* **32**, 1–22 (2003).

164. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).

165. Smith, G. D. & Ebrahim, S. Epidemiology—is it time to call it a day? *Int. J. Epidemiol.* **30**, 1–11 (2001).

166. Smith, G. D. & Ebrahim, S. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ* **330**, 1076–1079 (2005).

167. Nitsch, D. *et al.* Limits to Causal Inference based on Mendelian Randomization: A Comparison with Randomized Controlled Trials. *Am. J. Epidemiol.* **163**, 397–403 (2006).

168. Greenland, S. An introduction to instrumental variables for epidemiologists. *Int. J. Epidemiol.* **29**, 722–729 (2000).

169. Pierce, B. L. & Burgess, S. Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).

170. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).

171. Wald, A. The Fitting of Straight Lines if Both Variables are Subject to Error. *Ann. Math. Stat.* **11**, 284–300 (1940).

172. Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V. & Klungel, O. H. Instrumental Variables: Application and Limitations. *Epidemiology* **17**, 260–267 (2006).

173. Bowden, J., Smith, G. D., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).

174. Egger, M., Smith, G. D., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**, 629–634 (1997).

175. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).

176. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* **32**, 377–389 (2017).

177. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* **37**, 658–665 (2013).

178. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

179. Hemani, G., Tilling, K. & Smith, G. D. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLOS Genet.* **13**, e1007081 (2017).

180. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018) doi:10.1101/447367.

181. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).

182. Lloyd-Jones, L. R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am. J. Hum. Genet.* **100**, 228–237 (2017).

183. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

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

185. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* (2018) doi:10.1038/s41586-018-0175-2.

186. McQuillan, R. *et al.* Runs of Homozygosity in European Populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).

187. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).

188. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLOS Genet.* **10**, e1004234 (2014).

189. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).

190. Smith, B. H. *et al.* Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int. J. Epidemiol.* **42**, 689–700 (2013).

191. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).

192. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).

193. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).

194. Breiman, L. Heuristics of Instability and Stabilization in Model Selection. *Ann. Stat.* **24**, 2350–2383 (1996).

195. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55–67 (1970).

196. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).

197. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).

198. Haller, T., Kals, M., Esko, T., Mägi, R. & Fischer, K. RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Brief. Bioinform.* **16**, 39–44 (2015).

199. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558 (2002).

200. Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327**, 557–560 (2003).

201. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

202. Horvath, S. *et al.* Reversing age: dual species measurement of epigenetic age with a single clock. doi:10.1101/2020.05.07.082917.

203. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16**, 25 (2015).

204. Hillary, R. F. *et al.* Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden. *Clin. Epigenetics* **12**, 115 (2020).

205. Zheng, Y. *et al.* Blood Epigenetic Age may Predict Cancer Incidence and Mortality. *EBioMedicine* **5**, 68–73 (2016).

206. Horvath, S. *et al.* An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol.* **17**, 171 (2016).

207. Franke, K., Ziegler, G., Klöppel, S. & Gaser, C. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage* **50**, 883–892 (2010).

208. Pyrkov, T. V. & Fedichev, P. O. Biological age is a universal marker of aging, stress, and frailty. doi:10.1101/578245.

209. Alsaleh, H. & Haddrill, P. R. Identifying blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC® BeadChip. *Forensic Sci. Int.* **303**, 109944 (2019).

210. Hofman, A. *et al.* DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging* **8**, 1844–1865 (2016).

211. Christiansen, L. *et al.* DNA methylation age is associated with mortality in a longitudinal Danish twin study. *Aging Cell* **15**, 149–154 (2016).

212. Perna, L. *et al.* Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin. Epigenetics* **8**, 64 (2016).

213. Singh, S. S. *et al.* Association of the IgG N-glycome with the course of kidney function in type 2 diabetes. *BMJ Open Diabetes Res. Care* **8**, e001026 (2020).

214. Lemmers, R. F. H. *et al.* IgG glycan patterns are associated with type 2 diabetes in independent European populations. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1861**, 2240–2249 (2017).

215. Liu, Z., Leung, D. & Levine, M. Comparative analysis of epigenetic aging clocks from CpG characteristics to functional associations. 29.

216. Lee, Y. *et al.* Blood-based epigenetic estimators of chronological age in human adults using DNA methylation data from the Illumina MethylationEPIC array. *BMC Genomics* **21**, 747 (2020).

217. McCrory, C. *et al.* GrimAge outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality. *J. Gerontol. Ser. A* doi:10.1093/gerona/glaa286.

218. Pucić, M. *et al.* High throughput isolation and glycosylation analysis of IgG-variability and heritability of the IgG glycome in three isolated human populations. *Mol. Cell. Proteomics MCP* **10**, M111.010090-M111.010090 (2011).

219. Klarić, L. *et al.* Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. *Sci. Adv.* **6**, eaax0301 (2020).

220. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983–3989 (2018).

221. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).

222. Sales, S. *et al.* Gender, Contraceptives and Individual Metabolic Predisposition Shape a Healthy Plasma Lipidome. *Sci. Rep.* **6**, 27710 (2016).

223. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, 161–161 (2014).

224. Lundberg, M., Eriksson, A., Tran, B., Assarsson, E. & Fredriksson, S. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. doi:10.1093/nar/gkr424.

225. Long, T. *et al.* Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.* **49**, 568–578 (2017).

226. Kim, S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Commun. Stat. Appl. Methods* **22**, 665–674 (2015).

227. Cox, D. R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B Methodol.* **34**, 187–220 (1972).

228. Jiang, M. *et al.* Frailty index as a predictor of all-cause and cause-specific mortality in a Swedish population-based cohort. *Aging* **9**, 2629–2646 (2017).

229. Lu, A. T. *et al.* DNA methylation-based estimator of telomere length. *Aging* (2019) doi:10.18632/aging.102173.

230. Goff, D. C. *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* **129**, S49–S73 (2014).

231. Therneau, T. M., Grambsch, P. M. & Fleming, T. R. Martingale-Based Residuals for Survival Models. *Biometrika* **77**, 147–160 (1990).

232. Joshi, P. K. *et al.* Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nat. Commun.* **8**, 910 (2017).

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

234. Nally, R. M. & Walsh, C. J. Hierarchical Partitioning Public-domain Software. *Biodivers. Conserv.* **13**, 659–660 (2004).

235. Chevan, A. & Sutherland, M. Hierarchical Partitioning. *Am. Stat.* **45**, 90–96 (1991).

236. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* **40**, 161–169 (2008).

237. UK Biobank. *Neale lab* http://www.nealelab.is/uk-biobank.

238. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).

239. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).

240. Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLOS Med.* **17**, e1003062 (2020).

241. Matsunaga, H. *et al.* Transethnic Meta-Analysis of Genome-Wide Association Studies Identifies Three New Loci and Characterizes Population-Specific Differences for Coronary Artery Disease. *Circ. Genomic Precis. Med.* **13**, e002670 (2020).

242. Klimentidis, Y. C. *et al.* Phenotypic and Genetic Characterization of Lower LDL Cholesterol and Increased Type 2 Diabetes Risk in the UK Biobank. *Diabetes* **69**, 2194–2205 (2020).

243. Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **51**, 51–62 (2019).

244. Langefeld, C. D. *et al.* Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).

245. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).

246. van der Harst Pim & Verweij Niek. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).

247. Gerstein, H. C. Glucose: a continuous risk factor for cardiovascular disease. *Diabet. Med. J. Br. Diabet. Assoc.* **14 Suppl 3**, S25-31 (1997).

248. Tran, D. H. & Wang, Z. V. Glucose Metabolism in Cardiac Hypertrophy and Heart Failure. *J. Am. Heart Assoc.* **8**, e012673 (2019).

249. Zöchbauer-Müller, S. *et al.* Aberrant methylation of multiple genes in the upper aerodigestive tract epithelium of heavy smokers. *Int. J. Cancer* **107**, 612–616 (2003).

250. Marsit, C. J., Houseman, E. A., Schned, A. R., Karagas, M. R. & Kelsey, K. T. Promoter hypermethylation is associated with current smoking, age, gender and survival in bladder cancer. *Carcinogenesis* **28**, 1745–1751 (2007).

251. Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging* **9**, 2504–2520 (2017).

252. Timmers, P. R. *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife* **8**, e39856 (2019).

253. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

254. Pietzner, M. *et al.* Cross-platform proteomics to advance genetic prioritisation strategies. *bioRxiv* 2021.03.18.435919 (2021) doi:10.1101/2021.03.18.435919.

255. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).

256. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).

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

258. Williams, S. A. *et al.* Plasma protein patterns as comprehensive indicators of health. *Nat. Med.* **25**, 1851–1857 (2019).

259. Lehallier, B. *et al.* Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).

260. Smith, J. G. & Gerszten, R. E. Emerging Affinity-Based Proteomic Technologies for Large-Scale Plasma Profiling in Cardiovascular Disease. *Circulation* **135**, 1651–1664 (2017).

261. Gashaw, I., Ellinghaus, P., Sommer, A. & Asadullah, K. What makes a good drug target? *Drug Discov. Today* **17**, S24–S30 (2012).

262. Liu, Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11**, 786 (2015).

263. Holmes, M. V., Ala-Korpela, M. & Smith, G. D. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat. Rev. Cardiol.* **14**, 577–590 (2017).

264. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

265. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).

266. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

267. Tenenbaum, D. *KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). R package version 1.30.1*. (2020).

268. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).

269. Katsiki, N., Mikhailidis, D. P. & Banach, M. Leptin, cardiovascular diseases and type 2 diabetes mellitus. *Acta Pharmacol. Sin.* **39**, 1176–1188 (2018).

270. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).

271. Timpson, N. J. *et al.* C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *Int. J. Obes.* **35**, 300–308 (2011).

272. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).

273. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* **32**, 377–389 (2017).

274. Seidah Nabil G., Awan Zuhier, Chrétien Michel, & Mbikay Majambu. Pcsk9. *Circ. Res.* **114**, 1022–1036 (2014).

275. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).

276. Andrews, C., McLean, M. H. & Durum, S. K. Interleukin-27 as a Novel Therapy for Inflammatory Bowel Disease: A Critical Review of the Literature. *Inflamm. Bowel Dis.* **22**, 2255–2264 (2016).

277. Porter, R. J., Andrews, C., Brice, D. P., Durum, S. K. & McLean, M. H. Can We Target Endogenous Anti-inflammatory Responses as a Therapeutic Strategy for Inflammatory Bowel Disease? *Inflamm. Bowel Dis.* **24**, 2123–2134 (2018).

278. Visperas, A., Do, J. S., Bulek, K., Li, X. & Min, B. IL-27, targeting antigen-presenting cells, promotes Th17 differentiation and colitis in mice. *Mucosal Immunol.* **7**, 625–633 (2014).

279. Furuzawa Carballeda, J., Fonseca Camarillo, G. & Yamamoto-Furusho, J. K. Interleukin 27 is up-regulated in patients with active inflammatory bowel disease. *Immunol. Res.* **64**, 901–907 (2016).

280. Imielinski, M. *et al.* Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat. Genet.* **41**, 1335–1340 (2009).

281. Page, M. M. & Watts, G. F. Evolocumab in the treatment of dyslipidemia: pre-clinical and clinical pharmacology. *Expert Opin. Drug Metab. Toxicol.* **11**, 1505–1515 (2015).

282. German, C. A. & Shapiro, M. D. Small Interfering RNA Therapeutic Inclisiran: A New Approach to Targeting PCSK9. *BioDrugs* **34**, 1–9 (2020).

283. Creed, T. J. *et al.* Basiliximab for the treatment of steroid-resistant ulcerative colitis: further experience in moderate and severe disease. *Aliment. Pharmacol. Ther.* **23**, 1435–1442 (2006).

284. Sands, B. E. *et al.* Basiliximab Does Not Increase Efficacy of Corticosteroids in Patients With Steroid-Refractory Ulcerative Colitis. *Gastroenterology* **143**, 356-364.e1 (2012).

285. Yablecovitch, D. *et al.* Serum MMP-9: a novel biomarker for prediction of clinical relapse in patients with quiescent Crohn's disease, a post hoc analysis. *Ther. Adv. Gastroenterol.* **12**, 1756284819881590 (2019).

286. Schreiber, S. *et al.* A Phase 2, Randomized, Placebo-Controlled Study Evaluating Matrix Metalloproteinase-9 Inhibitor, Andecaliximab, in Patients With Moderately to Severely Active Crohn's Disease. *J. Crohns Colitis* **12**, 1014–1020 (2018).

287. Malmeström, C. *et al.* Serum levels of LIGHT in MS. *Mult. Scler. J.* **19**, 871–876 (2013).

288. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, (2015).

289. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *bioRxiv* 2021.02.23.432421 (2021) doi:10.1101/2021.02.23.432421.

290. Huan, T. *et al.* Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat. Commun.* **10**, 4267 (2019).

291. Klaric, L. *et al.* Mendelian randomisation identifies alternative splicing of the FAS death receptor as a mediator of severe COVID-19. *medRxiv* 2021.04.01.21254789 (2021) doi:10.1101/2021.04.01.21254789.

292. Zhao, Q., Wang, J., Hemani, G., Bowden, J. & Small, D. S. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann. Stat.* **48**, (2020).

293. Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M. & He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.* **52**, 740–747 (2020).

294. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

295. Staley, J. R. *et al.* PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).

296. Team, B. *Homo.sapiens: Annotation package for the Homo.sapiens object. R package version 1.3.1.* (2015).

297. Carlson, M. *org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2*. (2019).

298. Little, R. J. Missing data in Census Bureau surveys. *Proc. Second Annu. Census Bur. Res. Conf.* 442–454 (1986).

299. Monard, M.-C. A Study of K-Nearest Neighbour as an Imputation Method. *Second Int. Conf. Hybrid Intell. Syst.* **87**, 251–260 (2002).

300. Randall, J. C. *et al.* Sex-stratified Genome-wide Association Studies Including 270,000 Individuals Show Sexual Dimorphism in Genetic Loci for Anthropometric Traits. *PLOS Genet.* **9**, e1003500 (2013).

301. Myers, R. A. *et al.* Genome-wide interaction studies reveal sex-specific asthma risk alleles. *Hum. Mol. Genet.* **23**, 5251–5259 (2014).

302. Mitra, I. *et al.* Pleiotropic Mechanisms Indicated for Sex Differences in Autism. *PLOS Genet.* **12**, e1006425 (2016).

303. Frieden, T. R. & Centers for Disease Control and Prevention (CDC). CDC Health Disparities and Inequalities Report - United States, 2013. Foreword. *MMWR Suppl.* **62**, 1–2 (2013).

304. Anderson, G. D. Sex and Racial Differences in Pharmacological Response: Where Is the Evidence? Pharmacogenetics, Pharmacokinetics, and Pharmacodynamics. *J. Womens Health* **14**, 19–29 (2005).

305. Huffman, J. E. Examining the current standards for genetic discovery and replication in the era of mega-biobanks. *Nat. Commun.* **9**, 5054 (2018).

306. Zanetti, D. & Weale, M. E. Transethnic differences in GWAS signals: A simulation study. *Ann. Hum. Genet.* **82**, 280–286 (2018).

307. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

308. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging Genetic Variability across Populations for the Identification of Causal Variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).

309. Carlson, C. S. *et al.* Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLOS Biol.* **11**, e1001661 (2013).

310. Hellwege, J. N. *et al.* Population Stratification in Genetic Association Studies. *Curr. Protoc. Hum. Genet.* **95**, 1.22.1-1.22.23 (2017).

311. Wang, X. *et al.* Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **22**, 2303–2311 (2013).

312. Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).

313. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).

314. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).

315. Lander, E. S. The New Genomics: Global Views of Biology. *Science* **274**, 536–539 (1996).

316. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).

317. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease–common variant… or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).

318. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33**, 228–237 (2003).

319. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017).

320. Manolio, T. A. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* **14**, 549–558 (2013).

321. Wainschtein, P. *et al.* Recovery of trait heritability from whole genome sequence data. *bioRxiv* 588020 (2021) doi:10.1101/588020.

322. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nat. Rev. Genet.* **10**, 783–796 (2009).

323. Carson, A. R. *et al.* Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Nephrol.* **15**, 125 (2014).

324. Adelson, R. P. *et al.* Empirical design of a variant quality control pipeline for whole genome sequencing data using replicate discordance. *Sci. Rep.* **9**, 16156 (2019).

325. Ma, C., Blackwell, T., Boehnke, M. & Scott, L. J. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* **37**, 539–550 (2013).

326. Pulit, S. L., de With, S. A. J. & de Bakker, P. I. W. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. *Genet. Epidemiol.* **41**, 145–151 (2017).

327. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics* vol. 95 5–23 (2014).

328. General Data Protection Regulation (GDPR) – Official Legal Text. *General Data Protection Regulation (GDPR)* https://gdpr-info.eu/.

329. Timmers, P. R. H. J. *et al.* Trends in disease incidence and survival and their effect on mortality in Scotland: nationwide cohort study of linked hospital admission and death records 2001–2016. *BMJ Open* **10**, e034299 (2020).

330. DiMeglio, L. A., Evans-Molina, C. & Oram, R. A. Type 1 diabetes. *The Lancet* **391**, 2449–2462 (2018).

331. Chatterjee, S., Khunti, K. & Davies, M. J. Type 2 diabetes. *The Lancet* **389**, 2239–2251 (2017).

332. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).

333. Altman, D. G. & Bland, J. M. Statistics Notes: Diagnostic tests 2: predictive values. *BMJ* **309**, 102 (1994).

334. Brynedal, B. *et al.* Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *Am. J. Hum. Genet.* **100**, 581–591 (2017).

335. Kaufman, S., Rosset, S. & Perlich, C. Leakage in data mining: formulation, detection, and avoidance. in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* 556–563 (Association for Computing Machinery, 2011). doi:10.1145/2020408.2020496.

336. Zheng, A. & Casari, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. (O'Reilly Media, 2018).

337. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, (2017).

Appendix

Chapter 2 Supplementary Information

	Clinomics			DEXA			Hannum CpGs DNAme		
Outcome	Total	Cases	Controls	Total	Cases	Controls	Total	Cases	Controls
all	1388	463	925	881	308	573	812	314	498
c	1636	98	1538	1040	67	973	938	69	869
c00.c14	1662	3	1659	1058	1	1057	961	1	960
c15.c26	1657	22	1635	1053	17	1036	957	16	941
c30.c39	1661	5	1656	1058	4	1054	960	3	957
c43.c44	1657	11	1646	1055	7	1048	958	8	950
c45.c49	1661	2	1659	1057	2	1055	960	3	957
c50.c50	1658	18	1640	1055	14	1041	956	12	944
c51.c58	1661	8	1653	1058	3	1055	959	7	952
c60.c63	1655	12	1643	1054	8	1046	955	7	948
c64.c68	1661	10	1651	1057	9	1048	961	10	951
c69.c72	1661	2	1659	1057	2	1055	960	2	958
c73.c75	1662	2	1660	1058	2	1056	961	1	960
c76.c80	1659	34	1625	1057	24	1033	959	25	934
c81.c96	1662	16	1646	1058	12	1046	961	8	953
e	1602	206	1396	1015	131	884	936	142	794
e00.e07	1631	106	1525	1038	72	966	952	72	880
e10.e14	1650	49	1601	1049	29	1020	954	39	915
e15.e16	1661	2	1659	1058	2	1056	960	2	958
e20.e35	1661	4	1657	1057	2	1055	961	4	957
e50.e64	1662	10	1652	1058	8	1050	961	7	954
e65.e68	1660	23	1637	1058	6	1052	960	14	946
e70.e90	1640	61	1579	1040	35	1005	950	42	908
i	1477	341	1136	935	230	705	856	229	627
i05.i09	1661	14	1647	1058	12	1046	961	10	951
i10.i15	1577	241	1336	1003	155	848	912	165	747
i20.i25	1625	97	1528	1038	64	974	933	64	869
i26.i28	1661	18	1643	1057	11	1046	959	10	949
i30.i52	1632	122	1510	1038	81	957	942	80	862
i60.i69	1658	33	1625	1054	22	1032	959	21	938
i70.i79	1653	21	1632	1051	12	1039	957	18	939
i80.i89	1596	79	1517	1015	61	954	930	58	872
i95.i99	1660	23	1637	1057	13	1044	960	12	948
j	1586	185	1401	1018	113	905	917	123	794
j00.j06	1656	12	1644	1057	8	1049	960	10	950
j09.j18	1660	36	1624	1057	21	1036	959	24	935
j20.j22	1650	40	1610	1051	29	1022	954	24	930
j30.j39	1636	30	1606	1044	19	1025	945	20	925
j40.j47	1627	94	1533	1039	54	985	944	64	880
j60.j70	1659	12	1647	1056	6	1050	958	9	949

j80.j84	1662	9	1653	1058	7	1051	961	5	956
j85.j86	1662	2	1660	1058	1	1057	961	2	959
j90.j94	1656	15	1641	1056	8	1048	959	10	949
j95.j99	1662	14	1648	1058	8	1050	961	12	949

	Horvath CpGs DNAme			UPLC IgG Glyomics			NMR Metabolomics		
Outcome	Total	Cases	Controls	Total	Cases	Controls	Total	Cases	Controls
all	757	288	469	1474	484	990	1297	425	872
c	870	61	809	1735	97	1638	1525	89	1436
c00.c14	889	1	888	1762	3	1759	1548	3	1545
c15.c26	885	15	870	1757	21	1736	1544	20	1524
c30.c39	888	3	885	1761	5	1756	1548	5	1543
c43.c44	887	7	880	1757	9	1748	1543	10	1533
c45.c49	888	3	885	1762	3	1759	1548	3	1545
c50.c50	885	10	875	1757	19	1738	1544	17	1527
c51.c58	888	6	882	1760	9	1751	1546	6	1540
c60.c63	883	7	876	1755	12	1743	1542	9	1533
c64.c68	889	8	881	1761	10	1751	1547	9	1538
c69.c72	889	1	888	1761	2	1759	1547	2	1545
c73.c75	889	1	888	1762	2	1760	1548	2	1546
c76.c80	888	24	864	1759	38	1721	1545	31	1514
c81.c96	889	6	883	1762	13	1749	1548	15	1533
e	865	133	732	1691	217	1474	1488	187	1301
e00.e07	880	65	815	1727	109	1618	1516	98	1418
e10.e14	882	37	845	1743	57	1686	1538	43	1495
e15.e16	888	2	886	1761	4	1757	1548	2	1546
e20.e35	889	4	885	1761	4	1757	1548	4	1544
e50.e64	889	7	882	1762	8	1754	1548	9	1539
e65.e68	888	14	874	1759	23	1736	1545	20	1525
e70.e90	879	40	839	1740	67	1673	1529	57	1472
i	796	208	588	1570	361	1209	1383	315	1068
i05.i09	889	9	880	1761	15	1746	1547	12	1535
i10.i15	846	150	696	1674	250	1424	1471	216	1255
i20.i25	865	57	808	1724	107	1617	1516	86	1430
i26.i28	887	10	877	1760	18	1742	1547	14	1533
i30.i52	873	70	803	1731	128	1603	1521	111	1410
i60.i69	887	21	866	1756	38	1718	1546	29	1517
i70.i79	885	15	870	1753	23	1730	1539	19	1520
i80.i89	861	52	809	1695	87	1608	1488	81	1407
i95.i99	888	12	876	1761	23	1738	1546	19	1527
j	850	117	733	1686	196	1490	1482	173	1309
j00.j06	889	10	879	1756	15	1741	1543	13	1530
j09.j18	887	22	865	1761	40	1721	1546	33	1513
j20.j22	882	21	861	1749	43	1706	1538	38	1500
j30.j39	875	19	856	1738	32	1706	1526	29	1497
j40.j47	873	62	811	1725	94	1631	1516	83	1433
j60.j70	886	9	877	1760	12	1748	1546	11	1535
j80.j84	889	5	884	1762	11	1751	1548	8	1540

j85.j86	889	2	887	1762	2	1760	1548	2	1546
j90.j94	888	8	880	1756	18	1738	1544	15	1529
j95.j99	889	11	878	1762	14	1748	1548	14	1534

	MS Fatty Acids Lipidomics			MS Metabolomics			MS Complex Lipidomics		
Outcome	Total	Cases	Controls	Total	Cases	Controls	Total	Cases	Controls
all	751	296	455	685	262	423	741	291	450
c	864	60	804	785	55	730	852	60	792
c00.c14	883	1	882	803	1	802	873	1	872
c15.c26	878	15	863	799	13	786	869	16	853
c30.c39	882	2	880	803	2	801	872	3	869
c43.c44	881	6	875	801	5	796	870	5	865
c45.c49	882	3	879	803	2	801	872	3	869
c50.c50	881	12	869	799	11	788	870	12	858
c51.c58	881	7	874	801	6	795	871	6	865
c60.c63	878	6	872	798	6	792	867	6	861
c64.c68	883	9	874	803	9	794	873	9	864
c69.c72	882	2	880	802	1	801	872	1	871
c73.c75	883	1	882	803	1	802	873	1	872
c76.c80	881	20	861	801	19	782	871	20	851
c81.c96	883	6	877	803	8	795	873	7	866
e	863	132	731	782	115	667	852	127	725
e00.e07	875	65	810	794	59	735	865	64	801
e10.e14	877	35	842	799	30	769	867	36	831
e15.e16	882	2	880	803	1	802	872	2	870
e20.e35	883	4	879	803	3	800	873	3	870
e50.e64	883	7	876	803	4	799	873	7	866
e65.e68	882	13	869	802	10	792	873	10	863
e70.e90	875	40	835	795	34	761	865	39	826
i	791	220	571	722	187	535	781	215	566
i05.i09	883	11	872	803	9	794	873	12	861
i10.i15	838	158	680	767	132	635	831	156	675
i20.i25	858	63	795	785	55	730	852	62	790
i26.i28	883	9	874	801	7	794	872	8	864
i30.i52	866	77	789	789	61	728	855	74	781
i60.i69	882	20	862	802	16	786	872	19	853
i70.i79	879	17	862	799	14	785	869	18	851
i80.i89	855	51	804	776	46	730	843	53	790
i95.i99	882	13	869	803	8	795	872	11	861
j	842	116	726	768	98	670	831	110	721
j00.j06	882	11	871	802	9	793	872	11	861
j09.j18	882	21	861	801	16	785	871	20	851
j20.j22	875	26	849	797	19	778	866	23	843
j30.j39	868	18	850	789	16	773	858	16	842
j40.j47	868	60	808	790	53	737	856	57	799
j60.j70	880	8	872	802	5	797	870	9	861
j80.j84	883	4	879	803	3	800	873	3	870

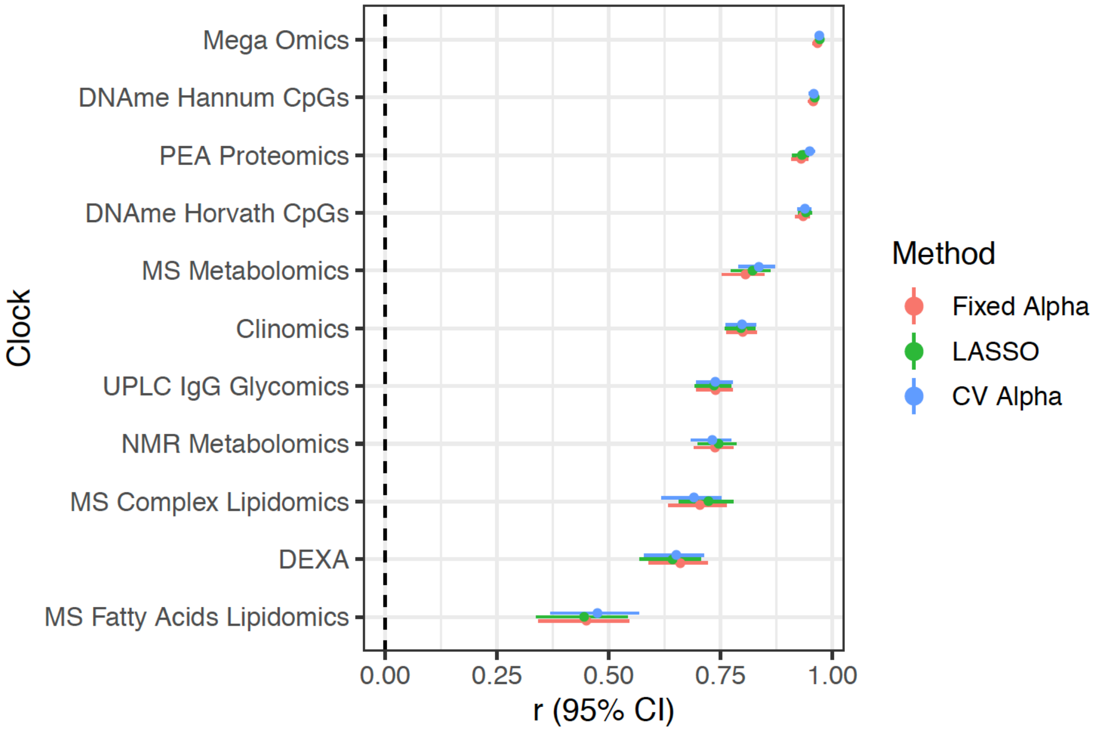
j85.j86	883	2	881	803	2	801	873	2	871
j90.j94	881	10	871	801	7	794	872	9	863
j95.j99	883	12	871	803	10	793	873	13	860

	PEA Proteomics			Mega Omics		
Outcome	Total	Cases	Controls	Total	Cases	Controls
all	751	296	455	685	262	423
c	864	60	804	785	55	730
c00.c14	883	1	882	803	1	802
c15.c26	878	15	863	799	13	786
c30.c39	882	2	880	803	2	801
c43.c44	881	6	875	801	5	796
c45.c49	882	3	879	803	2	801
c50.c50	881	12	869	799	11	788
c51.c58	881	7	874	801	6	795
c60.c63	878	6	872	798	6	792
c64.c68	883	9	874	803	9	794
c69.c72	882	2	880	802	1	801
c73.c75	883	1	882	803	1	802
c76.c80	881	20	861	801	19	782
c81.c96	883	6	877	803	8	795
e	863	132	731	782	115	667
e00.e07	875	65	810	794	59	735
e10.e14	877	35	842	799	30	769
e15.e16	882	2	880	803	1	802
e20.e35	883	4	879	803	3	800
e50.e64	883	7	876	803	4	799
e65.e68	882	13	869	802	10	792
e70.e90	875	40	835	795	34	761
i	791	220	571	722	187	535
i05.i09	883	11	872	803	9	794
i10.i15	838	158	680	767	132	635
i20.i25	858	63	795	785	55	730
i26.i28	883	9	874	801	7	794
i30.i52	866	77	789	789	61	728
i60.i69	882	20	862	802	16	786
i70.i79	879	17	862	799	14	785
i80.i89	855	51	804	776	46	730
i95.i99	882	13	869	803	8	795
j	842	116	726	768	98	670
j00.j06	882	11	871	802	9	793
j09.j18	882	21	861	801	16	785
j20.j22	875	26	849	797	19	778
j30.j39	868	18	850	789	16	773
j40.j47	868	60	808	790	53	737
j60.j70	880	8	872	802	5	797
j80.j84	883	4	879	803	3	800
j85.j86	883	2	881	803	2	801

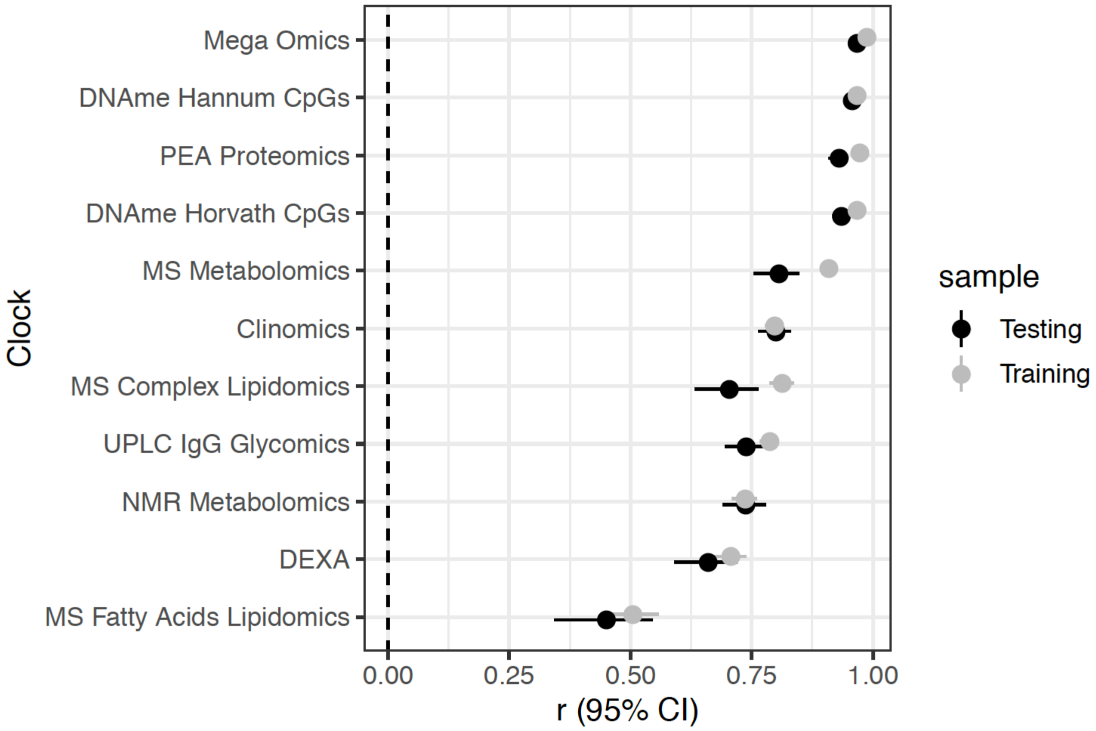
j90.j94	881	10	871	801	7	794
j95.j99	883	12	871	803	10	793

Supplementary Table 16. Summary Cases and Controls for Disease Blocks in ORCADES. *Outcome: the ICD10 codes defining a block. Total: Total number of samples in ORCADES with Martingale residuals for each disease block with each omics assay. Controls: the number of controls in the total. Cases: the number of cases in the total.*

Chapter 3 Supplementary Information



Supplementary Figure 26. Correlation of ChronAge and OCA were consistent, independent of penalised regression method. *Correlation (r) with 95% of confidence intervals of chronAge with omics clock estimated age (OCA) indicated on the y-axis via elastic net regression with a fixed alpha of 0.5, cross validated alpha and LASSO regression in the ORCADES testing sample.*



Supplementary Figure 27. Correlation of ChronAge and OCA in ORCADES Training and Testing Samples. *Correlation (r) with 95% of confidence intervals of chronAge with OCA indicated on the y-axis in the ORCADES Training and Testing samples.*

https://doi.org/10.7488/ds/3107

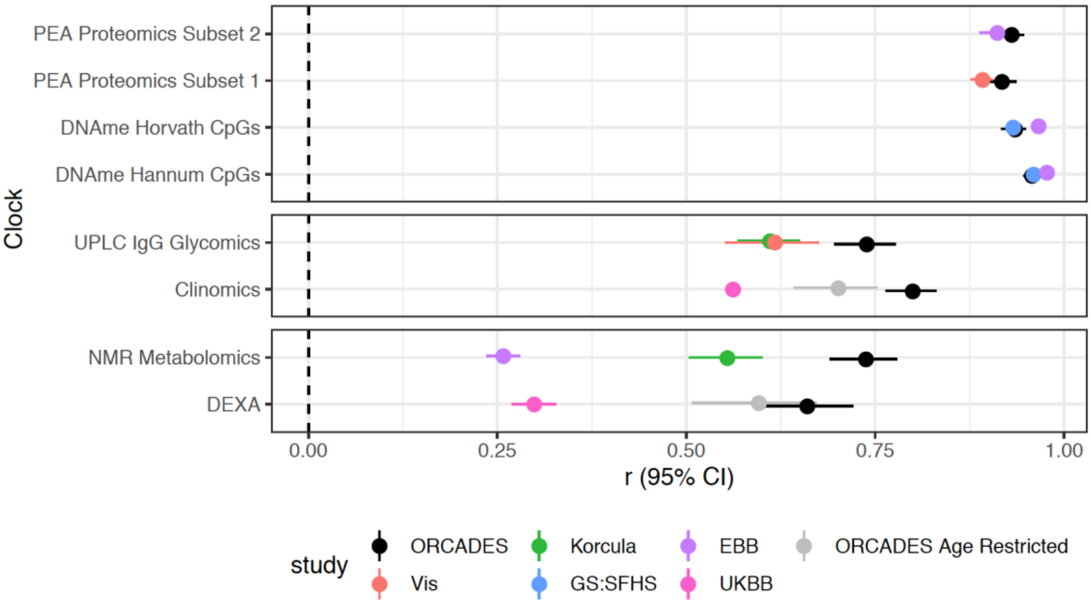
Supplementary Table 17. Biomarkers Selected for Model Inclusion Across Assays. *For each assay the description of each biomarker, variable_name: the biomarker ID for analysis. pass_qc: either 1 or 0 indicating whether each biomarker passed (1) or failed (0) quality control and therefore available for selection. selected: indicating whether the biomarker was selected for model inclusion (1) or not (0). available_core: if the biomarker was available for selection for the core model (1 available, 0 not) i.e. the biomarkers was selected for model inclusion in >95% of 500 iterations of clock construction. selected_core: if the biomarker was selected (1) or not (0) for inclusion in the core model. For the PEA Proteomics assay avail_sub_x: indicates that the biomarker was available for selection in the indicated protein subset, subset 1 being in the inflammation 1, cardiovascular II or cardiovascular III Olink panels used for validation in Croatia-Vis, subset 2 being in the Inflammation 1, cardiovascular II, cardiovascular III and Oncology II Olink panels used for validation in EGCUT. selected_sub_x: indicating that the biomarker was selected for model inclusion in the relevant protein subset clock, Panel: Olink panel.*

https://doi.org/10.7488/ds/3107

Supplementary Table 18. Clock Coefficients. *For each assay, predictor: the biomarker ID, coefficient: the clock coefficient.*

Omic	N	Mean Age	SD Age	Min Age	Max Age	% Female
DEXA	1158	55.85	14.19	18.02	88	59.93
DNAme Horvath CpGs	957	52.93	15.66	17.12	100.18	55.38
MS Fatty Acids Lipidomics	952	53.41	15.49	16.84	91.47	55.78
MS Metabolomics	861	52.81	15.05	17.12	90.79	57.38
Clinomics	1815	53.35	15.03	16.5	91.47	59.56
DNAme Hannum CpGs	1033	53.43	15.68	17.12	100.18	55.86
UPLC IgG Glycomics	1937	53.13	15.29	16.5	100.18	60.51
MS Complex Lipidomics	940	53.54	15.27	17.12	91.47	55.74
NMR Metabolomics	1643	52.96	14.91	16.5	91.47	59.95
PEA Proteomics	805	52.88	15.59	17.12	91.47	54.91
Mega Omics	796	53.1	15.31	17.12	91.47	56.78

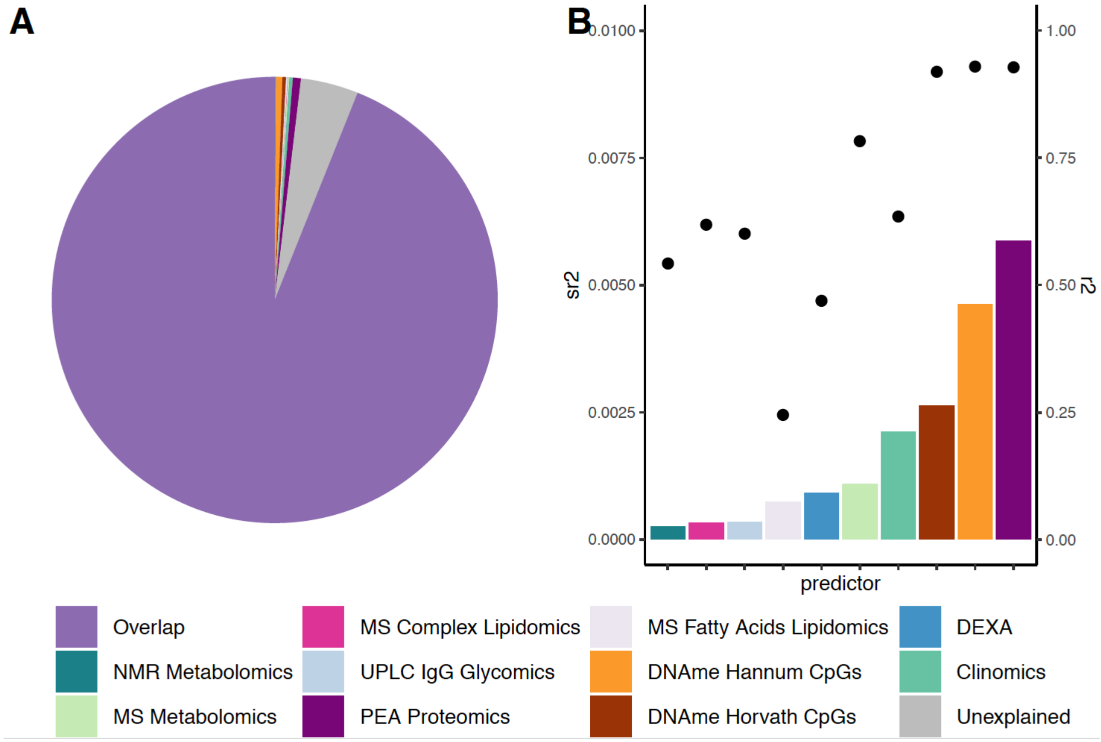
Supplementary Table 19. Age Characteristic of ORCADES Cohort. *Omic: Omic assay. N: number of individuals in ORCADES with the omics assay passing quality control. Mean Age: mean chronological age at venepuncture of ORCADES subset. SD Age: standard deviation of chronological age at venepuncture of ORCADES subset. Min Age: minimum chronological age at venepuncture of ORCADES subset. Max Age: maximum chronological age at venepuncture of ORCADES subset. % Female: percentage of ORCADES subset that is female.*



Supplementary Figure 28. Omics clocks trained in ORCADES predict chronAge in unrelated cohorts. *The correlation of OCA with ChronAge (x-axis) by the specified clock (y-axis). With the correlation in the ORCADES testing sample in black and additional populations as specified. The correlation in a restricted age range (40-75) ORCADES testing sample is shown in comparisons involving the UKBB shown in grey.*

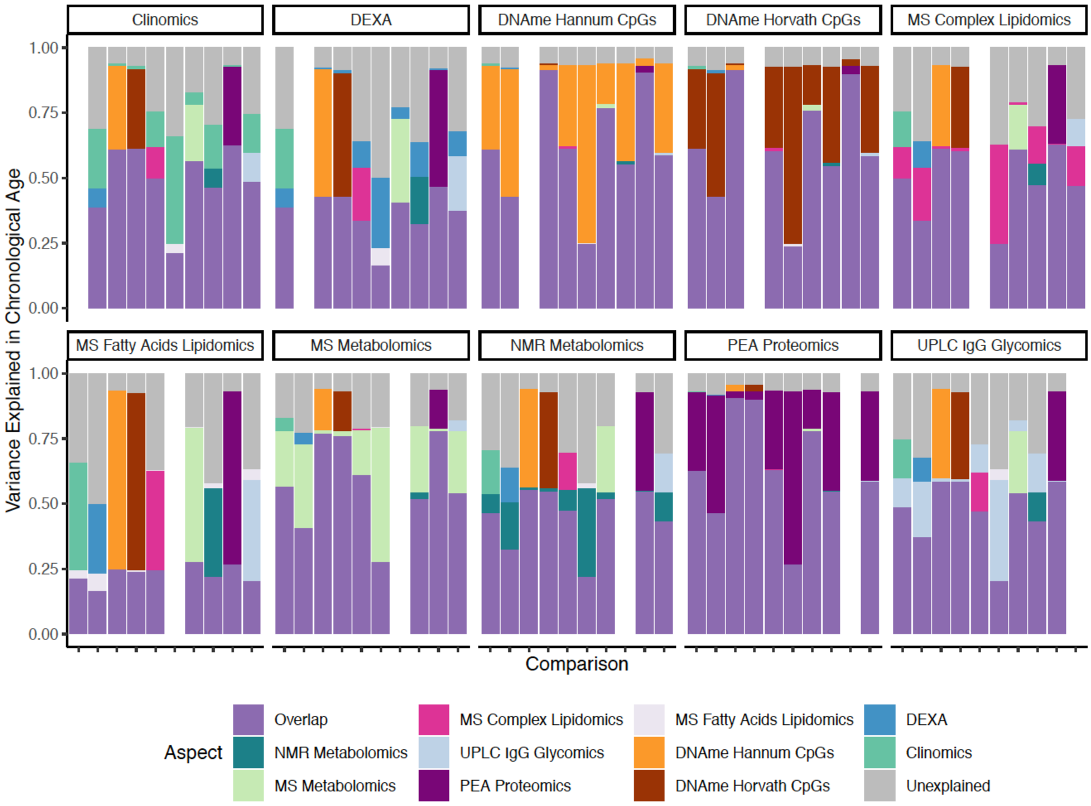
We found clocks built using the subsets of PEA proteomics measures available in our validation cohorts correlating with chronAge nearly as highly in Croatia-Vis (r=0.89) and EBB (r=0.91) as in the ORCADES testing sample (r=0.91 and r=0.93). Similarly, both Hannum and Horvath CpG based clocks achieved comparable correlations between OCA and chronAge in EBB (Hannum: r=0.98, Horvath: r=0.97) and GS:SHFS (Hannum: r=0.96, Horvath r=0.93) as in the ORCADES testing sample (Hannum: r=0.96, Horvath r=0.93). The UPLC IgG glycomics and Clinomics OCA were still correlated with chronAge in independent cohorts (UPLC IgG glycomics: r= 0.62 Croatia-Vis, r=0.61 Croatia-Korčula, Clinomics: r=0.56 UKBB) but less than in the ORCADES testing sample (UPLC IgG glycomics: r=0.74, Clinomics: r=0.80). There was correlation between NMR metabolomics estimated and chronAge in Croatia-Korčula, r=0.55 compared to r=0.73 in ORCADES however only a correlation of r=0.26 in EBB. Similarly, we found that the DEXA estimated age in UKBB correlated substantially lower with chronAge than in ORCADES (UKBB: r=0.30, ORCADES: r=0.66).

To assess whether the poor correlation of DEXA OCA and chronAge in UKBB was due to the difference in the ranges of chronAge of individuals in ORCADES compared to the UKBB we also compared a clock that was evaluated in ORCADES individuals between 40-75 (the recruiting age range of UKBB, compared to the 16-100 in the full ORCADES dataset). Despite the DEXA OCA having a lower correlation with chronAge in the age restricted ORCADES sample, r=0.60 compared with r=0.66 in the full age range sample, it is still drastically higher than the r=0.30 found in UKBB.



Supplementary *Figure 29*. Overlapping and unique variance in ChronAge explained across 10 omics clocks. *A) Partition of variance in ChronAge explained into that explained by 2 or more clocks (overlap), that not explained by any clock (unexplained), and that explained by each of the 10 clocks uniquely. Segments coloured by component explaining the variance in chronAge. B) squared part correlations (sr^2) (bars): unique variance in chronAge explained by each of the 10 clocks from figure A on the left-hand y-axis. R^2 (points) indicate the total variance explained in chronAge by each clock (right hand y-axis).*

Interestingly, the proportion of unique variance in chronAge explained by each OCA does not entirely mirror the univariate R^2 (black dots). It is important to note that the similarity between assays likely influences the proportion of unique variance in chronAge explained (at its most extreme, were a clock duplicated, it would explain no unique variance). This may explain why NMR metabolomics and MS complex lipidomics clocks have some of the lowest proportions of unique variance explained, despite NMR metabolomics and MS complex lipidomics having an R^2 higher than DEXA OCA and comparable to Clinomics. Interestingly, the DEXA and MS fatty acids lipidomics OCAs explain more unique variance than several clocks with higher R^2.

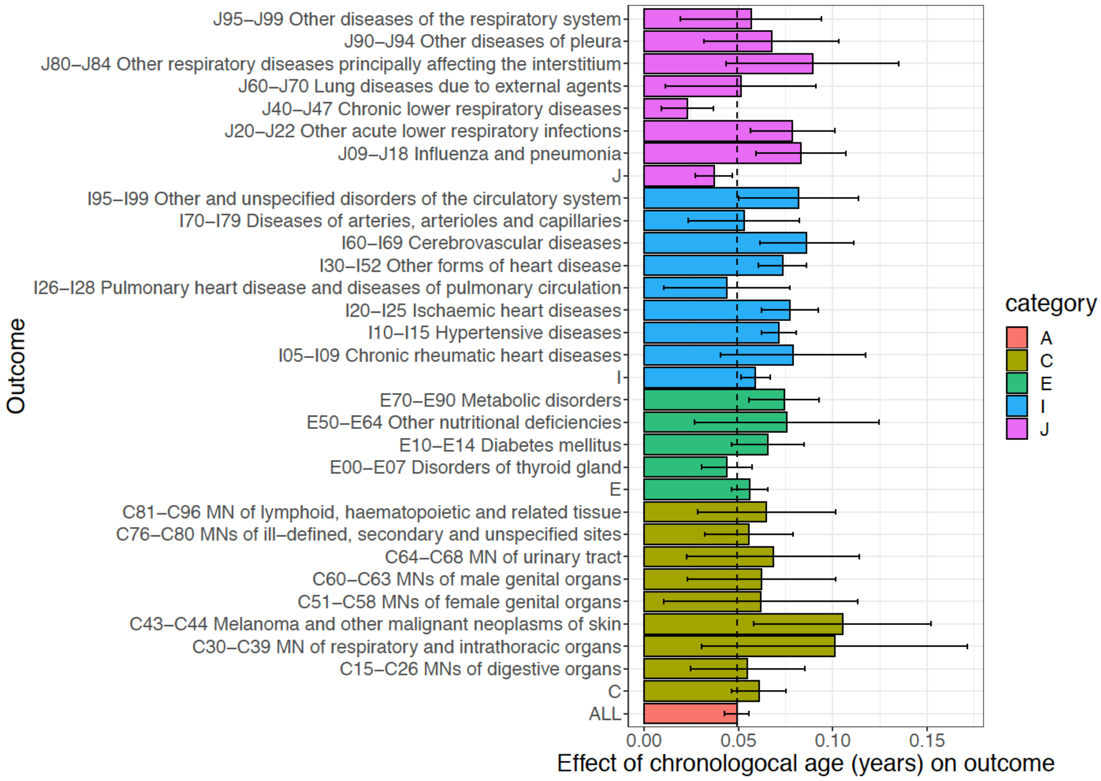


Supplementary Figure 30. Pairwise Comparisons of Variance Explained in ChronAge. *Pairwise comparison of variance in chronAge explained by OCA of the pairs of clocks in ORCADES. Comparison indicated on the x-axis, with the variance in chronAge explained on the y-axis. The colour of the bar indicates the aspect explaining the variance. For each comparison the proportion of variance explained by both clocks in the comparison (Overlap), the variance that remains unexplained fitting a bivariate model (unexplained) and the unique variance in chronAge explained by each of the two clocks in the comparison.*

Partly to consider the effect of two similar clocks affecting the unique variance explained, we performed pairwise comparisons, the unique variance in chronAge explained by each clock in the comparison was again calculated as the squared part correlation while controlling for the other clock in the pair. The overlap indicated is therefore the proportion of variance in chronAge explained by both clocks in the pair. Reiterating the results in **Supplementary Figure 29a**, **Supplementary Figure 30** shows that for 8 out of 10 clocks the mean percentage of variance explained in chronAge by both clocks (the overlap) is greater than 45%. The MS Fatty Acids Lipidomics and DEXA clocks had lower mean overlap, 23.2% and 36.9% respectively. Interestingly clocks that had higher correlations between OCA and chronAge, such as PEA Proteomics and DNAme based clocks were found to be contributing most of the additional variance in chronAge not explained by the overlap of both clocks. Conversely, the MS Fatty Acids Lipidomics clock, the clock with the lowest correlation between OCA and chronAge appears to contribute little of the additional variance in chronAge not already explained by the other clock across all comparisons.

Block	Title
J95-J99	Other diseases of the respiratory system
J90-J94	Other diseases of pleura
J85-J86	Suppurative and necrotic conditions of lower respiratory tract
J80-J84	Other respiratory diseases principally affecting the interstitium
J60-J70	Lung diseases due to external agents
J40-J47	Chronic lower respiratory diseases
J30-J39	Other diseases of upper respiratory tract
J20-J22	Other acute lower respiratory infections
J09-J18	Influenza and pneumonia
J00-J06	Acute respiratory infections
I95-I99	Other and unspecified disorders of the circulatory system
I80-I89	Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
I70-I79	Diseases of arteries, arterioles and capillaries
I60-I69	Cerebrovascular diseases
I30-I52	Other forms of heart disease
I26-I28	Pulmonary heart disease and diseases of pulmonary circulation
I20-I25	Ischaemic heart diseases
I10-I15	Hypertensive diseases
I05-I09	Chronic rheumatic heart diseases
E70-E90	Metabolic disorders
E65-E68	Obesity and other hyperalimentation
E50-E64	Other nutritional deficiencies
E20-E35	Disorders of other endocrine glands
E15-E16	Other disorders of glucose regulation and pancreatic internal secretion
E10-E14	Diabetes mellitus
E00-E07	Disorders of thyroid gland
C81-C96	Malignant neoplasm of lymphoid, haematopoietic and related tissue
C76-C80	Malignant neoplasms of ill-defined, secondary and unspecified sites
C73-C75	Malignant neoplasms of thyroid and other endocrine glands
C69-C72	Malignant neoplasms of eye, brain and other parts of central nervous system
C64-C68	Malignant neoplasm of urinary tract
C60-C63	Malignant neoplasms of male genital organs
C51-C58	Malignant neoplasms of female genital organs
C50-C50	Malignant neoplasm of breast
C45-C49	Malignant neoplasms of mesothelial and soft tissue
C43-C44	Melanoma and other malignant neoplasms of skin
C30-C39	Malignant neoplasm of respiratory and intrathoracic organs
C15-C26	Malignant neoplasms of digestive organs
C00-C14	Malignant neoplasms of lip, oral cavity and pharynx

Supplementary Table 20. ICD Block Definitions. *Block: block coding. Title: ICD meaning.*

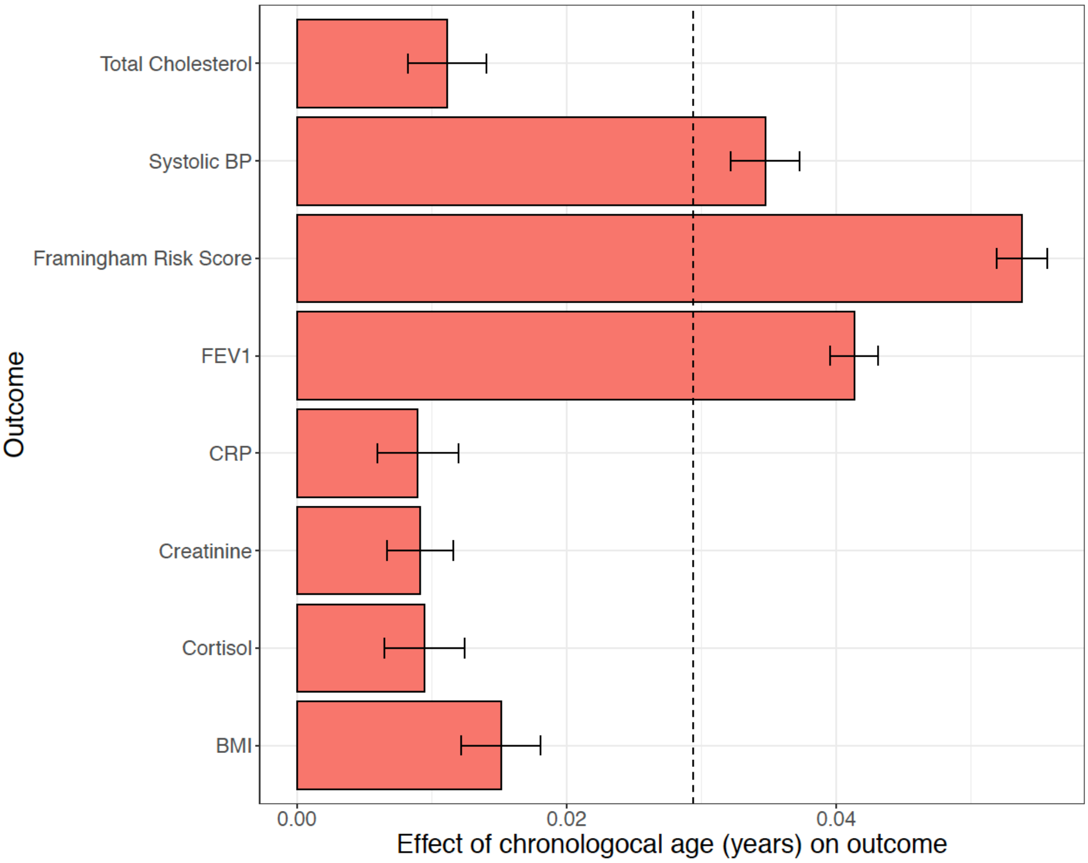


Supplementary Figure 31. Associations of disease incidence with chronAge. *Effect and its 95% CI: the log$_e$HR of chronAge on the incidence of the disease since participation, using a Cox Model. ICD 10 Chapters (i.e. whole Categories) count the first occurrence (post assessment) of any disease within the letter/category/chapter (including those blocks dropped from the individual block analysis due to lack of power) as incidence. Participants prevalent at assessment (i.e. a recorded prior incidence) within any grouping at assessment were excluded from the analysis of that grouping. The dashed line represents the hazard of age on any occurrence of the disease chapters under consideration, a hazard ratio of 0.0492, representing a doubling of incidence rate every 14 years. Distinctions in observed individual effects sizes from this were (visually) judged more materially due to sampling variance than true effects, and so that single factor was chosen as our best estimate of the age effect on each disease. MNs: Malignant neoplasms. Associations are only shown for those disease groups that passed QC and were taken forward to association testing with OCAA.*

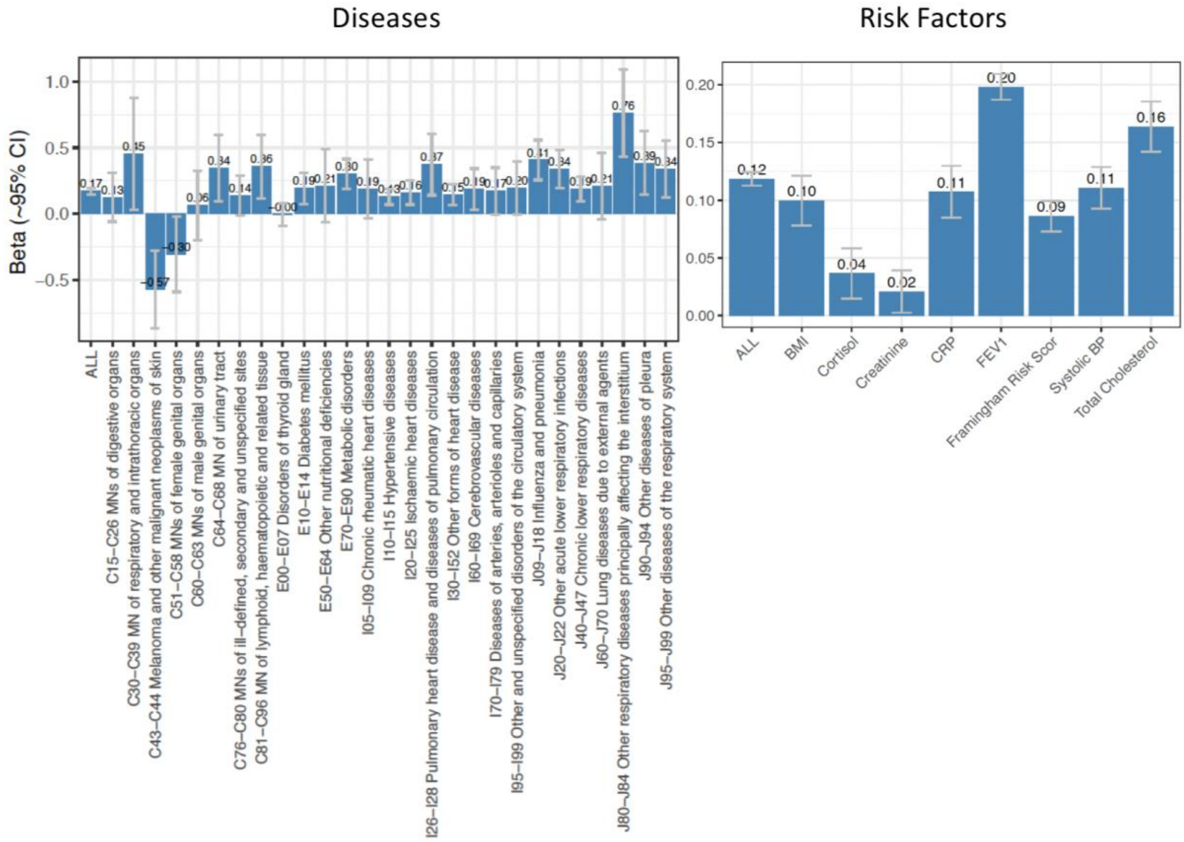
8/8 risk factors and 43/44 disease groupings associated with chronAge in the expected positive direction, except for cortisol and FEV1 which decline with chronAge. The disease exception, J00-J06 Acute respiratory infections, was not nominally significantly different from zero (log$_e$HR/SE -0.025/0.017). All the risk factors, and 34 of the disease groupings associations were significant after allowing for multiple testing (passed FDR 10% risk factors and diseases considered separately, one sided test H_1:b>0). For these 34 groupings there thus was reasonable power to detect associations with chronAge and so potentially biological OCAA. 2 disease groups had fewer than 5 cases and were excluded from the subsequent analysis, to further limit the burden of multiple testing.

The effect (log$_e$HR/SE) of one year of chronAge at outset on the first incidence of any of the diseases was 0.0492/0.00323, a doubling roughly every 14 years. This pattern was generally similar to the estimated effects for each disease individually, noting these are on the same (logistic) scale. With the largest observed differences arising from diseases with larger standard errors. However, the effect ($\log_e$HR/SE) of one year of chronAge on the risk factors varied more, although again they were on the same (standardised) scale. FRS and FEV1 (0.054/0.00083 and -0.041/0.00088) were most sensitive, whilst CRP and creatinine were less sensitive (effect/SE of 1 year of chronAge on standardised trait 0.0092/0.0012 and 0.0090/0.0015 respectively) as shown in Supplementary Figure 6b, whilst standard errors of the effect sizes were generally smaller (as a proportion of the effect).

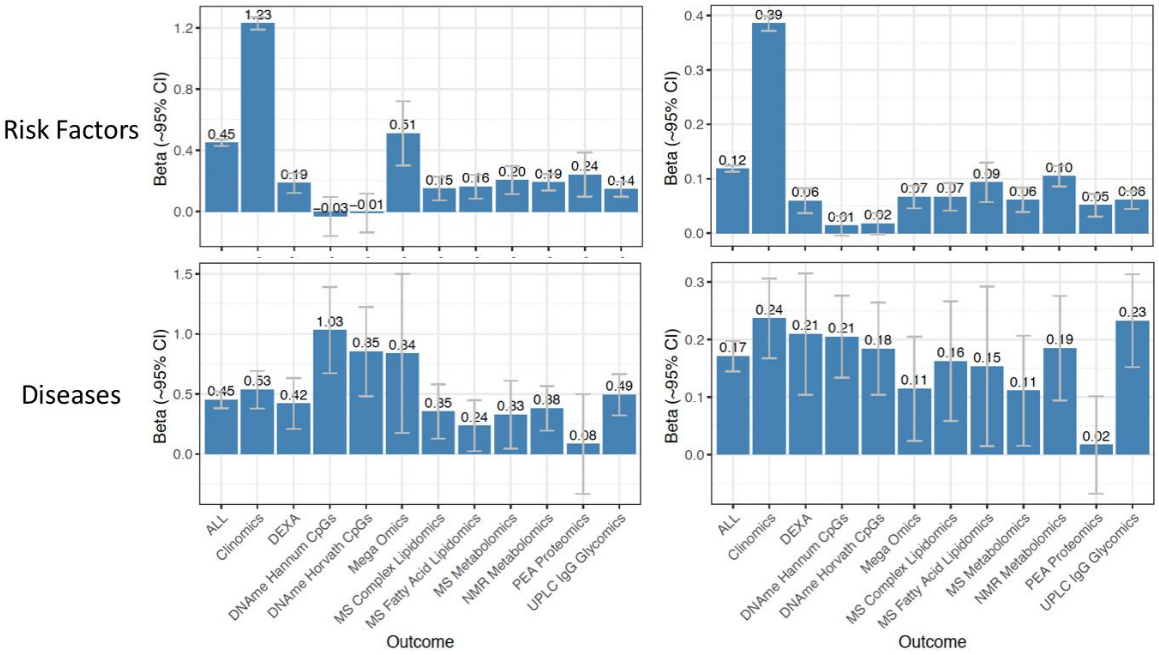
The remaining 32 disease groups along with all the risk factors were taken forward to association testing with OCAA, using the same models, with age and sex as covariates. Power was expected to be lower, due to lower variation and attenuation in OCAA compared to chronAge. As our principal purpose was to examine the effect of OCAA compared to chronAge, effect sizes of OCAA were rescaled so that the effect of one year of chronAge was one. This was done by dividing the observed effect of OCAA by the effect of chronAge on the outcomes. Given the similarities of the chronAge effects (and wide SEs) for diseases, this was done using the single factor 0.049187 $\log_e$HR. Whereas for risk factors this was done trait-by-trait (the effect of chronAge on the single trait) as these effects varied more and had lower standard errors.



Supplementary Figure 32. The strength of associations of risk factors with chronAge varies. *FEV1: Forced expiratory volume one second, CRP: C-reactive Protein, BMI: Body Mass Index. Effect: the estimated increase (and 95% CI) in standardised trait per year of chronAge using a linear model, with sex as a covariate. Traits which decrease as age increases (FEV1, cortisol) have been converted to ageing traits, by reversing their signs.*



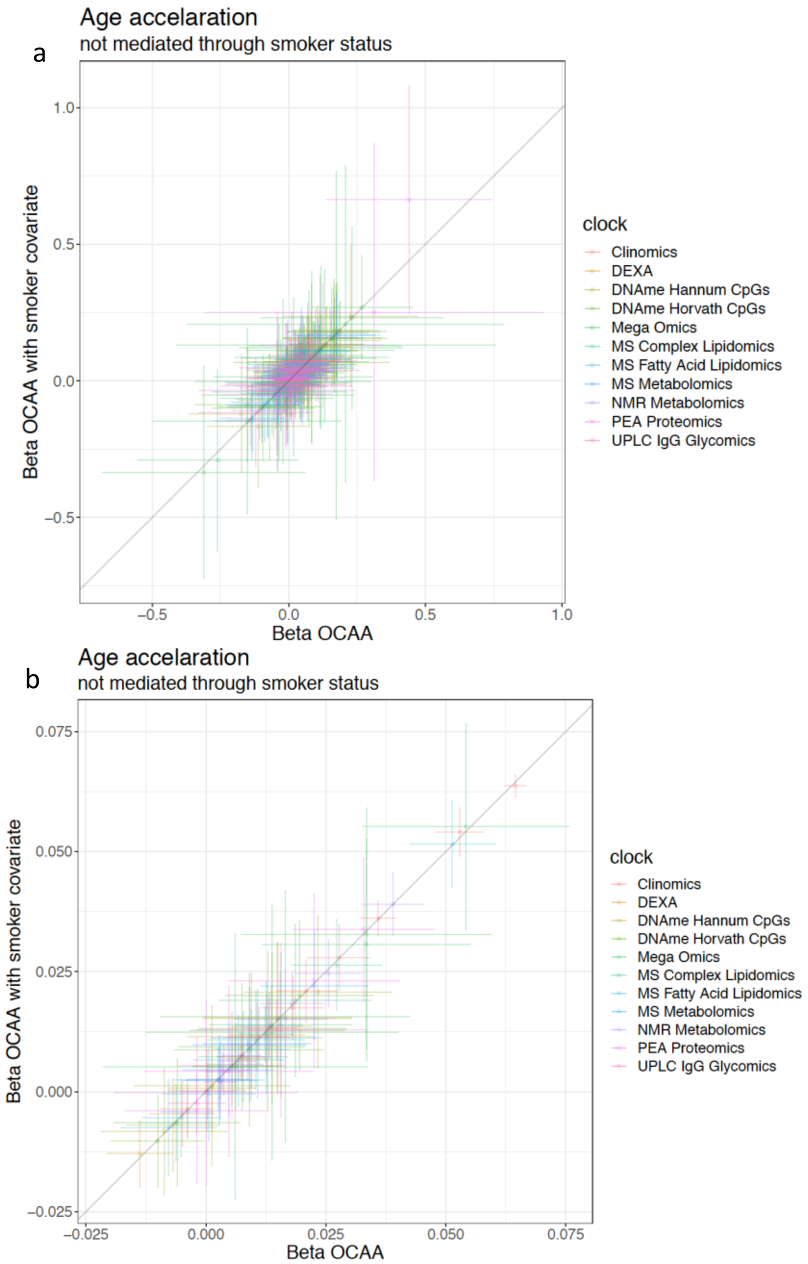
Supplementary Figure 33. Average effect across clocks of standardised OCAA upon outcome. *Beta: the observed effect of OCAAs on outcome. Beta was IVW averaged across OCAAs. SEs were calculated as the inverse root sum of the precisions (not strictly valid given correlated tests). Error bars shown are ± 2SEs. OCAA: omics clock estimated age acceleration.*



OCAA (years) effect per year ChronAge Effect

Standardised OCAA (years) effect – $\log_e$HR/standardised Risk Factor

Supplementary Figure 34. Averaged effects of OCAA across diseases and risk factors. *The Left-hand side shows the effect of OCAA in years per year of chronAge effect (OCAA effect divided by chronAge effect) IVW averaged across outcomes (either risk factors or diseases as specified on the y-axis). The right-hand side shows the effect of standardised OCAA (units of phenotypic standard deviation) IVW averaged across outcomes. Beta: the observed effect of OCAA on outcome. Beta was IVW averaged across outcomes. SEs were calculated as the inverse root sum of the precisions (not strictly valid given correlated tests). Error bars shown are ± 2SEs. OCAA: omics clock estimated age acceleration.*

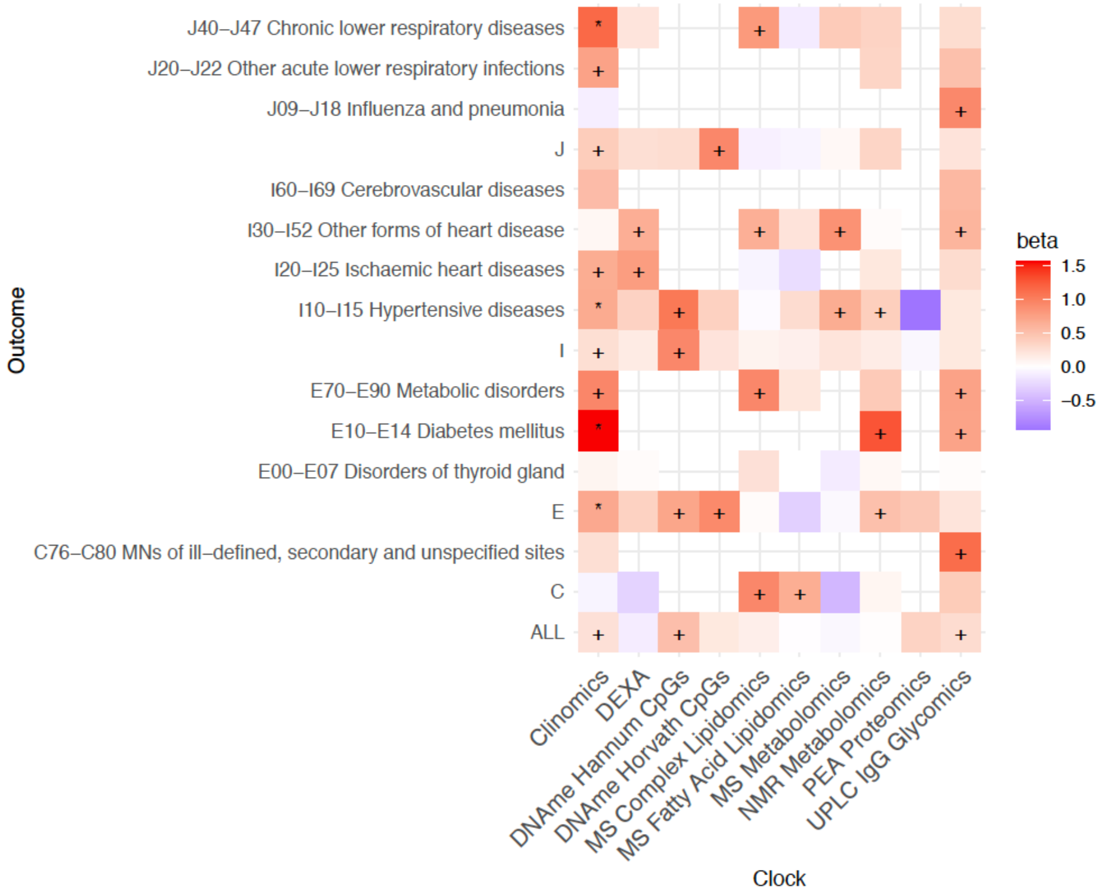


Supplementary Figure 35. Fitting smoking as a covariate does not appear to materially affect the association between OCAA and a) diseases or b) risk factors. *Beta OCAA - the observed effect of OCAA on the outcome under the models (see main text). Beta OCAA with smoker covariate - the observed effect of OCAA on the outcome under the same model, but with smoking fitted.*

Across all the associations studied for 11 clocks against 32 diseases and 8 risk factors, we found that the IVW ratio of the estimated effect of OCAA with and without smoking fitted as a covariate were 1.023 and 1.008 respectively. Individual test p-values for the ratio of the effects not being one all exceeded 0.4. Visual analysis confirmed these results: that smoking was not a material confounder of health-OCAA associations.

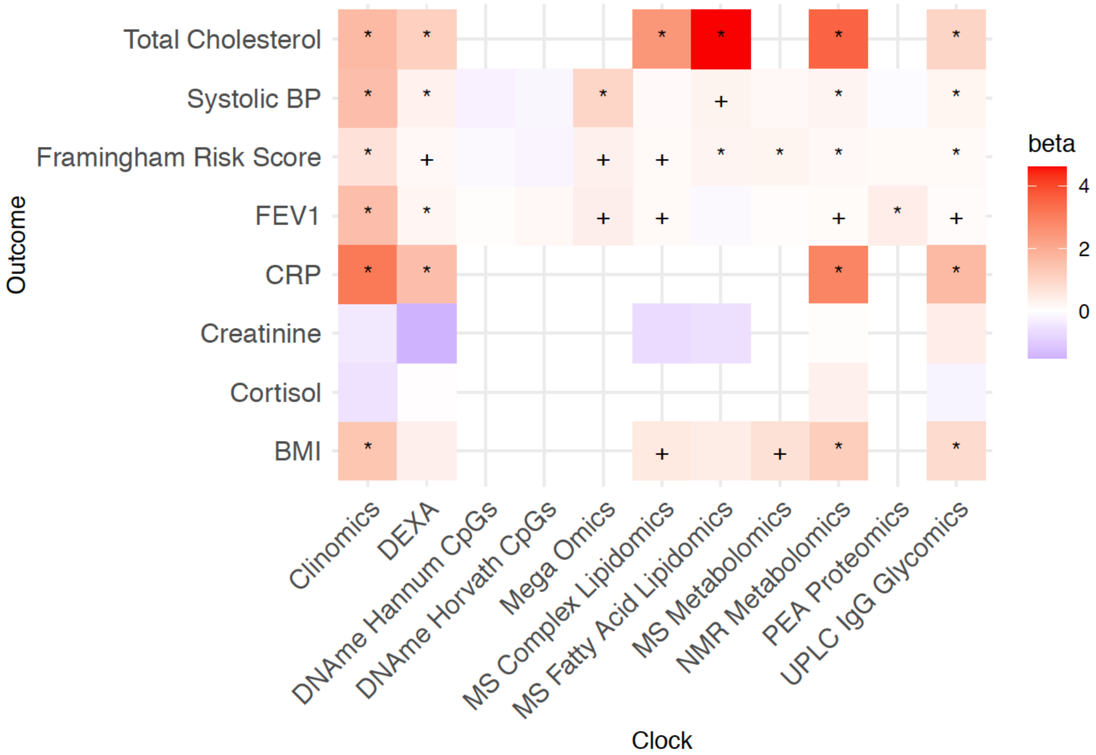
https://doi.org/10.7488/ds/3107

Supplementary Table 21. Summary of the association test between health outcomes, Chronological age and OCAA. *For diseases, A Cox model was fitted with time since outset specifying the base hazard and age as a proportional hazard, with first occurrence of hospitalisation for any disease in the ICD-10 block as event. Prevalent cases at outset were excluded. For quantitative traits, a linear model testing the effect of age on the outcome was tested. FEV1 and Cortisol were made negative to give traits which associate positively with age. Outcome: the quantitative trait or disease outcome under consideration. ICD blocks were analysed as a whole as were Chapters (single letters) for the sets of blocks considered, in which case first occurrence of any disease in the chapter was considered the event. ALL was defined as any disease in the chapters considered. N: the number of subjects included. Cases: the number of cases observed. BETA: the linear effect/log$_e$ Cox hazard ratio effect for the hazard (age). SE: the standard error of beta. Z: the z-statistic for the test of association (BETA/SE). P: the one-sided p-value for the test of positive association. Trait Mean: mean of the trait across the subjects in the analysis. Trait SD: the standard deviation of the trait across the subjects in the analysis. Q: the Benjamini-Hochberg FDR, allowing for all tests of age, within the disease and risk factors separately.*

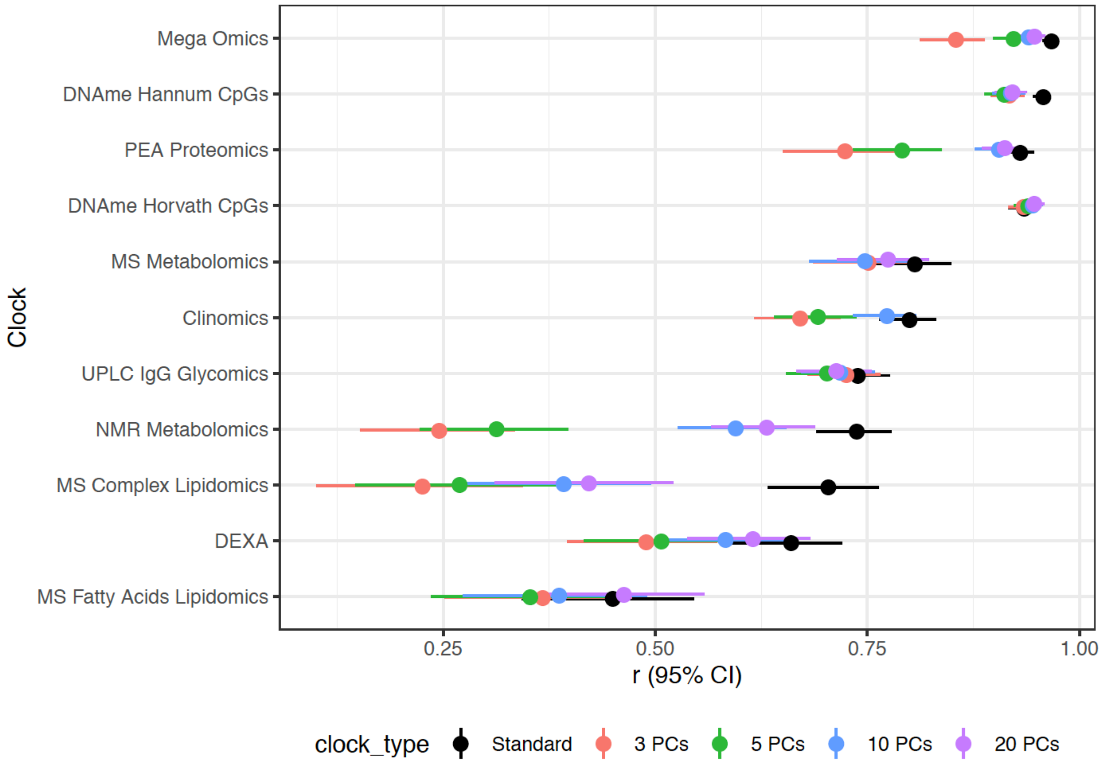


Supplementary Figure 36. OCAA positively (pink), and often significantly, associates with disease incidence in most cases where there is reasonable power. *+/* Association nominally (p<5%)/FDR 10% significant in the frequentist test of H_1: b>0 (FDR is determined across all tests shown in* ***Supplementary Table 21****, not just those shown here). Beta: the relative effect of a year of OCAA to a year on chronAge on outcome (measured in $\log_e$ hazard ratios). A value of one means the estimate of the effect of chronAge and OCAA are the same. Clock: the omics clock on which OCAA was measured. The mega-omics clock is not shown as it never met the SE<0.5 criterion. Disease group: the set of diseases (defined by ICD 10 codes) which were tested for first incidence after assessment against the clock (already prevalent cases were excluded).*

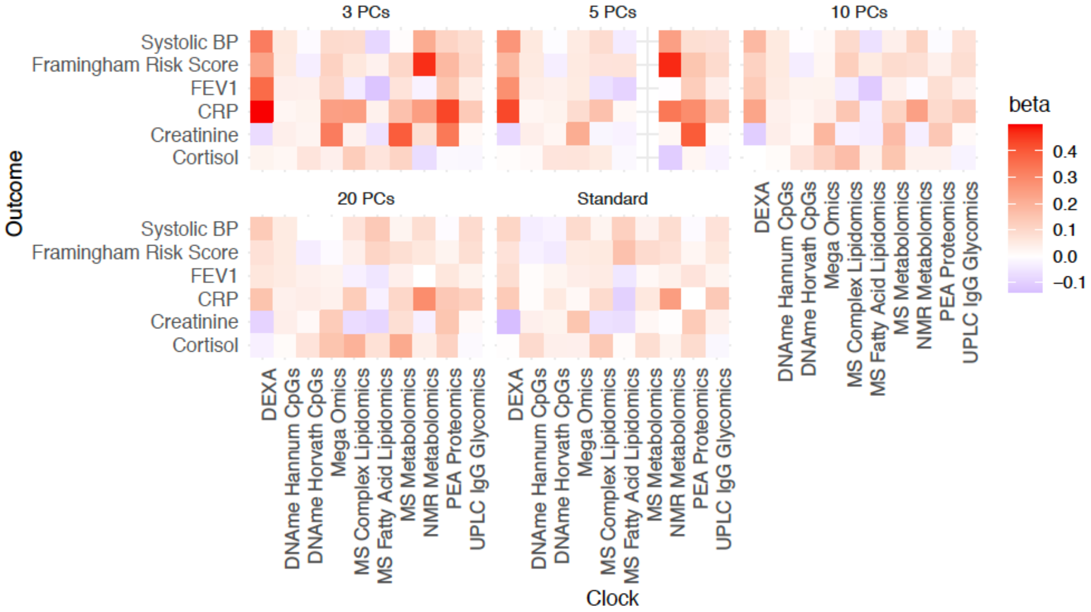
The presence of an entry in this figure denotes power, whilst its intensity denotes the size of the effect. Clinomics OCAA is thus relatively powerful and has large effects, as does the UPLC IgG Glycomics OCAA, albeit to a lesser extent. The more accurate clocks at estimating chronAge such as DNAme and PEA Proteomics based clocks on the other hand, show less power, although the Horvath CpGs clock does show some reasonably strong effect sizes.



Supplementary Figure 37. OCAA positively (pink) and often significantly associates with risk factors in most cases where there is reasonable power. *+/* Association nominally (p<5%)/FDR10 % significant in the frequentist test of* H_1*: b>0 (FDR is determined across all tests shown in* ***Supplementary Table 21****, not just those shown here). Beta: the relative effect of a year of OCAA to a year on chronAge on outcome. A value of one means the estimate of the effect of chronAge and OCAA are the same.*



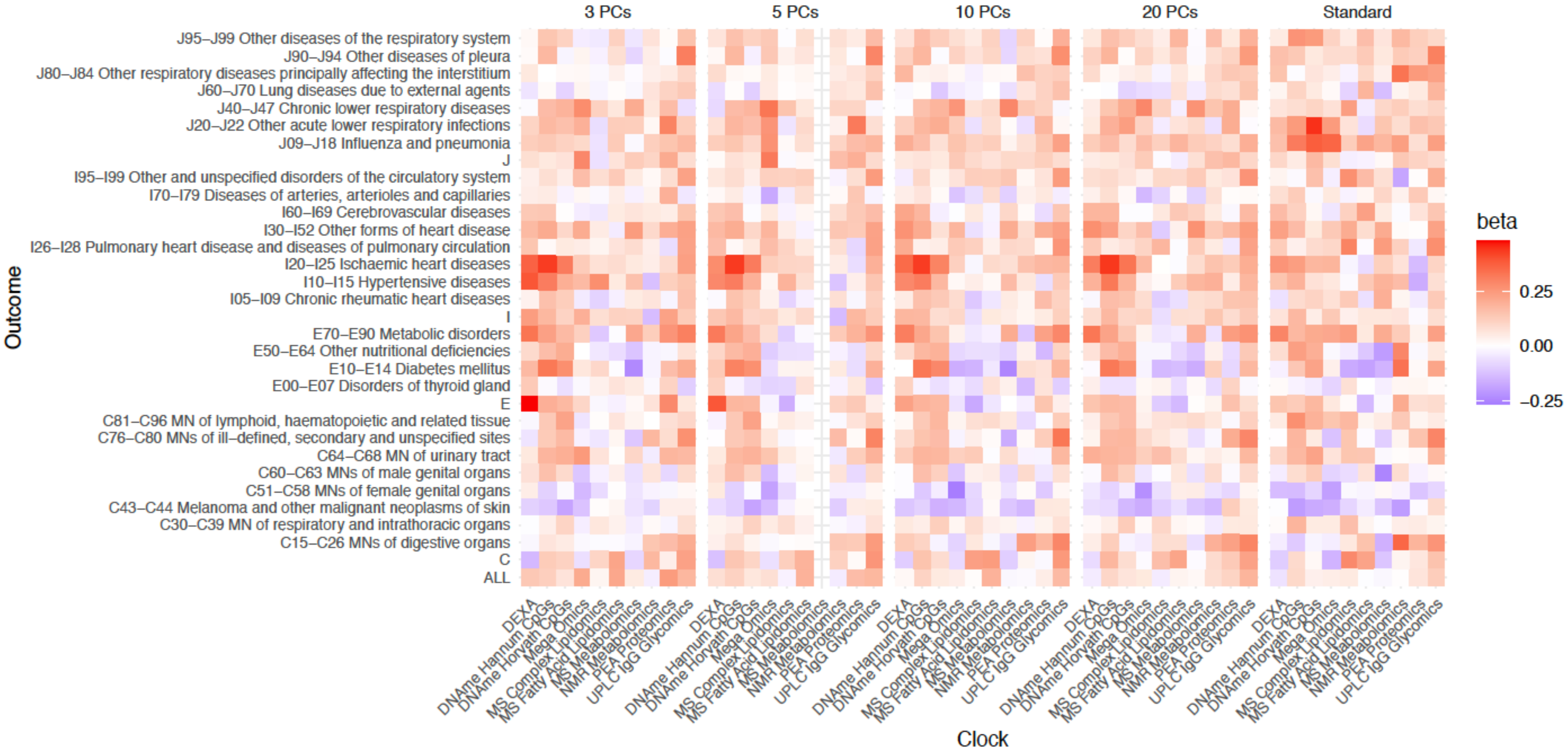
Supplementary Figure 38. Correlation of chronAge and OCA from clocks built using 3, 5, 10, 20 PCs. *Correlation (r) and 95% confidence interval of chronAge and OCA indicated on the y-axis using models constructed from 3, 5, 10 and 20 principal components of the assay in the ORCADES testing sample compared to the standard clock (black).*



Supplementary Figure 39. Reducing dimensionality of omics dataset used to build clocks increases the predictive ability of OCAA for risk factors. *Beta: the effect of a year of standardised (within clock) OCAA on outcome (effect sizes for standardised risk factors). Estimates were shrunk using a prior to reduce the possibility that frequentist best estimate beta was predominantly a consequence of a large SE. Clock: the omics clock on which OCAA was measured. Cholesterol/BMI which showed a particularly large effect from MS Fatty Acids Lipidomics/DEXA OCAA, excluded here to aid visualisation. X PCs: the number of PCs of the omic used as predictors to create the chronAge and OCAA measures.*

OCAAs (and risk factors) were standardised, whilst hazards were left on their $\log_e$ scale. The resultant measures of the effect of OCAA on outcome gave a measure of the ability of the OCAA to distinguish amongst individuals.

Clinomics was excluded from this analysis as it was based on only 12 predictors. We continued to exclude Total Cholesterol, but also excluded BMI as too close in nature to some of the predictors used.



Supplementary Figure 40. Reducing dimensionality of omics dataset to train ChronAge makes little difference to the predictive ability of OCAA for diseases. *Beta: the effect of a year of standardised (within clock) OCAA on outcome (measured in log$_e$ hazard ratios). Estimates were shrunk using a prior to reduce the possibility that frequentist best estimate beta was predominantly a consequence of a large SE. Clock: the omics clock on which OCAA was measured. Disease group: the set of diseases (defined by ICD 10 codes) which were tested for first incidence after assessment against the clock (already prevalent cases were excluded). X PCs: the number of PCs of the omic used as predictors to create the chronAge and OCAA measures.*

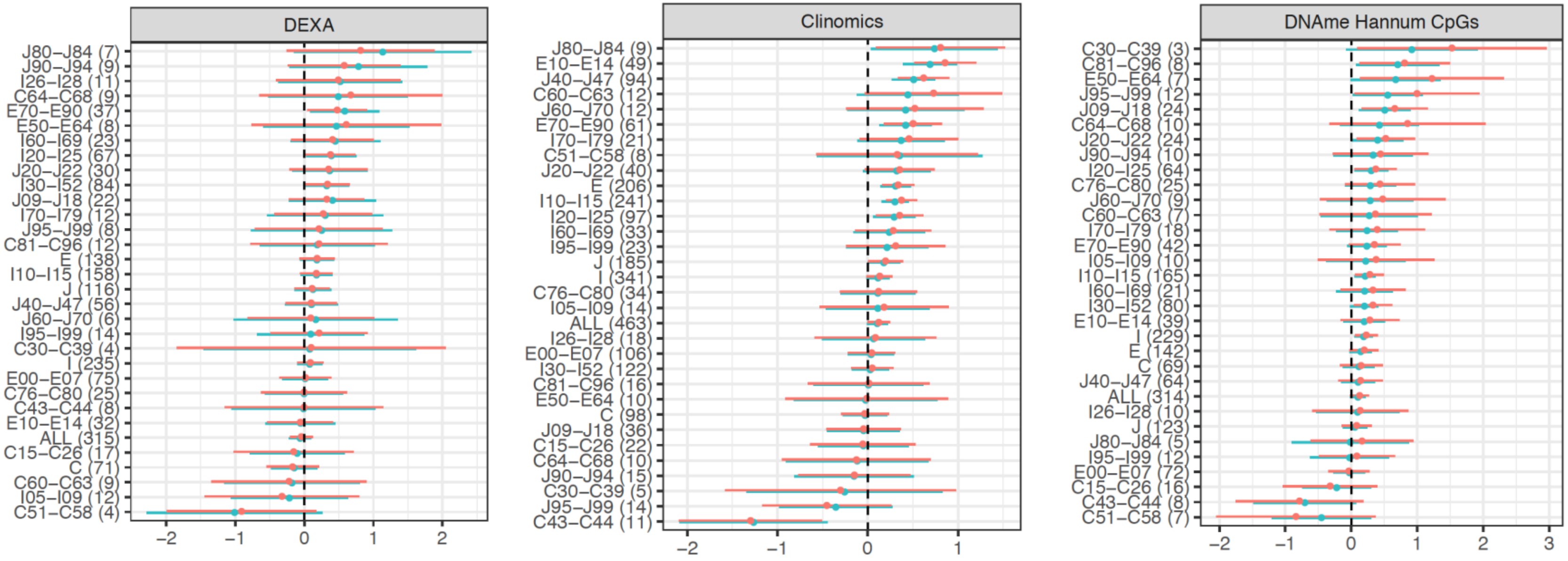
Chapter 4 Supplementary Information

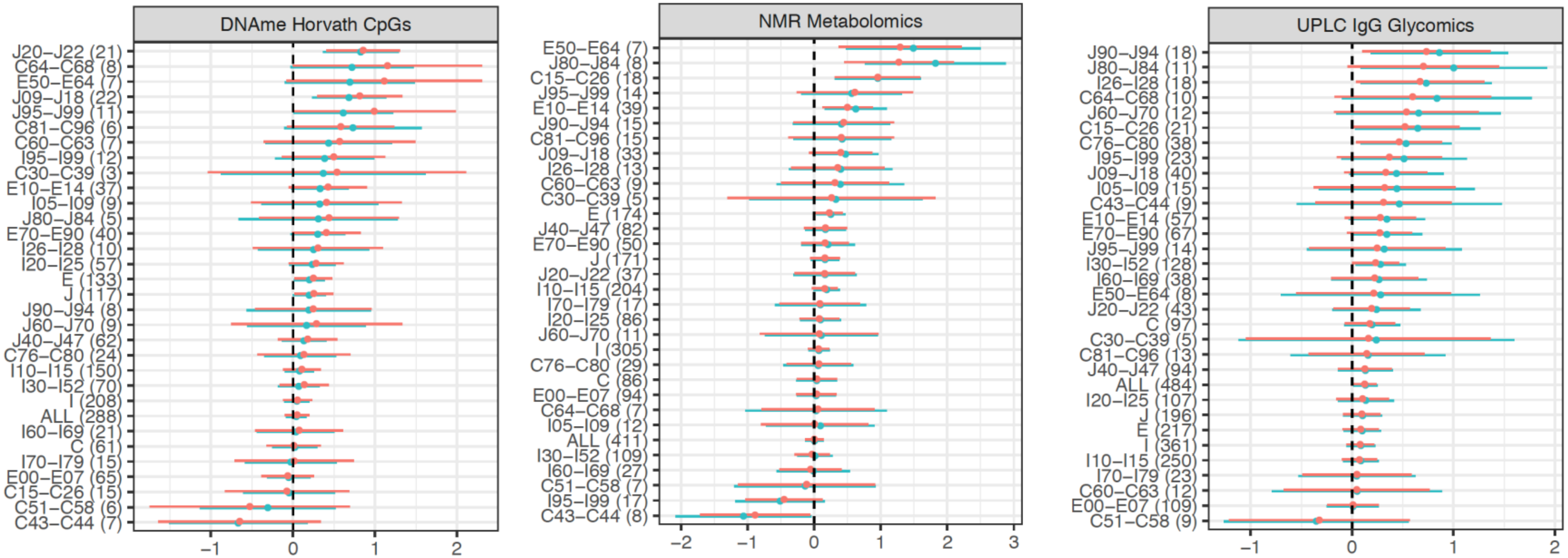
https://doi.org/10.7488/ds/3107

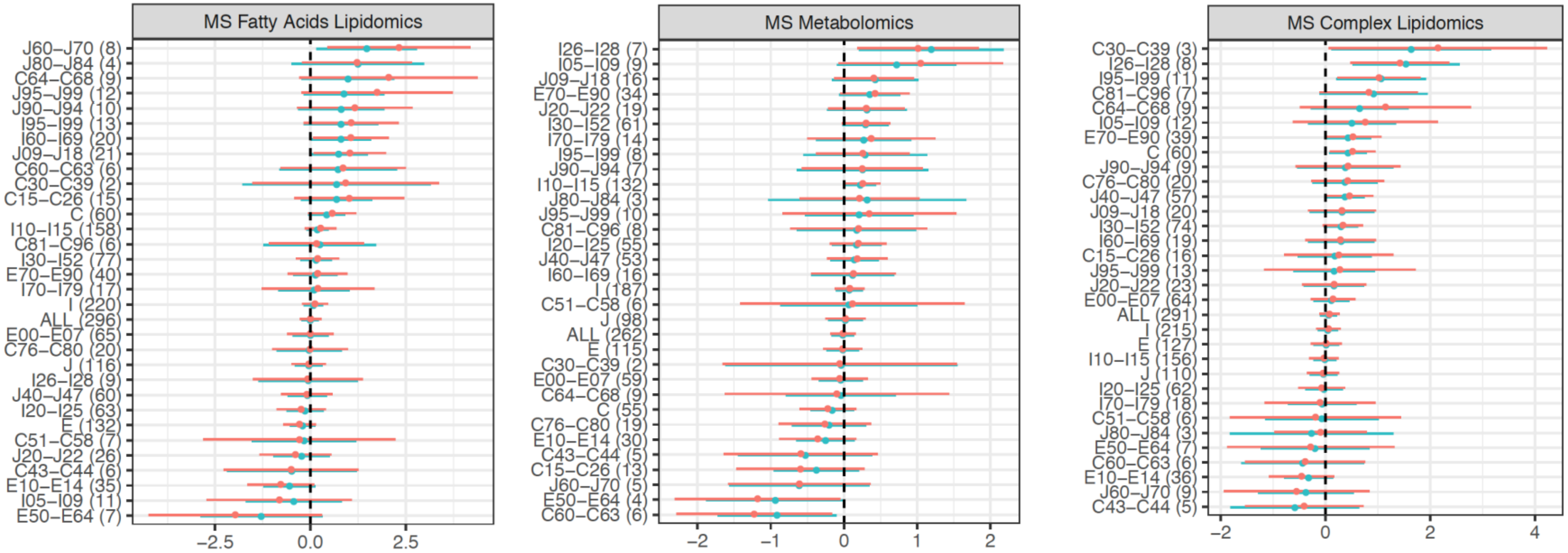
Supplementary Table 22. Biomarkers from all Assays Passing QC. *For each assay the predictor: the biomarker ID for analysis. available: either 1 or 0 indicating whether each biomarker passed (1) or failed (0) quality control and therefore available for selection across any of the outcomes.*

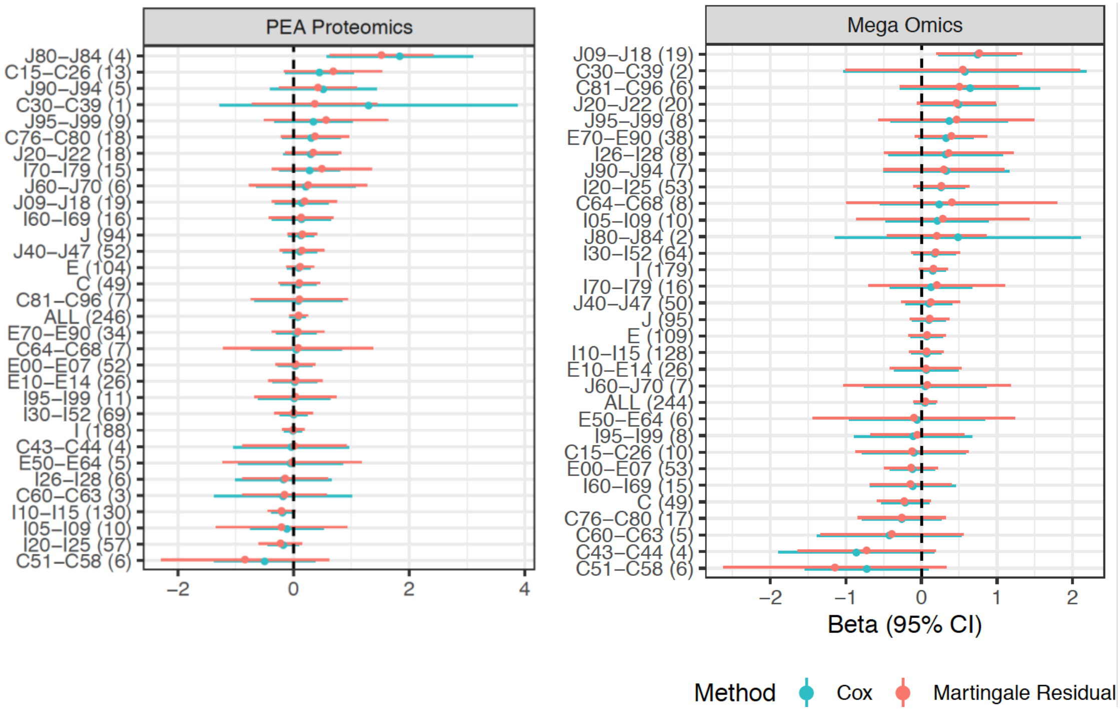
https://doi.org/10.7488/ds/3107

Supplementary Table 23. Coefficients for all Assay-Outcome LASSO regression Models.

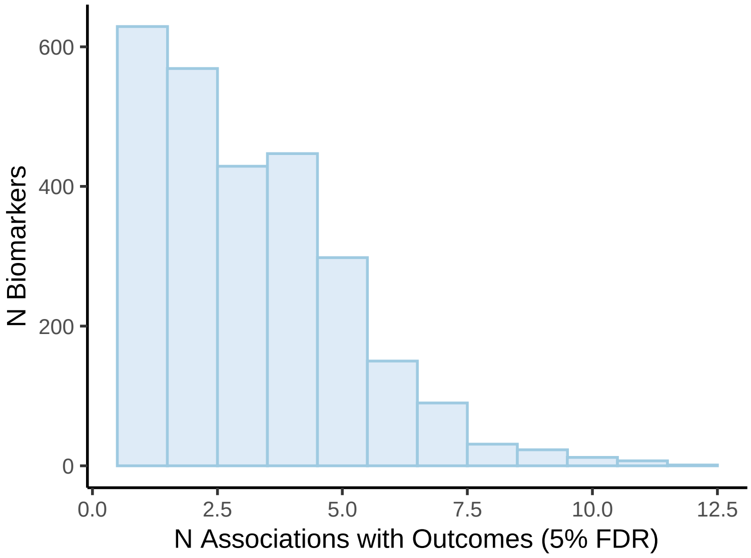

DEXA
J80–J84 (7)
J90–J94 (9)
I26–I28 (11)
C64–C68 (9)
E70–E90 (37)
E50–E64 (8)
I60–I69 (23)
I20–I25 (67)
J20–J22 (30)
I30–I52 (84)
J09–J18 (22)
I70–I79 (12)
J95–J99 (8)
C81–C96 (12)
E (138)
I10–I15 (158)
J (116)
J40–J47 (56)
J60–J70 (6)
I95–I99 (14)
C30–C39 (4)
I (235)
E00–E07 (75)
C76–C80 (25)
C43–C44 (8)
E10–E14 (32)
ALL (315)
C15–C26 (17)
C (71)
C60–C63 (9)
I05–I09 (12)
C51–C58 (4)
−2
−1
0
1
2
Clinomics
J80–J84 (9)
E10–E14 (49)
J40–J47 (94)
C60–C63 (12)
J60–J70 (12)
E70–E90 (61)
I70–I79 (21)
C51–C58 (8)
J20–J22 (40)
E (206)
I10–I15 (241)
I20–I25 (97)
I60–I69 (33)
I95–I99 (23)
J (185)
I (341)
C76–C80 (34)
I05–I09 (14)
ALL (463)
I26–I28 (18)
E00–E07 (106)
I30–I52 (122)
C81–C96 (16)
E50–E64 (10)
C (98)
J09–J18 (36)
C15–C26 (22)
C64–C68 (10)
J90–J94 (15)
C30–C39 (5)
J95–J99 (14)
C43–C44 (11)
−2
−1
0
1
DNAme Hannum CpGs
C30–C39 (3)
C81–C96 (8)
E50–E64 (7)
J95–J99 (12)
J09–J18 (24)
C64–C68 (10)
J20–J22 (24)
J90–J94 (10)
I20–I25 (64)
C76–C80 (25)
J60–J70 (9)
C60–C63 (7)
I70–I79 (18)
E70–E90 (42)
I05–I09 (10)
I10–I15 (165)
I60–I69 (21)
I30–I52 (80)
E10–E14 (39)
I (229)
E (142)
C (69)
J40–J47 (64)
ALL (314)
I26–I28 (10)
J (123)
J80–J84 (5)
I95–I99 (12)
E00–E07 (72)
C15–C26 (16)
C43–C44 (8)
C51–C58 (7)
−2
−1
0
1
2
3


DNAme Horvath CpGs
J20–J22 (21)
C64–C68 (8)
E50–E64 (7)
J09–J18 (22)
J95–J99 (11)
C81–C96 (6)
C60–C63 (7)
I95–I99 (12)
C30–C39 (3)
E10–E14 (37)
I05–I09 (9)
J80–J84 (5)
E70–E90 (40)
I26–I28 (10)
I20–I25 (57)
E (133)
J (117)
J90–J94 (8)
J60–J70 (9)
J40–J47 (62)
C76–C80 (24)
I10–I15 (150)
I30–I52 (70)
I (208)
ALL (288)
I60–I69 (21)
C (61)
I70–I79 (15)
E00–E07 (65)
C15–C26 (15)
C51–C58 (6)
C43–C44 (7)
−1
0
1
2
NMR Metabolomics
E50–E64 (7)
J80–J84 (8)
C15–C26 (18)
J95–J99 (14)
E10–E14 (39)
J90–J94 (15)
C81–C96 (15)
J09–J18 (33)
I26–I28 (13)
C60–C63 (9)
C30–C39 (5)
E (174)
J40–J47 (82)
E70–E90 (50)
J (171)
J20–J22 (37)
I10–I15 (204)
I70–I79 (17)
I20–I25 (86)
J60–J70 (11)
I (305)
C76–C80 (29)
C (86)
E00–E07 (94)
C64–C68 (7)
I05–I09 (12)
ALL (411)
I30–I52 (109)
I60–I69 (27)
C51–C58 (7)
I95–I99 (17)
C43–C44 (8)
−2
−1
0
1
2
3
UPLC IgG Glycomics
J90–J94 (18)
J80–J84 (11)
I26–I28 (18)
C64–C68 (10)
J60–J70 (12)
C15–C26 (21)
C76–C80 (38)
I95–I99 (23)
J09–J18 (40)
I05–I09 (15)
C43–C44 (9)
E10–E14 (57)
E70–E90 (67)
J95–J99 (14)
I30–I52 (128)
I60–I69 (38)
E50–E64 (8)
J20–J22 (43)
C (97)
C30–C39 (5)
C81–C96 (13)
J40–J47 (94)
ALL (484)
I20–I25 (107)
J (196)
E (217)
I (361)
I10–I15 (250)
I70–I79 (23)
C60–C63 (12)
E00–E07 (109)
C51–C58 (9)
−1
0
1
2


MS Fatty Acids Lipidomics
J60–J70 (8)
J80–J84 (4)
C64–C68 (9)
J95–J99 (12)
J90–J94 (10)
I95–I99 (13)
I60–I69 (20)
J09–J18 (21)
C60–C63 (6)
C30–C39 (2)
C15–C26 (15)
C (60)
I10–I15 (158)
C81–C96 (6)
I30–I52 (77)
E70–E90 (40)
I70–I79 (17)
I (220)
ALL (296)
E00–E07 (65)
C76–C80 (20)
J (116)
I26–I28 (9)
J40–J47 (60)
I20–I25 (63)
E (132)
C51–C58 (7)
J20–J22 (26)
C43–C44 (6)
E10–E14 (35)
I05–I09 (11)
E50–E64 (7)
−2.5
0.0
2.5
MS Metabolomics
I26–I28 (7)
I05–I09 (9)
J09–J18 (16)
E70–E90 (34)
J20–J22 (19)
I30–I52 (61)
I70–I79 (14)
I95–I99 (8)
J90–J94 (7)
I10–I15 (132)
J80–J84 (3)
J95–J99 (10)
C81–C96 (8)
I20–I25 (55)
J40–J47 (53)
I60–I69 (16)
I (187)
C51–C58 (6)
J (98)
ALL (262)
E (115)
C30–C39 (2)
E00–E07 (59)
C64–C68 (9)
C (55)
C76–C80 (19)
E10–E14 (30)
C43–C44 (5)
C15–C26 (13)
J60–J70 (5)
E50–E64 (4)
C60–C63 (6)
−2
−1
0
1
2
MS Complex Lipidomics
C30–C39 (3)
I26–I28 (8)
I95–I99 (11)
C81–C96 (7)
C64–C68 (9)
I05–I09 (12)
E70–E90 (39)
C (60)
J90–J94 (9)
C76–C80 (20)
J40–J47 (57)
J09–J18 (20)
I30–I52 (74)
I60–I69 (19)
C15–C26 (16)
J95–J99 (13)
J20–J22 (23)
E00–E07 (64)
ALL (291)
I (215)
E (127)
I10–I15 (156)
J (110)
I20–I25 (62)
I70–I79 (18)
C51–C58 (6)
J80–J84 (3)
E50–E64 (7)
C60–C63 (6)
E10–E14 (36)
J60–J70 (9)
C43–C44 (5)
−2
0
2
4



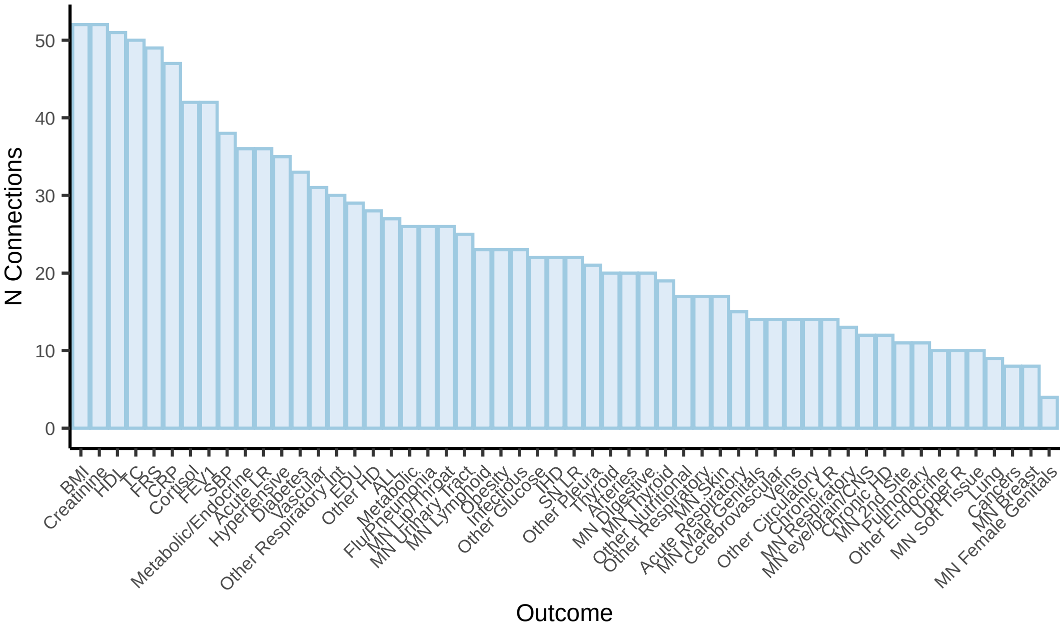
Supplementary Figure 41. Effect sizes using full Cox model and Martingale residuals are concordant for EHR modelling in the ORCADES dataset. *The disease block and the number of cases of that disease block in parenthesis are indicated on the y-axis. The effect size estimate and 95% confidence intervals for the effect of OCAA on disease block (log$_e$HR/standard deviation of OCAA) for each of the 11 OCAAs. Estimates from Cox models are in green and Martingale residuals in red.*



Supplementary Figure 42. Histogram of Significant Outcome Associations Across Single Omics Biomarkers. *Number of biomarkers (y axis) which had the number of significant outcome associations indicated on the x axis. Results shown for biomarker-outcome associations that passed 5% FDR significance.*

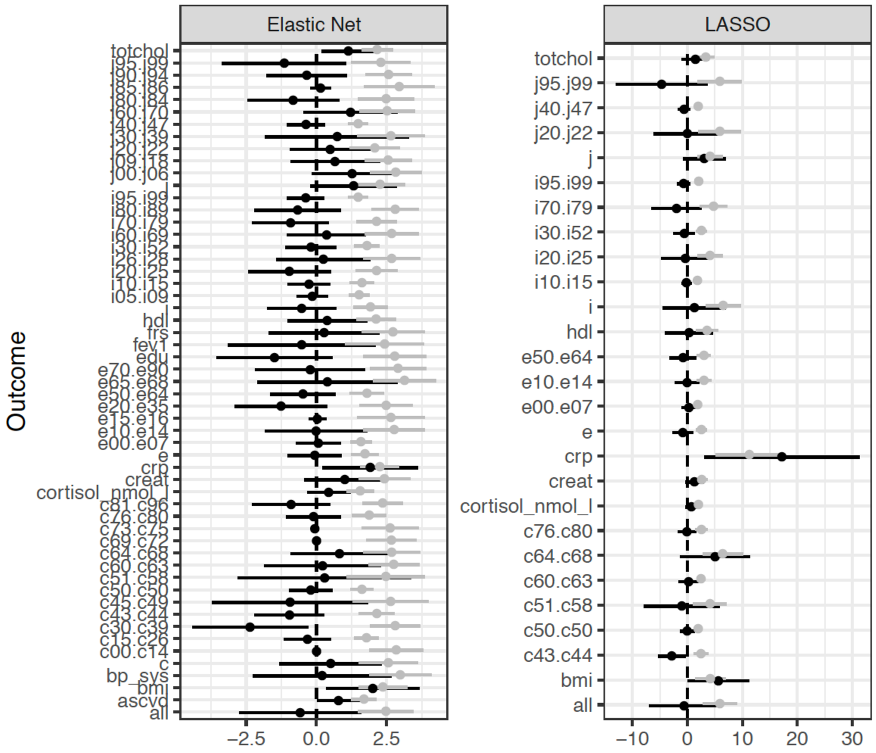
Trait	N Associations	N Unique Associations	N Shared Associations	Trait	N Associations	N Unique Associations	N Shared Associations
c60.c63	6	1	5	i60.i69	15	2	13
c64.c68	61	4	57	i70.i79	14	2	12
c69.c72	98	6	92	i80.i89	23	5	18
c73.c75	12	3	9	i95.i99	9	3	6
c76.c80	10	6	4	j	28	2	26
c81.c96	89	9	80	j00.j06	33	5	28
e	128	4	124	j09.j18	32	6	26
e00.e07	42	2	40	j20.j22	86	6	80
e10.e14	140	2	138	j30.j39	12	4	8
e15.e16	28	4	24	j40.j47	19	2	17
all	35	2	33	c00.c14	42	10	32
e20.e35	5	1	4	j60.j70	10	2	8
e50.e64	18	1	17	j80.j84	103	12	91
e65.e68	50	4	46	j85.j86	18	3	15
e70.e90	71	3	68	j90.j94	11	2	9
i	27	1	26	j95.j99	9	1	8
i05.i09	12	1	11	FRS	945	9	936
i10.i15	76	3	73	BMI	1231	47	1184
i20.i25	15	2	13	EDU	65	7	58
i26.i28	9	2	7	HDL	1256	28	1228
i30.i52	31	7	24	c15.c26	11	0	11
c	3	1	2	TC	1351	142	1209

Supplementary Table 24. Shared vs Unique Biomarker Associations Across Outcomes*. Trait: outcome. N Associations: number significant (5% FDR) biomarker-outcome associations. N Unique Associations: number of significantly associated biomarkers that are only associated with that trait. N Shared Associations: number of associated biomarkers that are associated with at least one other outcome.*

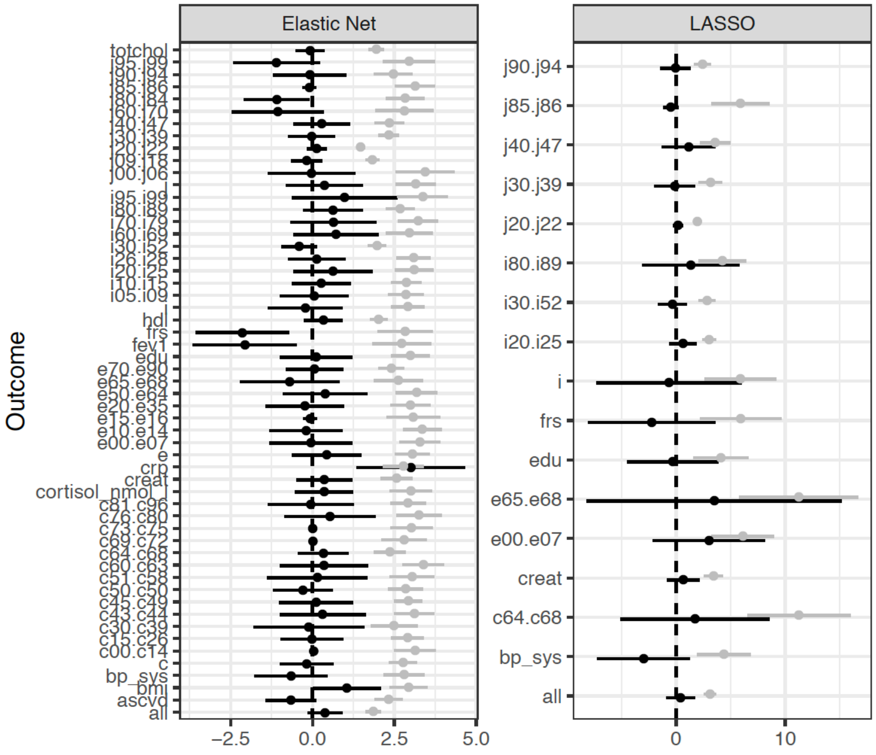


Supplementary Figure 43. Connectivity Across Outcomes. *Showing the number of other outcomes (y axis) that each outcome (x axis) shares significant biomarker associations with.*

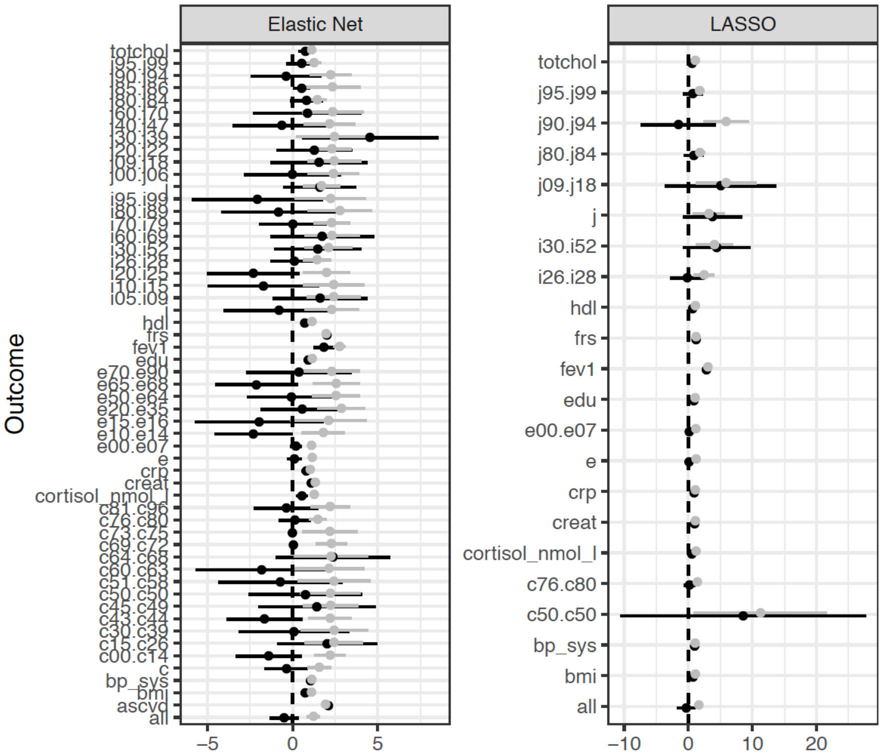
Hannum CpGs DNAme



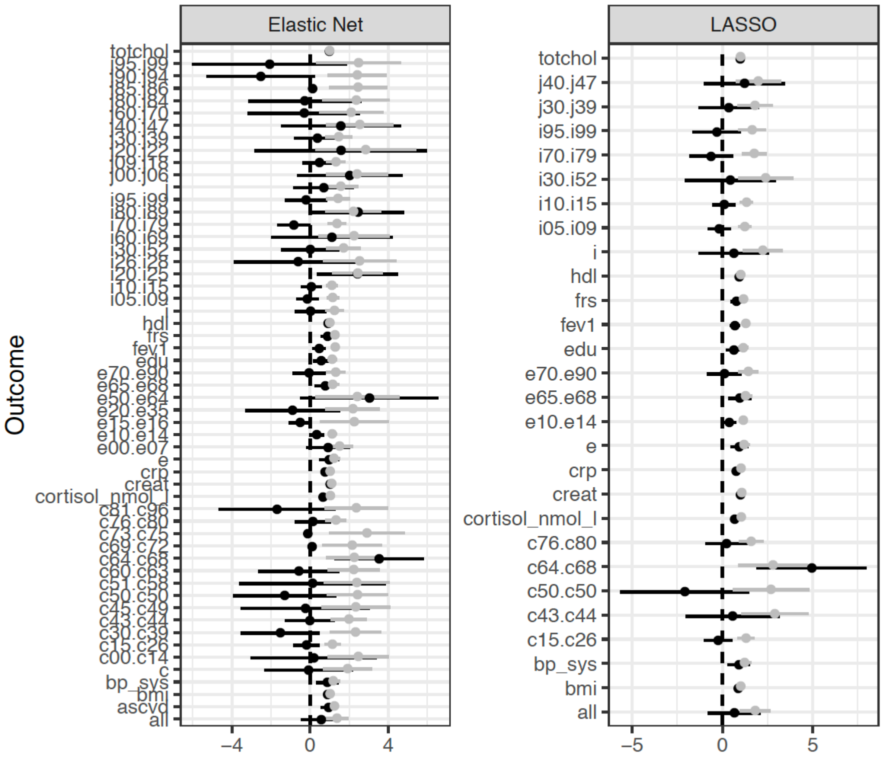
Horvath CpGs DNAme



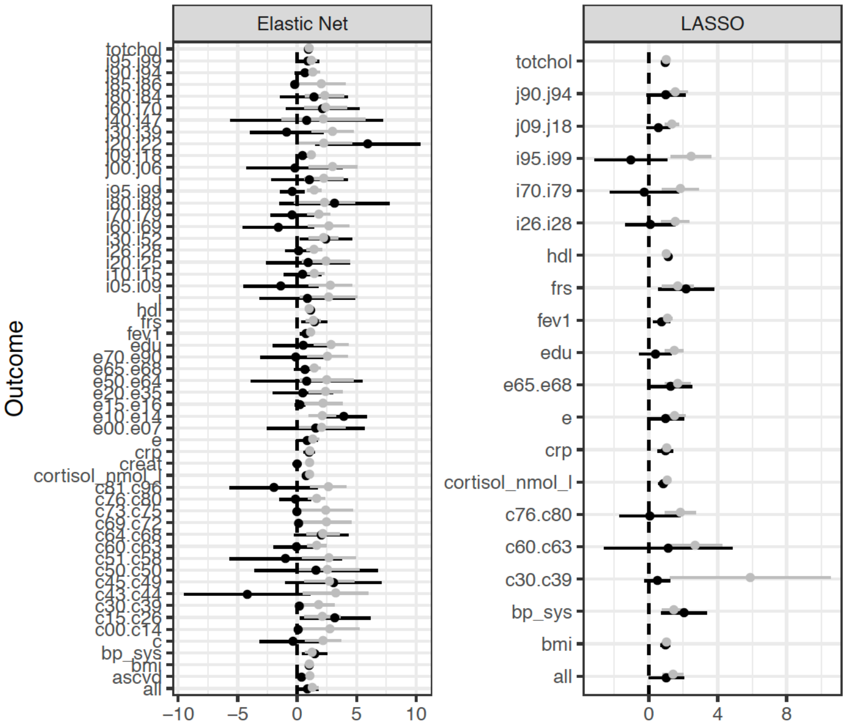
UPLC IgG Glycomics



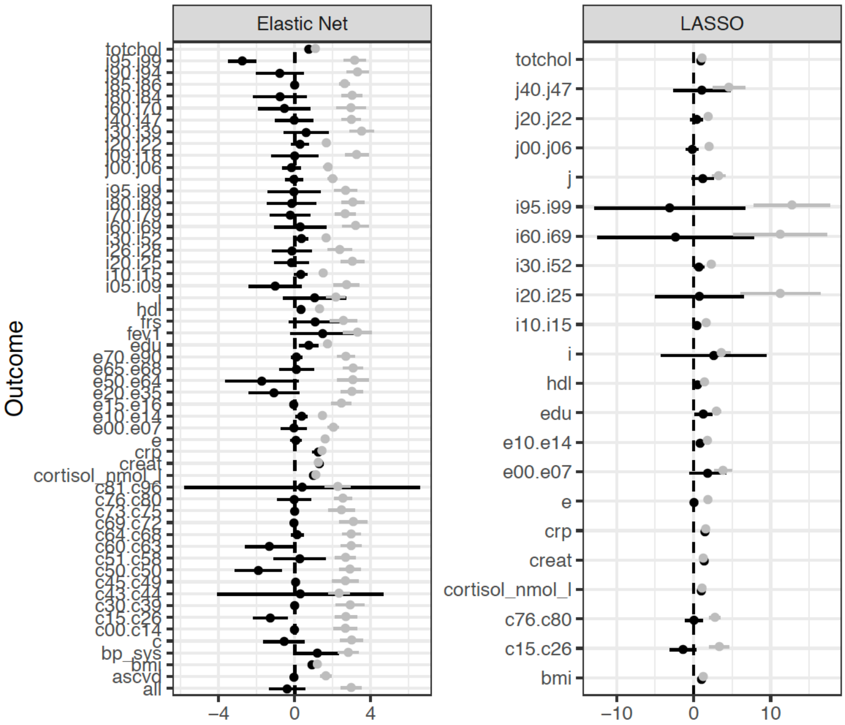
NMR Metabolomics



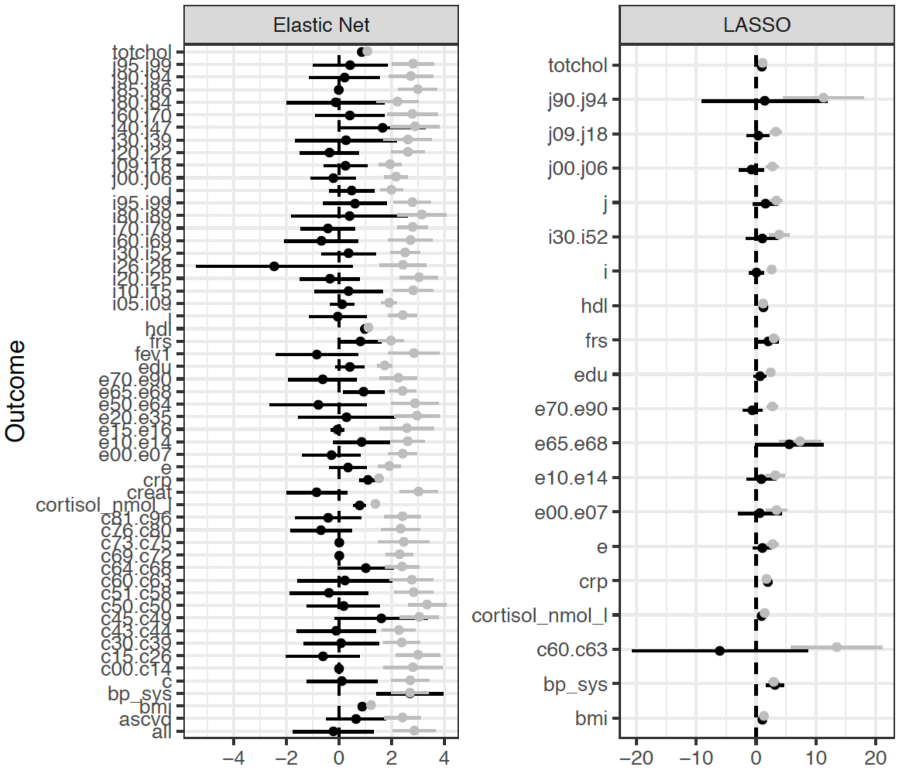
MS Fatty Acids Lipidomics



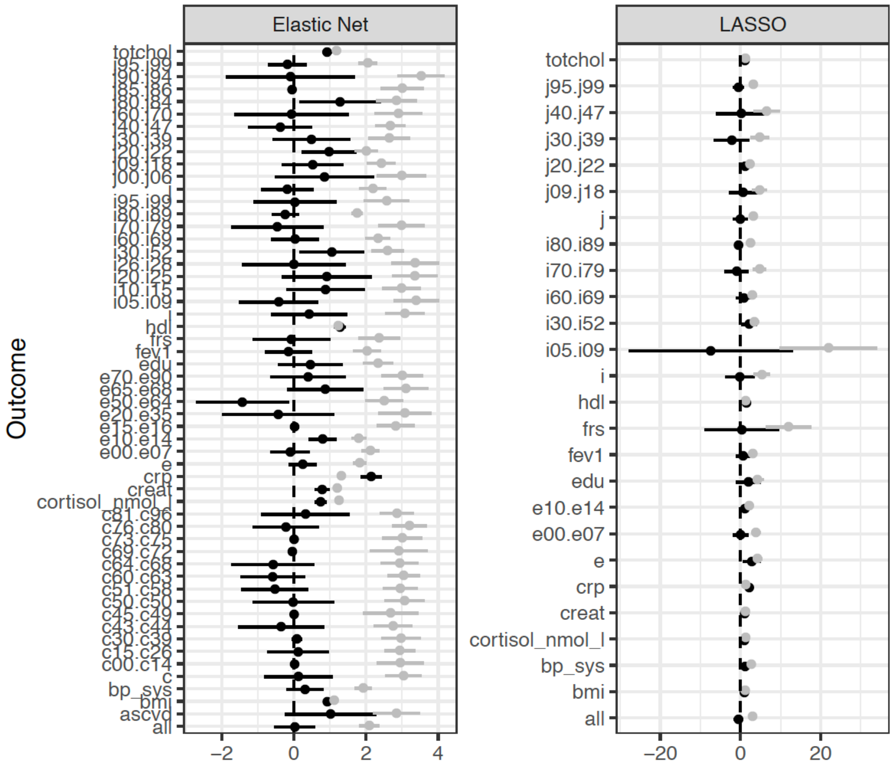
MS Metabolomics

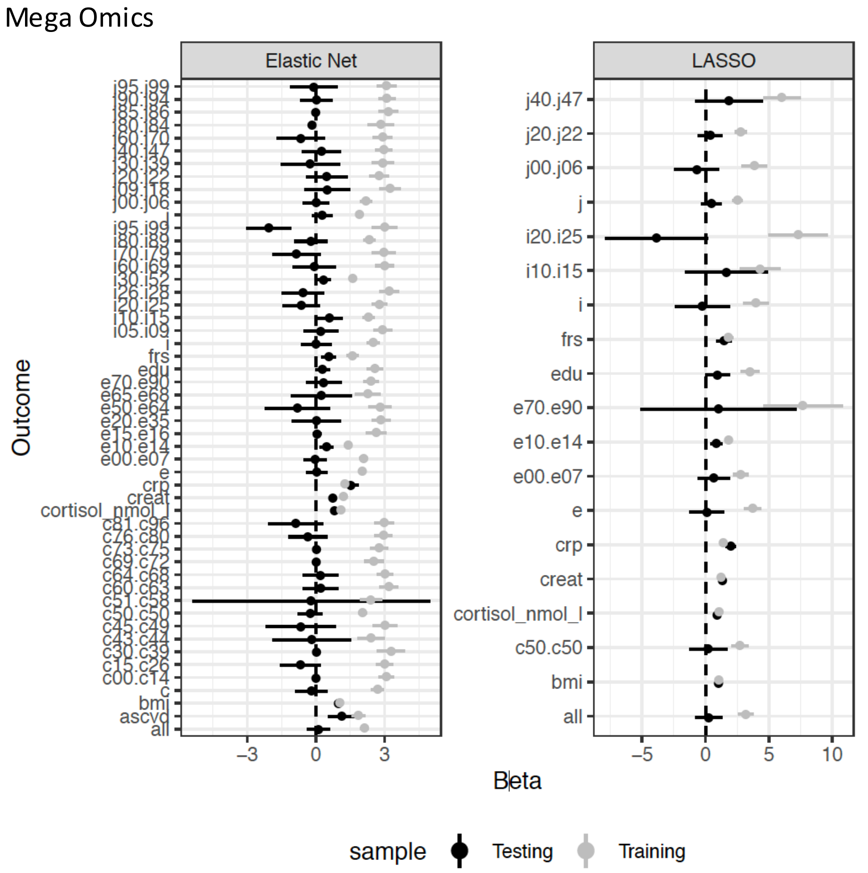


MS Complex Lipidomics

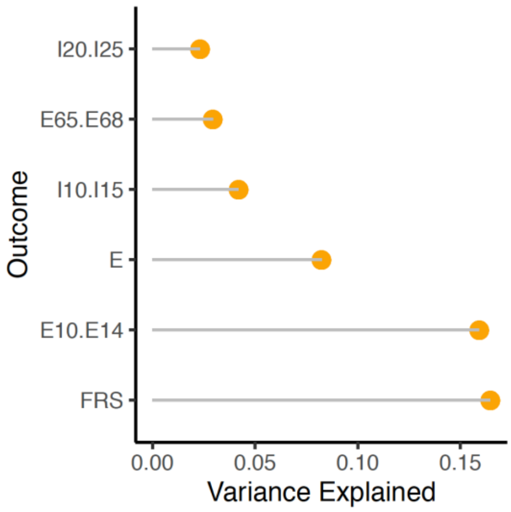


PEA Proteomics





Supplementary Figure 44. Effect size and 95% confidence intervals from regression of outcome predicted by the model and observed outcome in training and testing samples. *These estimates are across outcomes (y axis) and between methods (panels) across omics.*



Supplementary Figure 45. Variance Explained in Outcome by Clinomics Scores. *FRS: Framingham Risk Score. E10.E14: Diabetes mellitus. E: All block E metabolism related disorders. I10.I15: Hypertensive diseases. E65.E68: Obesity and other hyperalimentation. I20.I25: Ischaemic heart diseases.*

Chapter 5 Supplementary Information

https://doi.org/10.7488/ds/3107

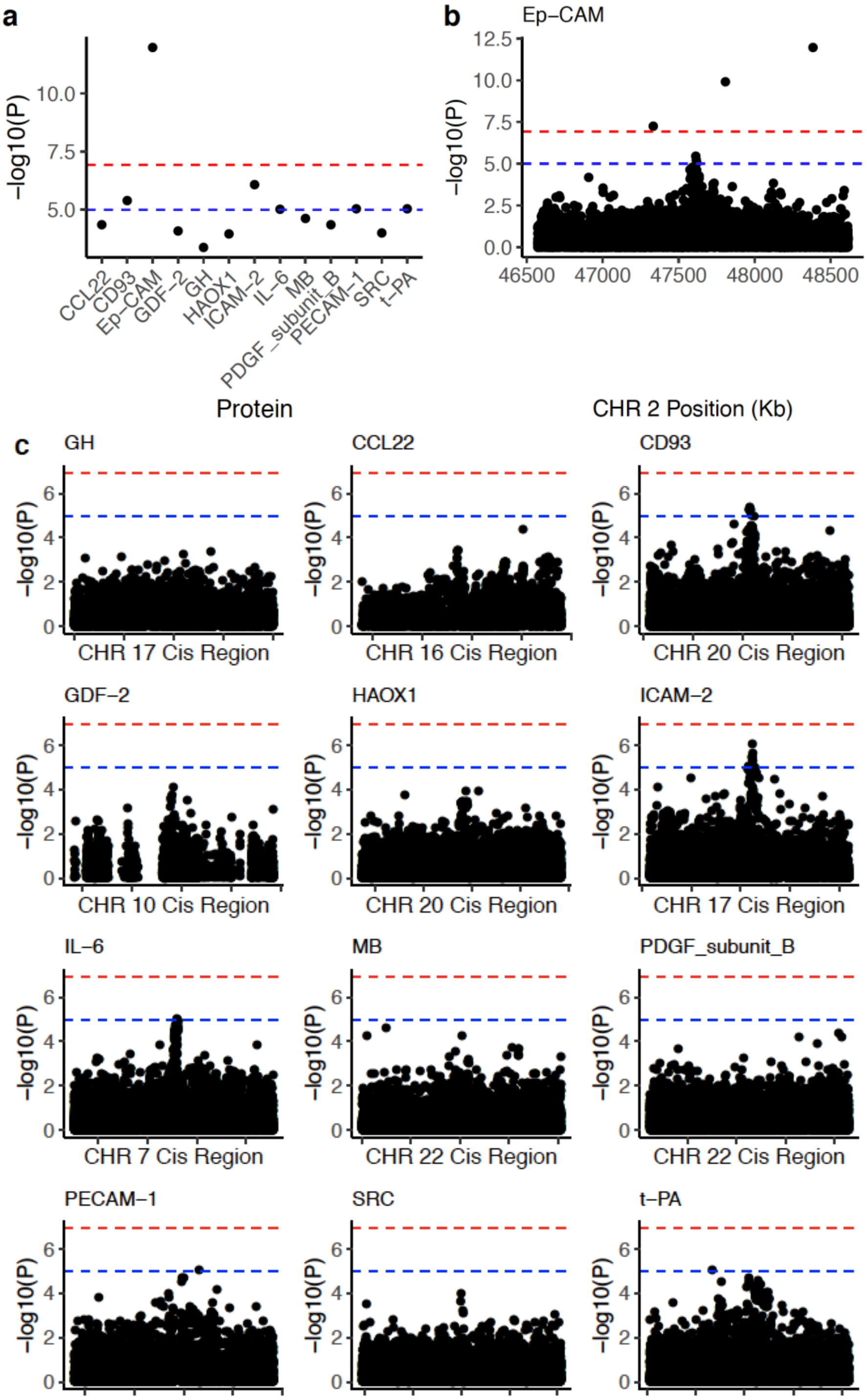
Supplementary Table 25. Contributing Cohorts. *For each cohort the full cohort name, study design, PMID, cohort description, ethics, matrix for proteomics, genotyping array, imputation panel, phasing software, imputation software, GWAS software and cohort specific acknowledgements.*

https://doi.org/10.7488/ds/3107

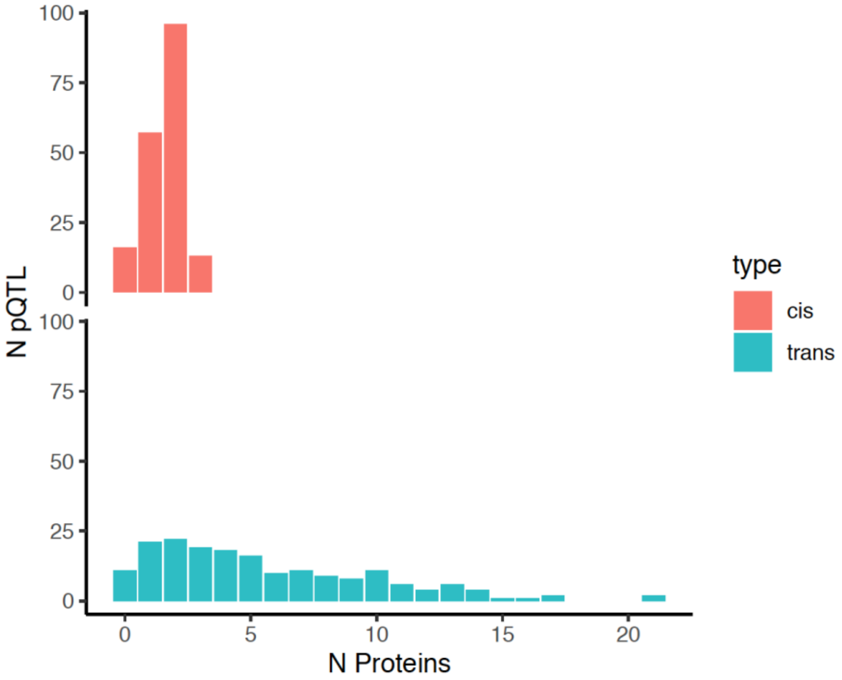
Supplementary Table 26. 1,308 Significant SNP-Protein Associations. *RSID: rsid, MARKERID: indicating chromosome position (b37) with the two alleles (alphabetical order), SNPID: chromosome_position (b37), EFFECT_ALLELE: the effect allele, REFERECE_ALLELE: other allele, FREQ1: frequency of the effect allele, FREQSE: standard error of the average frequency across cohorts, MIN_FREQ: minimum frequency of the effect allele across cohorts, MAX_FREQ: maximum frequency of the effect allele across cohorts, BETA: effect size of effect allele on protein, SE: standard error of the coefficient estimate, PVAL: p-value, DIRECTION: direction of cohort level effect, HETISQ: heterogeneity I^2 statistic, HETCHISQ: heterogeneity chi squared statistic, HETDF: heterogeneity degrees of freedom, HETPVAL: heterogeneity P-value, N: sample size, RQC_IMP: imputation quality score, CHR: chromosome, POS: position (b37), TRAIT: protein the variant is associated with, TYPE: cis- or trans-association, with cis defined as any variant within ± 1 Mb surrounding the coding region of the gene encoding the protein, p_s: whether the variant was discovered in the primary analysis or secondary (conditional analysis), pJ: the conditional joint p-value, NA if variant found in primary analysis, locus: indicating the independent locus the SNP-Protein association was assigned to, the name is the MARKERID of the top SNP of the assigned locus in* ***Supplementary Table 27****.*

https://doi.org/10.7488/ds/3107

Supplementary Table 27. 592 Significant Loci. The association result with the lowest P-value is indicated for each loci. *RSID: rsid, MARKERID: indicating chromosome position (b37) with the two alleles (alphabetical order), SNPID: chromosome_position (b37), EFFECT_ALLELE: the effect allele, REFERECE_ALLELE: other allele, FREQ1: frequency of the effect allele, FREQSE: standard error of the average frequency across cohorts, MIN_FREQ: minimum frequency of the effect allele across cohorts ,MAX_FREQ: maximum frequency of the effect allele across cohorts, BETA: effect size of effect allele on protein, SE: standard error of the coefficient estimate, PVAL: p-value, DIRECTION: direction of cohort level effect, HETISQ: heterogeneity I^2 statistic, HETCHISQ: heterogeneity chi squared statistic, HETDF: heterogeneity degrees of freedom, HETPVAL: heterogeneity P-value, N: sample size, RQC_IMP: imputation quality score, CHR: chromosome, POS: position (b37), TRAIT: protein the variant is associated with, TYPE: cis- or trans-association, with cis defined as any variant within ± 1 Mb surrounding the coding region of the gene encoding the protein, p_s: whether the variant was discovered in the primary analysis or secondary (conditional analysis), pJ: the conditional joint p-value, NA if variant found in primary analysis, locus: indicating the independent locus the SNP-Protein association was assigned to, the name is the MARKERID of the top SNP.*



Supplementary Figure 46. Underpowered to find cis-pQTL for 13 proteins. *a) Indicates the SNP with the lowest P-value in the cis region for the protein specified. b) Manhattan plot for the cis region for the protein Ep-CAM with points indicating the -log10(p-value). c) Manhattan plot for the cis regions of the 12 remaining proteins with no significant cis-signals. Genome-wide significance threshold for the cis region (1 x 10^{-5}) indicated in blue and Bonferroni significance threshold (1.18 x 10^{-7}) indicated in red.*



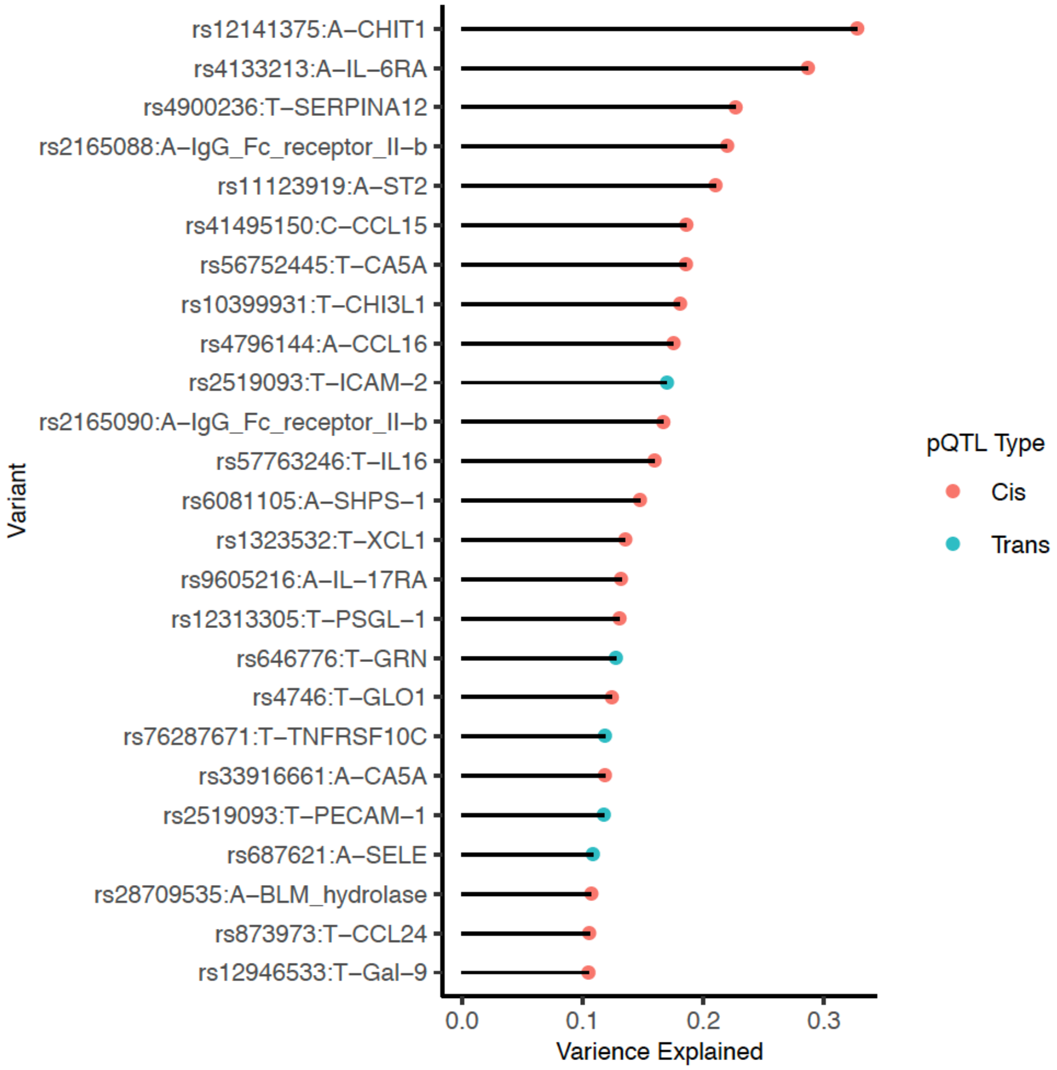
Supplementary Figure 47. Number of Cis & Trans pQTL per Protein. *Histogram of the number of Cis and trans pQTL across proteins.*

https://doi.org/10.7488/ds/3107

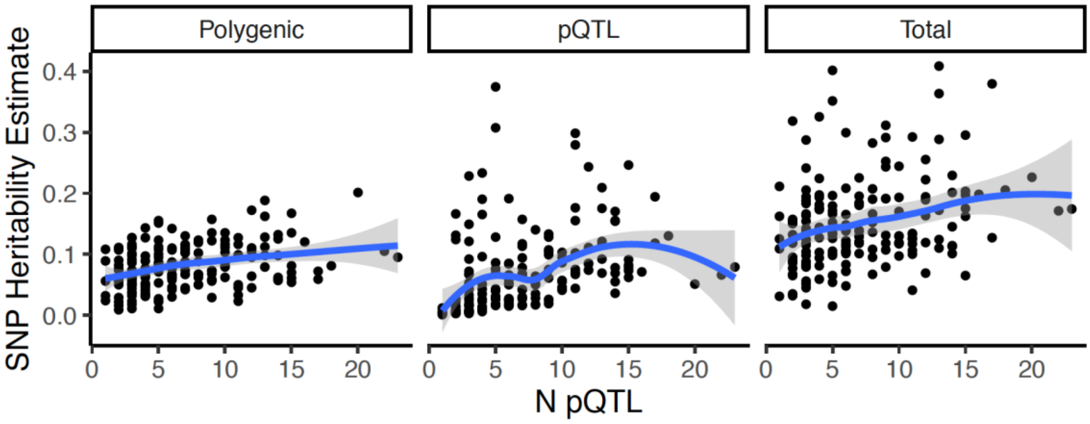
Supplementary Table 28. Pleiotropic Loci. *RSID: rsid of lead variant, Locus: chromosome_position build 37, indicating the 592 independent loci in* ***Supplementary Table 27****, count: Number significant SNP-Protein associations assigned to each independent locus.*

https://doi.org/10.7488/ds/3107

Supplementary Table 29. pQTL from 22 previous GWAS of Plasma Protein Levels. *hgnc_protein: human gene nomenclature committee (HGNC) protein names, snp: rsid, pub_p: P-value in publication, hgnc_gene: human gene nomenclature committee (HGNC) gene names, study: publication, n: sample size, chr: chromosome, pos38: position (b38).*



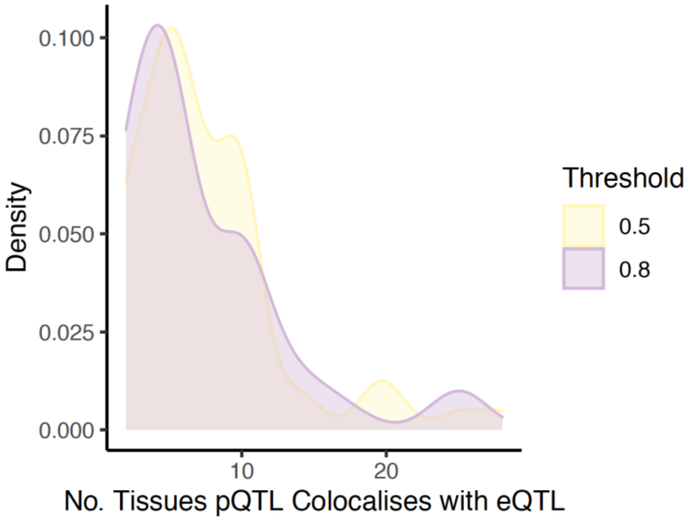
Supplementary Figure 48. Variance Explained by pQTL. *Variance explained in protein level on the x axis and the SNP (effect allele indicated) and the associated protein on the y axis. Points are coloured based on whether the variants are cis- or trans-associated with the plasma protein level. Only variants that have a variance explained of >10% are shown.*



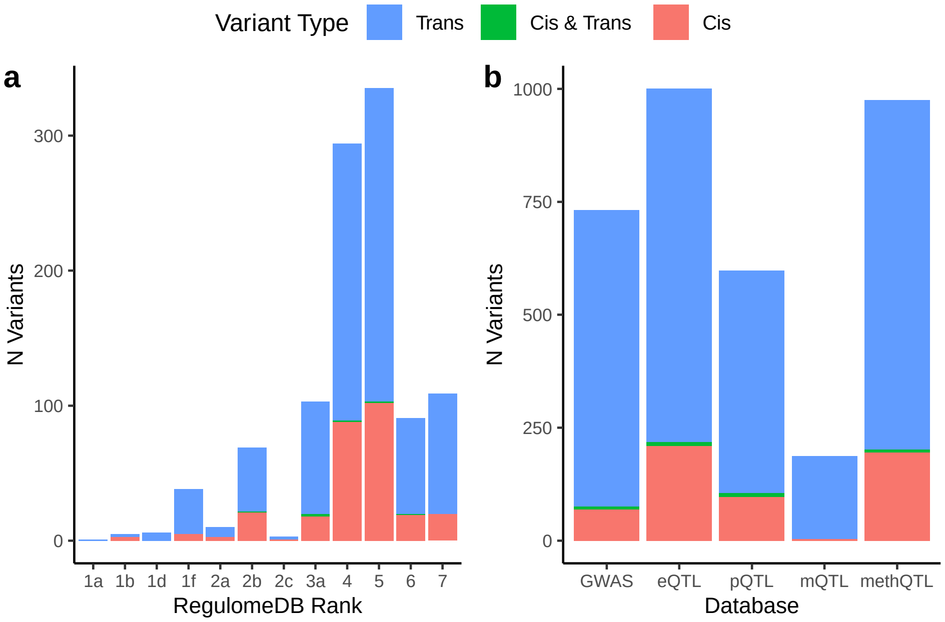
Supplementary Figure 49. Number of pQTL increases with increasing SNP heritability. *Estimates of heritability of the plasma protein levels on the y-axis with number of significant pQTL on the x-axis. Panels indicate the component of heritability. pQTL: the sum of the estimated variance explained in protein level from each of the lead variants. Polygenic: LDSC estimated SNP heritability excluding variants indexed by the lead variants. Total: the sum of the pQTL and Polygenic SNP heritability estimates.*

eQTL Dataset	N Associations	N Genes	N Proteins	N Gene encoding Protein	N Proteins other Genes (Not gene encoding protein)	N other Genes
Võsa (Cis)	431	265	102	5	101	261
Võsa (Trans)	6852	765	90	2	90	765
Cage	136	92	52	3	51	89
Westra	147	121	82	9	82	121
GTEx	965	642	156	72	150	590

Supplementary Table 30. SMR-HEIDI: Association of Gene Expression and Protein Levels. *eQTL Dataset: the eQTL dataset used for analysis. N Associations: the number of significant (within Dataset Bonferroni P_{SMR}, $P_{HEIDI} \geq 0.01$) probe-protein associations. N Genes: the number of genes that probes in significant associations map to. N Proteins: the number of plasma protein levels that had significant associations. N Gene encoding protein: The number of associations where the probe maps to the gene that encodes the protein it was found associated with. N Proteins other Genes (Not gene encoding protein): number of proteins that were associated with the expression levels of genes other than the coding gene. N other Genes: the number of genes other than the coding gene the protein was significantly associated with.*



Supplementary Figure 50. Density Plot of the number of different tissues pQTL colocalise with eQTL in. *Results for strong evidence of colocalisation PP>0.8 and likely to colocalise PP>0.5.*



Supplementary Figure 51. Lead Variant Annotation. *a) number of pQTL lead variants assigned each rank by RegulomeDB indicating evidence for being located in a regulatory/functional region. b) Number of lead variants that have previously been reported as significantly associated ($p<5 \times 10^{-8}$) with complex trait GWAS, eQTL: expression QTL, pQTL: protein QTL, mQTL: metabolite QTL, methQTL: CpG dinucleotide methylation QTL. Type of variant indicated by colour: Cis, Trans or both Cis & trans for different proteins.*

https://doi.org/10.7488/ds/3107

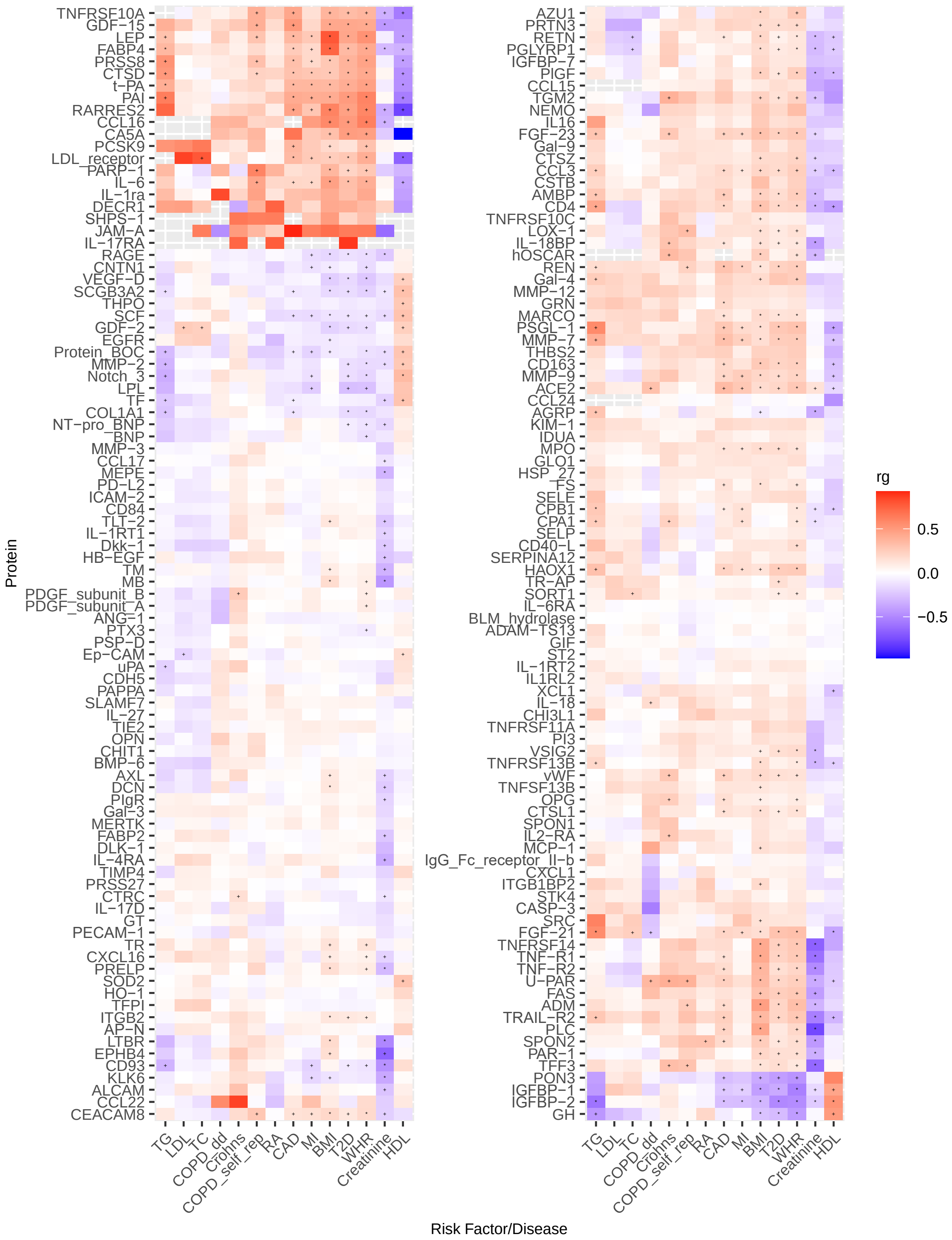
Supplementary Table 31. Trans gene-Protein Relationships. *TRAIT: the protein of interest, RSID: the rsid of the lead variants of the trans-pQTL in question, symbol: the symbols for the trans genes for that protein. Trans genes were defined as any gene whose coding region overlaps with a 1 MB window surrounding the lead variant of a trans-pQTL for that protein. known_int: indicating whether the trans gene had a known interaction with the protein of interest (1) or not (NA) in the STRING protein-protein interaction database (version 10). data_base: source database. KEGG: the KEGG pathway database, GO: gene ontology (GO) database of GO terms, PMID: pubmed. common_item: the commonality between trans gene and protein of interest. For queries against the KEGG database this is the pathway name_KEGG pathway ID, for queries against the GO database this is the GO term that the trans gene had in common with the protein of interest, for queries in pubmed, this is the pubmed ID for publications that mention both the trans gene and the protein of interest.*

https://doi.org/10.7488/ds/3107

Supplementary Table 32. Phenoscanner Results.

https://doi.org/10.7488/ds/3107

Supplementary Table 33. Publicly Available Summary Statistics used for Genetic Correlations. *Phenotype: trait, Name of Phenotype: name of phenotype as appears in figures, Source: Consortium/cohort of summary statistics, Author: first author of publication or Neale indicating http://www.nealelab.is/uk-biobank. PMID: PubMed ID of publication, File: name of file downloaded from source, URL: download link.*



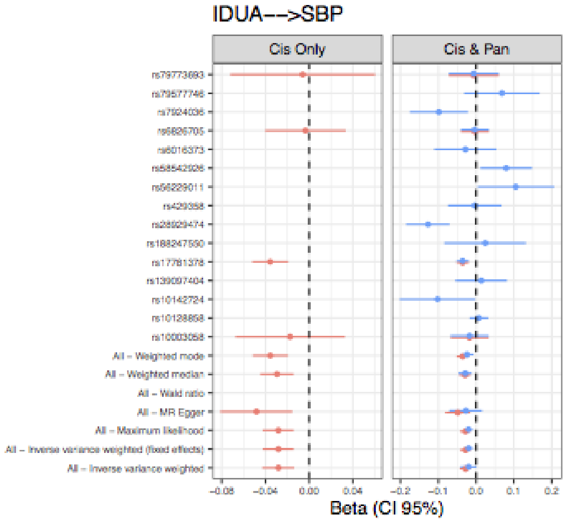
Supplementary Figure 52. Genetic correlations (r_g) of plasma protein levels and complex traits. *Estimates of genetic correlations of protein levels and complex traits and health risk factors calculated using high definition likelihood. The shade indicates the magnitude of the r_g estimate with the colour denoting the direction of the correlation. 5% FDR correlations and those statistically significant after Bonferroni correction for multiple testing are indicated with (+) and (*) respectively. Proteins and complex traits are ordered based on hierarchical clustering of the correlation coefficient (r_g).*

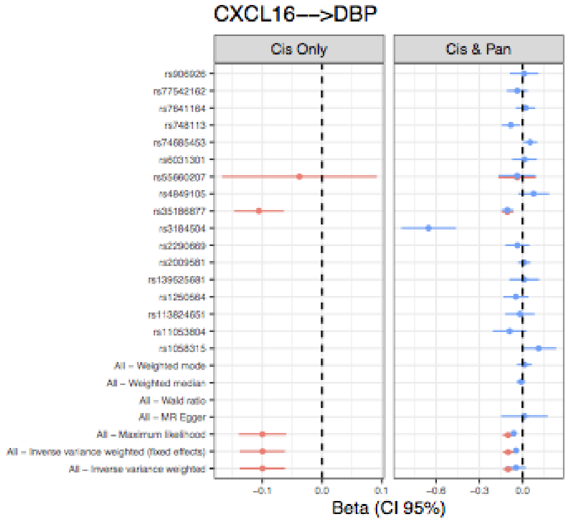
https://doi.org/10.7488/ds/3107

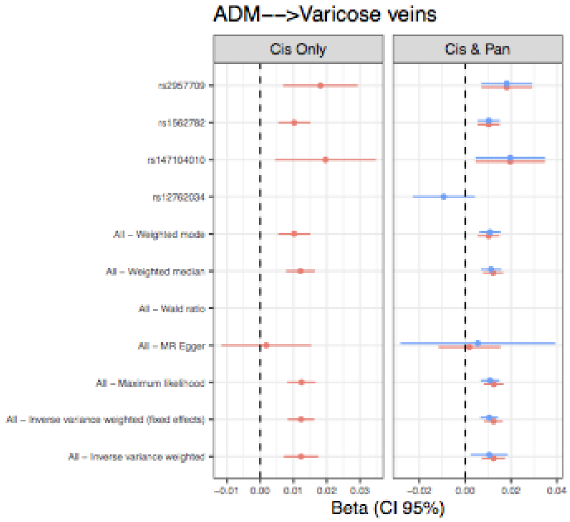
Supplementary Table 34. MR Outcomes. *Indicating the MRBase & TwoSampleMR outcome id and full outcome name. Full Name: descriptive name of outcome. id: MRBase ID. year: year of publication. author: author of publication. consortium: consortium if study used a consortium. sex: sex of participants used in outcome study. pmid: pubmed ID for publication. population: ancestry of outcome study population. nsnp: number of SNPs in outcome study. sample_size: total samples size for outcome study. build: genome build used in outcome study. ncase: number of cases if case control study. ncontrol: number of controls if case control study.*

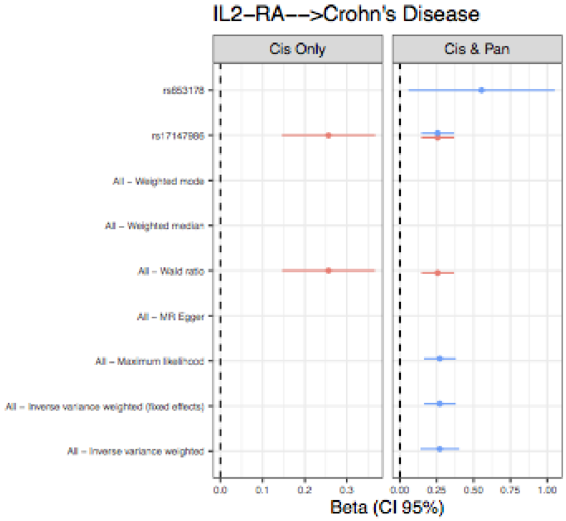
https://doi.org/10.7488/ds/3107

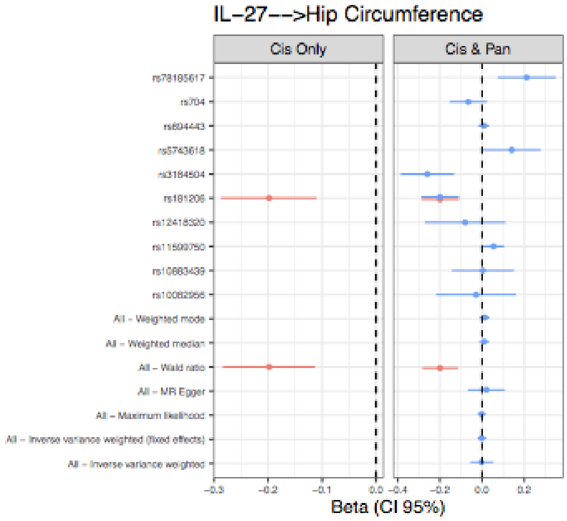
Supplementary Table 35. Significant Mendelian Randomisation Results. *exposure: exposure protein, outcome: full name of outcome (MRBase), method: indicating the MR method used, nsnp: the number of variants used as instruments in the MR analysis, beta: inferred causal effect estimate of protein level on outcome, se: standard error of effect size estimate, p: p-value of effect of protein level on outcome, correct_causal_direction: indicating the results of the Steiger test for directionality of the causal relationship, all were TRUE in this analysis indicating that the direction of effect is exposure to outcome, steiger_pval: the p-value for the Steiger test for directionality of causality, Q: Cochran's Q value for heterogeneity between instruments (NA for analyses with single SNPs as IVs), Q_df: the degrees of freedom for Cochran's Q statistic, Q_pval: p-value for Cochran's Q, with significant values indicating significant instrument heterogeneity therefore values of Q_pval>0.05 indicate no significant heterogeneity, egger_intercept: the intercept from MR-Egger, its divergence from zero indicating horizontal pleiotropy, pval: the p-value for the egger_intercept, with significant (p<0.05) indicating evidence of horizontal pleiotropy, bi_dir: indicating whether there was significant evidence of a causal effect of outcome on exposure in bidirectional MR was performed (all 20 significant protein-outcomes relationships showed no significant evidence of causal effect of outcome on exposure therefore passed this sensitivity analysis), outcome_id, q: the adjusted MR p-value using Benjamini-Hochberg method, coloc: the posterior probability of hypothesis 4 - that protein and outcome share a causal variant - results from colocalisation of pQTL used as instruments in the MR analysis and the outcome GWAS.*

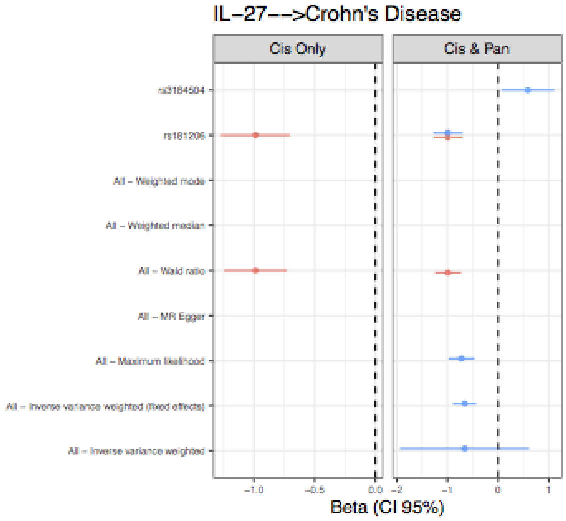

IDUA-->SBP
Cis Only
Cis & Pan
rs79773693
rs79577746
rs7924036
rs6826705
rs6016373
rs58542926
rs56229011
rs429358
rs28929474
rs188247550
rs17781378
rs139097404
rs10142724
rs10128858
rs10003058
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−0.08
−0.04
0.00
0.04
−0.2
−0.1
0.0
0.1
0.2
Beta (CI 95%)

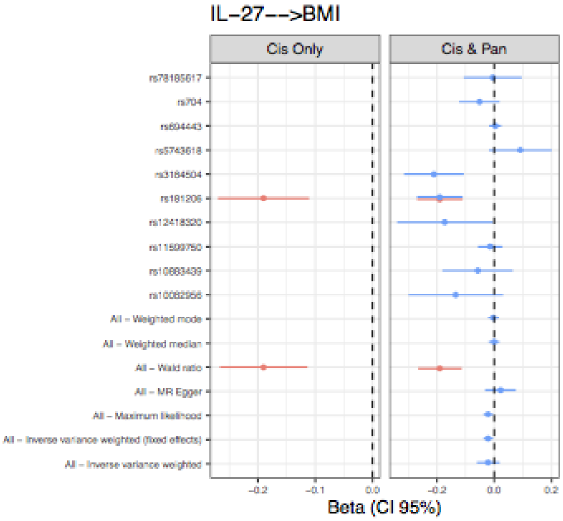

CXCL16-->DBP
Cis Only
Cis & Pan
rs906926
rs77542162
rs7841184
rs748113
rs74685453
rs6031301
rs55660207
rs4849105
rs35186877
rs3184504
rs2290669
rs2009581
rs139525681
rs1250584
rs113824651
rs11053804
rs1058315
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−0.1
0.0
0.1
−0.6
−0.3
0.0
Beta (CI 95%)

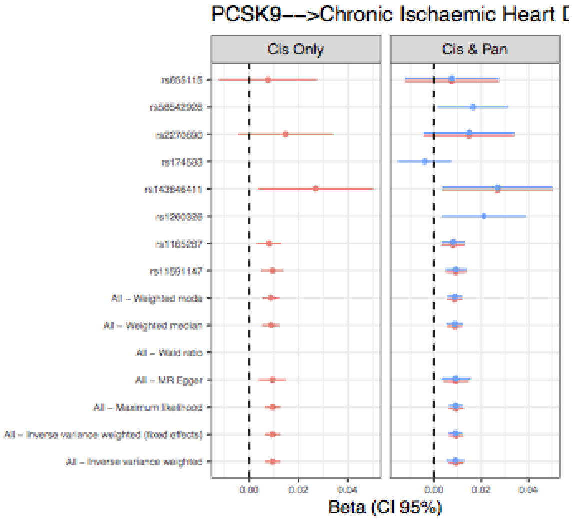

ADM-->Varicose veins
Cis Only
Cis & Pan
rs2957709
rs1562782
rs147104010
rs12762034
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−0.01
0.00
0.01
0.02
0.03
−0.02
0.00
0.02
0.04
Beta (CI 95%)

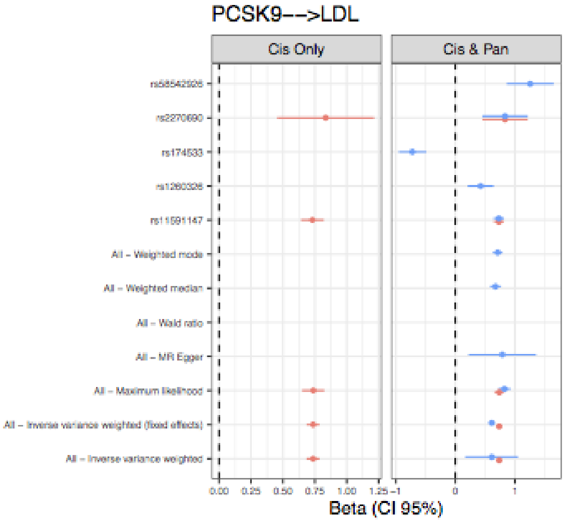

IL2-RA-->Crohn's Disease
Cis Only
Cis & Pan
rs653178
rs17147686
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
0.0
0.1
0.2
0.3
0.00
0.25
0.50
0.75
1.00
Beta (CI 95%)

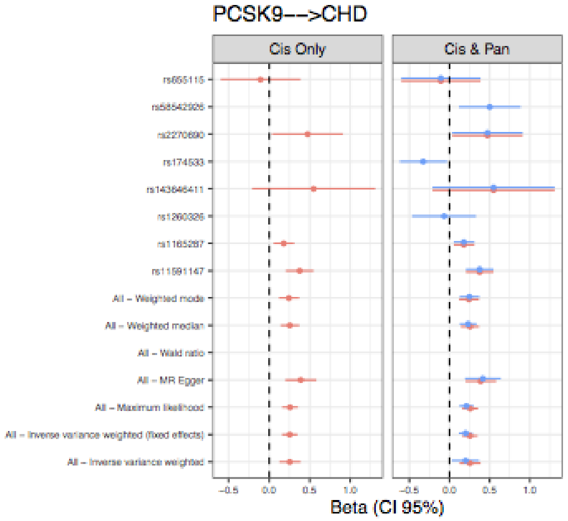

IL-27-->Hip Circumference
Cis Only
Cis & Pan
rs78185617
rs704
rs694443
rs5743618
rs3184504
rs181206
rs12418320
rs11599750
rs10883439
rs10082956
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−0.3
−0.2
−0.1
0.0
−0.4
−0.2
0.0
0.2
Beta (CI 95%)

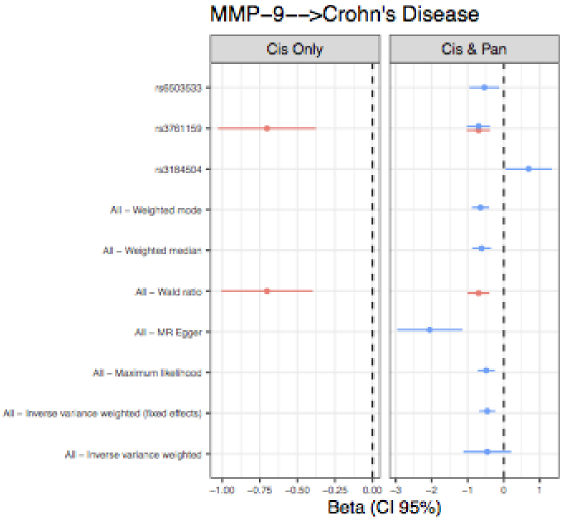

IL-27-->Crohn's Disease
Cis Only
Cis & Pan
rs3184504
rs181206
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−1.0
−0.5
0.0
−2
−1
0
1
Beta (CI 95%)

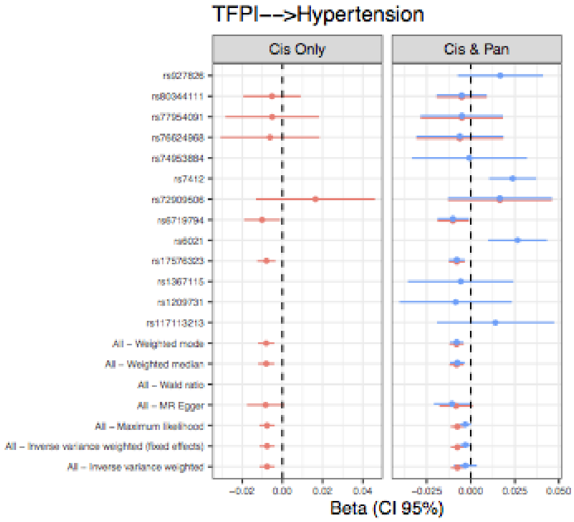

IL–27-->BMI
Cis Only
Cis & Pan
rs78185617
rs704
rs894443
rs5743618
rs3184504
rs181206
rs12418320
rs11599750
rs10883439
rs10082956
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
-0.2
-0.1
0.0
-0.2
0.0
0.2
Beta (CI 95%)

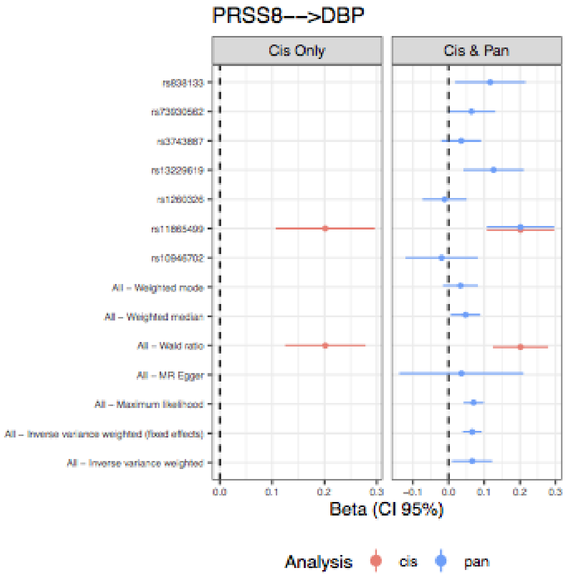

PCSK9-->Chronic Ischaemic Heart D
Cis Only
Cis & Pan
rs655115
rs58542926
rs2270690
rs174533
rs143646411
rs1260326
rs1165287
rs11591147
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
0.00
0.02
0.04
0.00
0.02
0.04
Beta (CI 95%)

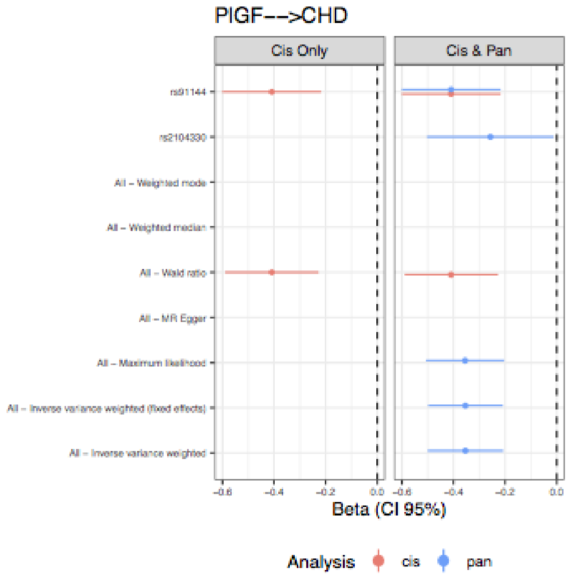

PCSK9-->LDL
Cis Only
Cis & Pan
rs58542926
rs2270690
rs174533
rs1260326
rs11591147
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
0.00
0.25
0.50
0.75
1.00
1.25
-1
0
1
Beta (CI 95%)

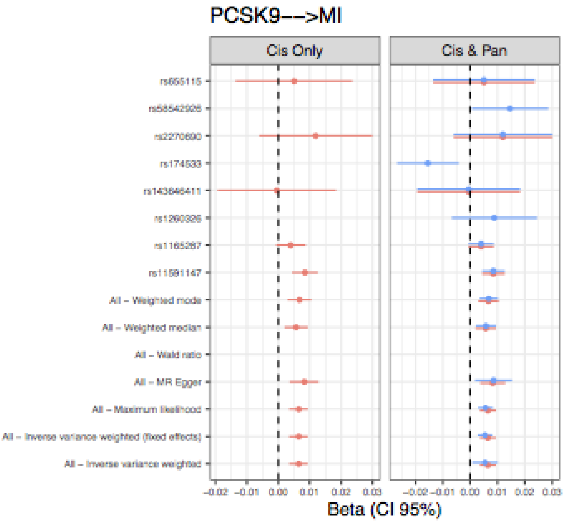

PCSK9-->CHD
Cis Only
Cis & Pan
rs655115
rs58542926
rs2270690
rs174533
rs143646411
rs1260326
rs1165287
rs11591147
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
-0.5
0.0
0.5
1.0
-0.5
0.0
0.5
1.0
Beta (CI 95%)

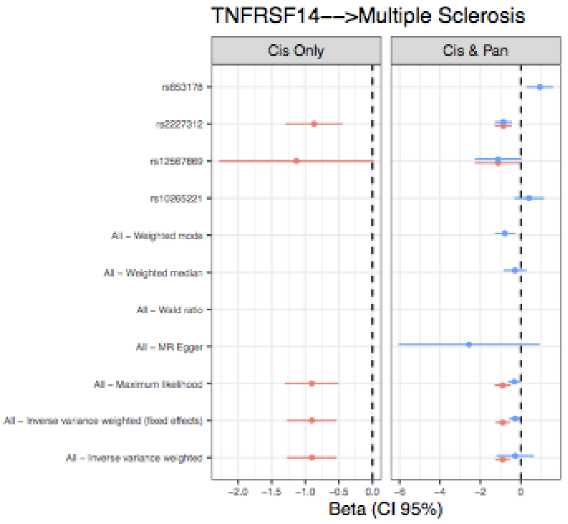

MMP–9-->Crohn's Disease
Cis Only
Cis & Pan
rs6503533
rs3761159
rs3184504
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
-1.00
-0.75
-0.50
-0.25
0.00
-3
-2
-1
0
1
Beta (CI 95%)

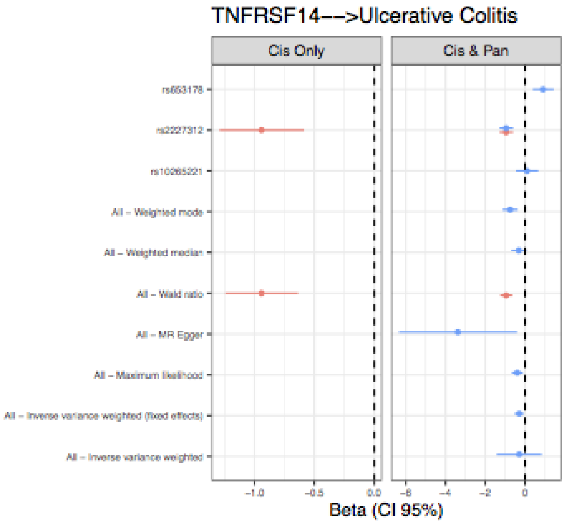

TFPI-->Hypertension
Cis Only
Cis & Pan
rs927826
rs80344111
rs77954091
rs76624968
rs74953884
rs7412
rs72909506
rs6719794
rs6021
rs17576323
rs1367115
rs1209731
rs117113213
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
-0.02
0.00
0.02
0.04
-0.025
0.000
0.025
0.050
Beta (CI 95%)

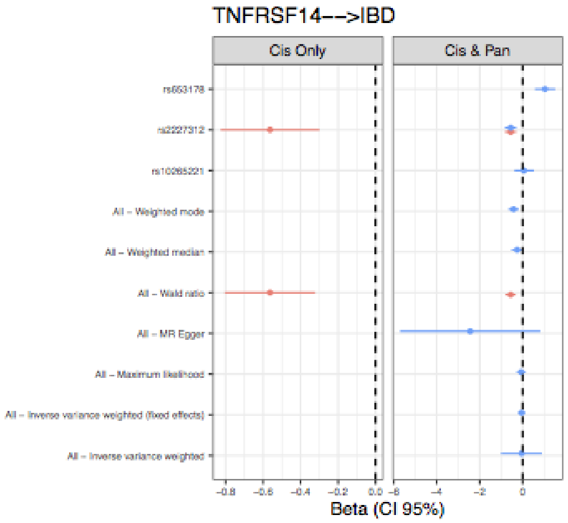

PRSS8-->DBP
Cis Only
Cis & Pan
rs838133
rs73930582
rs3743887
rs13229619
rs1260326
rs11865499
rs10946702
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
0.0
0.1
0.2
0.3
−0.1
0.0
0.1
0.2
0.3
Beta (CI 95%)
Analysis
cis
pan

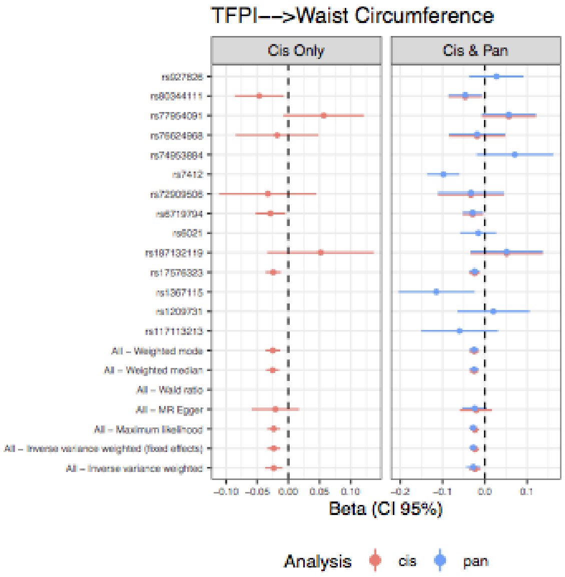

PIGF-->CHD
Cis Only
Cis & Pan
rs91144
rs2104330
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−0.6
−0.4
−0.2
0.0
−0.6
−0.4
−0.2
0.0
Beta (CI 95%)
Analysis
cis
pan


PCSK9-->MI
Cis Only
Cis & Pan
rs655115
rs58542926
rs2270690
rs174533
rs143846411
rs1260326
rs1165287
rs11591147
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−0.02
−0.01
0.00
0.01
0.02
0.03
−0.02
−0.01
0.00
0.01
0.02
0.03
Beta (CI 95%)


TNFRSF14-->Multiple Sclerosis
Cis Only
Cis & Pan
rs653178
rs2227312
rs12567869
rs10265221
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−2.0
−1.5
−1.0
−0.5
0.0
−6
−4
−2
0
Beta (CI 95%)


TNFRSF14-->Ulcerative Colitis
Cis Only
Cis & Pan
rs653178
rs2227312
rs10265221
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−1.0
−0.5
0.0
−6
−4
−2
0
Beta (CI 95%)


TNFRSF14-->IBD
Cis Only
Cis & Pan
rs653178
rs2227312
rs10265221
All – Weighted mode
All – Weighted median
All – Wald ratio
All – MR Egger
All – Maximum likelihood
All – Inverse variance weighted (fixed effects)
All – Inverse variance weighted
−0.8
−0.6
−0.4
−0.2
0.0
−6
−4
−2
0
Beta (CI 95%)



Supplementary Figure 53. Additional Sensitivity Analysis for Significant Proteins. *Panels for each significant MR estimate the effect size and 95% confidence intervals (x axis) for the MR method or individual IV (y axis). Left hand panels show results of cis IVs only analysis and right-hand panels show comparison of cis only and cis and trans combined termed pan.*

Protein	MR Outcome	Drugs	Drug Status	Drug Use	Protein Location
PlGF	CHD	Aflibercept	Approved	Branch Retinal Vein Occlusion With Macular Edema |Central Retinal Vein Occlusion With Macular Edema |Diabetic Macular Edema (DME) |Diabetic Retinopathy (DR) | Macular Edema |Metastatic Colorectal Cancer (MCRC) |Myopic Choroidal Neovascularization |Neovascular Age-Related Macular Degeneration |Wet Age-Related Macular Degeneration	Secreted to blood
PRSS8	DBP	1-[4-(hydroxymethyl)phenyl]guanidine	Experimental	-	Membrane/Secreted
IL2-RA	Crohns	Denileukin diftitox Basiliximab Aldesleukin	Approved Approved Approved	Cutaneous T-cell Lymphoma (CTCL) Kidney Transplant Rejection Renal Cell Carcinoma	Intracellular/Membrane

MMP-9	Crohns	Minocycline	Approved	Bartonellosis, Brucellosis | Campylobacter fetus |Chancroid | Cholera | Conjunctivitis, Inclusion |Granuloma Inguinale | Lymphogranuloma Venereum |Nongonococcal urethritis |Periodontitis | Plague |Psittacosis | Q Fever | Relapsing Fever |Respiratory Tract Infections (RTI) |Rickettsia Infections | Rickettsialpox |Rocky Mountain Spotted Fever | Trachoma |Tularemia | Typhus Fever | Inflammatory lesions	Secreted to Blood
		Captopril	Approved	Aldosteronism | Anatomic renal artery stenosis |Congestive Heart Failure (CHF) | Diabetic Nephropathy | Heart Failure | High Blood Pressure (Hypertension) |Hypertensive crisis |Non ST Segment Elevation Acute Coronary Syndrome |Raynaud's Phenomenon |Ejection fraction of 40% or less Left ventricular dysfunction	

PCSK9	CHD Chronic ischaemic heart disease MI LDL	Alirocumab	Approved	Heterozygous Familial Hypercholesterolemia | Myocardial Infarction | Stroke | Unstable Angina Pectoris | Primary Hyperlipidemia	Secreted to Blood
		Evolocumab	Approved	Atherosclerotic Cardiovascular Diseases | Heterozygous Familial Hypercholesterolemia | Homozygous Familial Hypercholesterolemia	
		Inclisiran	Approved	Mixed Dyslipidemias | Primary Hypercholesterolemia	
TFPI	Waist Circ Hypertension	Coagulation factor VIIa	Approved	Bleeding | Severe Bleeding	Secreted to Blood
		Recombinant Human Dalteparin	Approved	Cardiovascular Events | Clotting | Deep Vein Thrombosis | Symptomatic Venous Thromboembolism | Venous Thromboembolism	
		Andexanet alfa	Approved	Severe Life-threatening, uncontrollable Bleeding	
ADM	Varicose veins	-	-	-	Intracellular/Secreted to blood
CXCL16	DBP	-	-	-	Intracellular/Membrane/ Secreted to blood
IDUA	SBP	-	-	-	Intracellular

IL-27	BMI Hip Circ Crohn’s	-	-	-	Intracellular/Secreted to blood
TNFRSF14	IBD MS Ulcerative colitis	-	-	-	Intracellular/Membrane

Supplementary Table 36. Drug Target Status of Proteins with significant MR estimates. *For each of the 11 proteins highlighted in our MR analysis: MR outcome is associated with. Drugs: the drugs the protein is a current target for. Drug Status: the status of associated drugs. Drug Use: phenotypes associated drugs are used to combat. Protein Location: location of the expressed protein in vivo (Human Protein Atlas[337] available from http://www.proteinatlas.org).*

https://doi.org/10.7488/ds/3107

Supplementary Table 37. Cohort Sample Sizes.