

Integrative analysis of common and rare variants for a more comprehensive genetic risk assessment

Doctoral thesis
to obtain a doctorate (PhD)
from the Faculty of Medicine
of the University of Bonn

Rana Aldisi

from Doha, Qatar

2024

Written with authorization of
the Faculty of Medicine of the University of Bonn

First reviewer: Prof. Dr. med. Peter Krawitz

Second reviewer: Prof. Dr. Holger Fröhlich

Day of oral examination: 07.11.2024

From the Institute for Genomic Statistics and Bioinformatics

Director: Prof. Dr. med. Peter Krawitz

Dedication

I dedicate this thesis to myself, a reflection of my only slightly wavering belief in my abilities and the sheer stubbornness that carried me through this academic journey. Through many late nights, countless coffee cups, and more 'Eureka!' moments than I can count, I've made it here.

As I stand at this milestone, I am reminded that self-belief is the cornerstone of achievement, and this dedication is a celebration of my personal growth and determination. May it inspire others to trust in their own capabilities and persevere through their academic pursuits, just as I have on this remarkable journey.

Table of Contents

List of Abbreviations	6
1 Abstract	7
2 Introduction with aims and references	8
2.1 Genetics of complex phenotypes.....	8
2.2 The role of environment in complex traits.....	10
2.3 Genetic Risk Analysis and Assessment	10
2.4 Aims of the study	11
3 Publications	14
3.1 Publication 1 - GenRisk: a tool for comprehensive genetic risk modeling	14
3.1.1 Publication 1 - Appendix A	18
3.2 Publication 2 - Gene-based burden scores identify rare variant associa- tions for 28 blood biomarkers.....	55
3.2.1 Publication 2 - Appendix A	67
3.3 Publication 3 - Analysis of 72,469 UK Biobank exomes links rare variants to male-pattern hair loss	124
3.3.1 Publication 3 - Appendix A	138
3.3.2 Publication 3 - Appendix B	152
4 Discussion with references	153
4.1 Limitations and future outlook	155
5 Acknowledgment	158

List of abbreviations

GBS gene-based scores

GWAS Genome-Wide Association Studies

MAF minor allele frequency

ML machine learning

MPHL male-pattern hair loss

PRS Polygenic Risk Scores

QTL Quantitative Trait Loci

SKAT Sequence Kernel Association Test

1. Abstract

The etiology of complex traits is difficult to interpret because of their multifactorial nature. And while environmental factors play an important role in their development, genetic factors also have huge and crucial effect on the expression of complex phenotypes. However, a relevant part of the genetic landscape is yet to be discovered, despite the century long research and studies on the topic. The aim of this thesis is to investigate the role of rare pathogenic variants in complex traits and integrate their analysis with common variants for a more comprehensive genetic risk assessment.

In the first paper, we introduce an open source python package, GenRisk, that implements gene-based burden scores, focusing on rare deleterious variants, and polygenic risk scores (PRS), which are based on common variants. GenRisk's pipeline also contains an association analysis function using different regression models and a genetic risk modeling function utilizing multiple machine learning models. In this paper, we applied the pipeline on samples from UK Biobank as a usage case example.

The second paper employs the GenRisk framework to explore 28 blood biomarkers within the UK Biobank cohort. We performed the PRS calculation using genotyping data while exome data was used for the gene-based scores (GBS) calculation. Association analysis was done using linear regression and genetic risk prediction models were also generated with either PRS, GBS, or both. We were able to show that rare pathogenic variants play an important role at an individual level, but the traditional PRS could be more informative when predicting the genetic risk at a population level.

In the last paper, we conduct a more thorough analysis on 72,469 samples from UK Biobank to investigate the rare-variants influence on male-pattern hair loss (MPHL). Novel candidate genes were identified including *HEPH*, *CEPT1* and *EIF3F*, further proving that rare variants contribute to the genetic landscape of complex phenotypes like male-pattern hair loss.

In conclusion, our findings indicate that rare deleterious variants have an essential role in complex phenotypes, and can be analysed to discover new targets in these traits. Nonetheless, further investigation needs to be conducted to effectively integrate the effects of rare and common variants, ultimately improving comprehensive genetic risk assessment strategies.

2. Introduction with aims and references

Human genetic traits are generally split into two categories: monogenic and polygenic. Monogenic traits are typically determined by one gene or allele, and they follow clear patterns, like dominant or recessive inheritance (Cleyneen and Halfvarsson, 2019). On the other hand, polygenic traits, also known as complex traits, are caused by variations in multiple genes and can also be influenced by other external factors, like diet or smoking. Thus, they do not follow a specific pattern, which makes it difficult to identify their underlying factors (Muthuirulan and Capellini, 2019). Complex traits represent a vast and diverse array of human characteristics, including height, intelligence, susceptibility to common diseases, and even personality traits. Furthermore, it has been recently suggested by multiple sources that the traditional classifications of phenotypes into monogenic or polygenic traits is oversimplifying the underlying genetic causality (Katsanis, 2016; Kousi and Katsanis, 2015). For instance, many studied common diseases are characterized by both familial and sporadic forms, such as diabetes (Karges et al., 2020), cardiovascular diseases (Dai, 2016), and a number of neurodegenerative disorders (Piaceri, 2013; Tang et al., 2017). This adds yet another layer of complexity into identifying and understanding the causality of these traits.

2.1 Genetics of complex phenotypes

Heritability, also known as H^2 , is a measure of the proportion of phenotypic variation in a population that can be attributed to genetic factors (Visscher et al., 2008). Genetic variants can generally be categorized based on their minor allele frequency (MAF) and effect sizes, as seen in Figure 1. Alleles with high effect size are usually rare and disease causing. More common variants, or genetic regions, which are also known as Quantitative Trait Loci (QTL), mostly have low effect sizes and may influence a phenotype either positively or negatively. Individually, QTLs make subtle contributions to the trait, but the aggregate of these genetic contributions creates a spectrum of trait values across a population (Powder, 2019). Genome-Wide Association Studies (GWAS) have been instrumental in identifying these genetic variants associated with complex traits, shedding light on their polygenic nature. However, since common variants usually have small effect sizes individually, it is challenging to capture their collective impact (Uffelmann et al., 2021). Furthermore, while GWAS studies have identified numerous variants associated with complex traits, their cumulative effects often account for only a small fraction of the estimated heritability, which falls short of explaining the expected heritability based on familial studies. This is known as the 'missing heritability' problem (Golan et al., 2014).

Different hypotheses have been proposed and investigated to address this gap. It has even been suggested that the heritability observed in family and twin studies might be

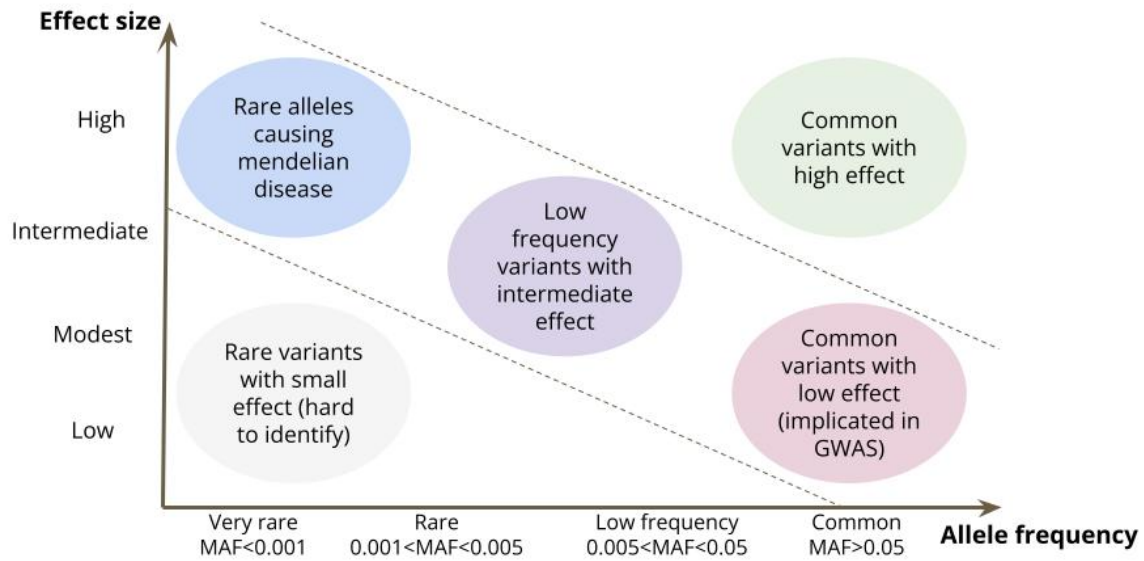


Figure 1: The relationship between the frequency of an allele and its effect size. Generally, rare and very rare alleles ($MAF < 0.005$) have high effect sizes while common variants ($MAF > 0.05$) have low to modest effect sizes. Less frequently, some common variants might have high effect sizes, and some rare variants might have low effect sizes. Image adapted from (Whitcomb et al., 2015)

simply overestimated (Felson, 2014). GWAS typically do not comprehensively capture the contribution of rare and low-frequency variants or structural variants (such as copy number variations). Thus, one hypothesis suggests that these types of genetic variations, which may have larger effect sizes, could account for a significant portion of the missing heritability (Lee et al., 2014). In fact, many studies have investigated the effect of rare variants and observed that they play a role in different complex phenotypes such as hypertension (Surendran et al., 2020), autism spectrum disorder (More et al., 2023), and diabetes (Deaton et al., 2021).

Other studies suggest that non-linear effects might be in play here. One of those effects would be epigenetic modifications, such as DNA methylation and histone modifications, which can add an additional layer of complexity to the etiology of complex traits. Epigenetic changes can modulate the activity of genes without altering the underlying DNA sequence. They can be influenced by both genetic and environmental factors, creating a dynamic and responsive system (Handy et al., 2011). These changes can persist across generations, potentially contributing to the transgenerational inheritance of complex traits (Kilpinen and Dermitzakis, 2012). Another non-linear effect that could contribute to complex traits is gene-gene interactions, also known as epistasis. Here the combination of specific genetic

variants may have a more substantial impact on the trait than each variant individually (Wei et al., 2014). Similarly, gene-environment interactions can modify the genetic effects on a trait (Laville et al., 2022).

2.2 The role of environment in complex traits

The environment plays a crucial role in the development and expression of complex traits. Environmental influences consist of a wide variety of factors, including diet, lifestyle, exposure to toxins, and socioeconomic conditions. These external factors can interact with genetic variants to shape the trait's ultimate expression. For example, a person with a genetic predisposition to obesity may or may not become overweight depending on their diet and physical activity level (van Vliet-Ostaptchouk et al., 2012). The impact of environmental factors can be immediate or occur over a longer time frame, such as in developmental or epigenetic processes.

2.3 Genetic Risk Analysis and Assessment

Genetic risk assessment is a comprehensive process that evaluates an individual's likelihood of developing a particular genetic condition or disease based on their genetic makeup (Igo et al., 2019). Traditionally, genetic risk assessment has focused on common genetic variants, often identified through GWAS. More recent approaches recognize the importance of rare variants, which, although individually infrequent, can have a significant impact on disease risk (Wainschtein et al., 2022).

The most commonly used method for genetic risk prediction is Polygenic Risk Scores (PRS). PRS aggregates the effects of numerous genetic variants to estimate an individual's overall risk for a certain disease or trait (Wang et al., 2022). These variants' effects are generally derived from previous GWAS analyses, which use a univariate approach (i.e., the association of each variant to a phenotype is done independently) (Uffelmann et al., 2021). However, more recent multivariate methods have been established to derive PRS, such as *snpnet*, which applies regression on the highly-dimensional genotyping data in batches (Klinkhammer et al., 2023). Nevertheless, these methods only account for the genetic disposition coming from common variants.

On the other hand, rare variants' contribution to phenotypes has been mostly studied using burden tests. These tests calculate the genetic burden of all variants in a genetic region into one score which is then analyzed. However, some limitations for burden tests include the assumption of the causality and directionality of the variants' effect on the traits (Kosmicki et al., 2016). Variance-component tests were developed to overcome these limitations. One of the most widely used of those tests is Sequence Kernel Association Test (SKAT), which is a supervised test that aggregates the statistics of multiple variants in a

region and evaluates their distribution (Wu et al., 2011). The method was further improved to incorporate both common and rare variants, in an extended method called SKAT-O (Lee et al., 2012). One disadvantage of SKAT and SKAT-O is that they do not provide scores or processed data at an individual-level.

2.4 Aims of the study

In this thesis, we hypothesize that rare and low-frequency variants with modest to high effect sizes could contribute to the genetic landscape of complex phenotypes. Thus, the aim of this project is to generate and optimize an analytical framework for the comprehensive evaluation of genetic risk by systematically integrating the effects of rare highly damaging variants and the polygenic component that is attributable to common variants. To achieve our goal, GenRisk, a python package for comprehensive common and rare variant analysis, was implemented (publication 1). GenRisk pipeline was used to perform association analyses and genetic risk modeling on 28 blood biomarkers as quantitative phenotypes (publication 2). The same analyses were also performed on male-pattern hair loss as binary phenotype (publication 3). The analyses from GenRisk were also compared with other previously established methods. Genetic risk modeling was performed using different machine learning models to account for associations that could arise for non-linear interactions and environmental factors.

References

- Isabelle Cleyngen and Jonas Halfvarsson (2019). How to approach understanding complex trait genetics – inflammatory bowel disease as a model complex trait. *United European Gastroenterology Journal*, 7(10):1426–1430.
- Xuming Dai (2016). Genetics of coronary artery disease and myocardial infarction. *World Journal of Cardiology*, 8(1):1.
- Aimee M. Deaton, Margaret M. Parker, Lucas D. Ward, Alexander O. Flynn-Carroll, Lucas BonDurant, et al. (2021). Gene-level analysis of rare variants in 379, 066 whole exome sequences identifies an association of *gigyl1* loss of function with type 2 diabetes. *Scientific Reports*, 11(1).
- Jacob Felson (2014). What can we learn from twin studies? a comprehensive evaluation of the equal environments assumption. *Social Science Research*, 43:184–199.
- David Golan, Eric S. Lander, and Saharon Rosset (2014). Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49).

- Diane E. Handy, Rita Castro, and Joseph Loscalzo (2011). Epigenetic modifications: Basic mechanisms and role in cardiovascular disease. *Circulation*, 123(19):2145–2156.
- Robert P. Igo, Tyler G. Kinzy, and Jessica N. Cooke Bailey (2019). Genetic risk scores. *Current Protocols in Human Genetics*, 104(1).
- Beate Karges, Nicole Prinz, Kerstin Placzek, Nicolin Datz, Matthias Papsch, et al. (2020). A comparison of familial and sporadic type 1 diabetes among young patients. *Diabetes Care*, 44(5):1116–1124.
- Nicholas Katsanis (2016). The continuum of causality in human genetic disorders. *Genome Biology*, 17(1).
- H. Kilpinen and E. T. Dermitzakis (2012). Genetic and epigenetic contribution to complex traits. *Human Molecular Genetics*, 21(R1):R24–R28.
- Hannah Klinkhammer, Christian Staerk, Carlo Maj, Peter Michael Krawitz, and Andreas Mayr (2023). A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Frontiers in Genetics*, 13.
- Jack A. Kosmicki, Claire L. Churchhouse, Manuel A. Rivas, and Benjamin M. Neale (2016). Discovery of rare variants for complex phenotypes. *Human Genetics*, 135(6):625–634.
- M. Kousi and N. Katsanis (2015). Genetic modifiers and oligogenic inheritance. *Cold Spring Harbor Perspectives in Medicine*, 5(6):a017145–a017145.
- Vincent Laville, Timothy Majarian, Yun J. Sung, Karen Schwander, Mary F. Feitosa, et al. (2022). Gene-lifestyle interactions in the genomics of human complex traits. *European Journal of Human Genetics*, 30(6):730–739.
- Seunggeun Lee, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin (2014). Rare-variant association analysis: Study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23.
- Seunggeun Lee, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, et al. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237.
- Ravi Prabhakar More, Varun Warriar, Helena Brunel, Clara Buckingham, Paula Smith, et al. (2023). Identifying rare genetic variants in 21 highly multiplex autism families: the role of diagnosis and autistic traits. *Molecular Psychiatry*, 28(5):2148–2157.
- Pushpanathan Muthurulan and Terence D. Capellini (2019). Complex phenotypes: Mechanisms underlying variation in human stature. *Current Osteoporosis Reports*, 17(5):301–323.

- Irene Piaceri (2013). Genetics of familial and sporadic alzheimer s disease. *Frontiers in Bioscience*, E5(1):167–177.
- Kara E. Powder (2019). *Quantitative Trait Loci (QTL) Mapping*, page 211–229. Springer US.
- Praveen Surendran, Elena V. Feofanova, Najim Lahrouchi, Ioanna Ntalla, Savita Karthikeyan, et al. (2020). Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals. *Nature Genetics*, 52(12):1314–1332.
- Yan Tang, Xue Xiao, Hua Xie, Chang-min Wan, Li Meng, et al. (2017). Altered functional brain connectomes between sporadic and familial parkinson's patients. *Frontiers in Neuroanatomy*, 11.
- Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, et al. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1).
- Jana V. van Vliet-Ostaptchouk, Harold Snieder, and Vasiliki Lagou (2012). Gene–lifestyle interactions in obesity. *Current Nutrition Reports*, 1(3):184–196.
- Peter M. Visscher, William G. Hill, and Naomi R. Wray (2008). Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.
- Pierrick Wainschein, Deepti Jain, Zhili Zheng, Stella Aslibekyan, Diane Becker, et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics*, 54(3):263–273.
- Ying Wang, Kristin Tsuo, Masahiro Kanai, Benjamin M. Neale, and Alicia R. Martin (2022). Challenges and opportunities for developing more generalizable polygenic risk scores. *Annual Review of Biomedical Data Science*, 5(1):293–320.
- Wen-Hua Wei, Gibran Hemani, and Chris S. Haley (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733.
- David C. Whitcomb, Celeste A. Shelton, and Randall E. Brand (2015). Genetics and genetic testing in pancreatic cancer. *Gastroenterology*, 149(5):1252–1264.e4.
- Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

3. Publications

3.1 Publication 1 - GenRisk: a tool for comprehensive genetic risk modeling

This publication describes an open source python package called GenRisk, which implements several modules for a comprehensive genetic risk analysis. The pipeline contains modules that perform gene-based scores calculations, association analyses, PRS calculation and prediction modeling. Unlike many other pipelines, this tool has many adjustable features, which allows flexible implementation depending on downstream analyses. The documentation for GenRisk can be found in subsection 3.1.1 (Publication 1 - Appendix A) as part of the supplementary material for this paper. As the first author, I was directly involved in the planning of the work in this article. Apart from developing the GenRisk package, I had collected all the data needed and performed most the data analyses, evaluation and interpretation of the results.



Genetics and population analysis

GenRisk: a tool for comprehensive genetic risk modeling

Rana Aldisi ^{1,*}, Emadeldin Hassanin¹, Sugirthan Sivalingam^{1,2,3},
 Andreas Bunes^{1,2,3}, Hannah Klinkhammer^{1,3}, Andreas Mayr³, Holger Fröhlich^{4,5},
 Peter Krawitz ¹ and Carlo Maj^{1,6}

¹Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn 53127, Germany, ²Core Unit for Bioinformatics Analysis, University Hospital Bonn, Bonn 53127, Germany, ³Institute of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn 53127, Germany, ⁴Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, 53757 Sankt Augustin, Germany, ⁵Bonn-Aachen International Center for IT (b-it), University of Bonn, Bonn 53115, Germany and ⁶Centre for Human Genetics, University of Marburg, Marburg 35033, Germany

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on September 24, 2021; revised on February 4, 2022; editorial decision on March 4, 2022; accepted on March 9, 2022

Abstract

Summary: The genetic architecture of complex traits can be influenced by both many common regulatory variants with small effect sizes and rare deleterious variants in coding regions with larger effect sizes. However, the two kinds of genetic contributions are typically analyzed independently. Here, we present GenRisk, a python package for the computation and the integration of gene scores based on the burden of rare deleterious variants and common-variants-based polygenic risk scores. The derived scores can be analyzed within GenRisk to perform association tests or to derive phenotype prediction models by testing multiple classification and regression approaches. GenRisk is compatible with VCF input file formats.

Availability and implementation: GenRisk is an open source publicly available python package that can be downloaded or installed from Github (<https://github.com/AldisiRana/GenRisk>).

Contact: s0raaldi@uni-bonn.de

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

In the past decade, genome-wide association studies (GWAS) have been used extensively to investigate the genetic architecture of complex traits and diseases (Uffelmann *et al.*, 2021). However, despite the identification of many disease-associated common variants which also led to the development of several accurate polygenic risk score (PRS) models, a substantial part of the genetic architecture of common traits remains unknown (Lee *et al.*, 2014). This is known as missing heritability, which is the difference between the heritability observed in twins studies and the measured heritability explained by common variants (Génin, 2020).

Different studies suggested that the missing heritability is mainly attributable to rare variants (Young, 2019). In line with this hypothesis, many studies have observed that rare variants play a role in complex phenotypes, such as hypertension (Russo *et al.*, 2018), schizophrenia (John *et al.*, 2019) and autism (Havdahl *et al.*, 2021). Burden tests are among the most applied methods to investigate rare variant effects starting from sequencing data. These methods typically collapse rare variants in a genetic region (e.g. gene) into a single burden variable and then regress the phenotype on the burden variable to test for the cumulative effects of rare variants (Bomba *et al.*,

2017). On the other hand, the genetic contribution of common variants is typically analyzed by mean of PRS, which is usually computed as the weighted sum of risk alleles with respect to a phenotype, where the risk alleles and the corresponding weights are derived from a reference GWAS (Choi *et al.*, 2020).

Generally, gene-based burden tests are applied on exome/target sequencing data while GWAS is performed on post-imputed chip-array data for the genotyping of high-frequent variants. In the light of the increasing availability of whole genome sequencing data, there is a need of bioinformatics solutions integrating different methodological approaches into a unique framework. With this aim in mind, we developed GenRisk, a python package that seamlessly combines different tools and libraries to analyze genotype–phenotype associations by considering both polygenic effects and the enrichment of rare deleterious variants at gene-based level.

2 Implementation

The GenRisk pipeline contains multiple modules, which can be run using a commandline interface or within a python environment. The modules can be run sequentially, so that the input of a module is the

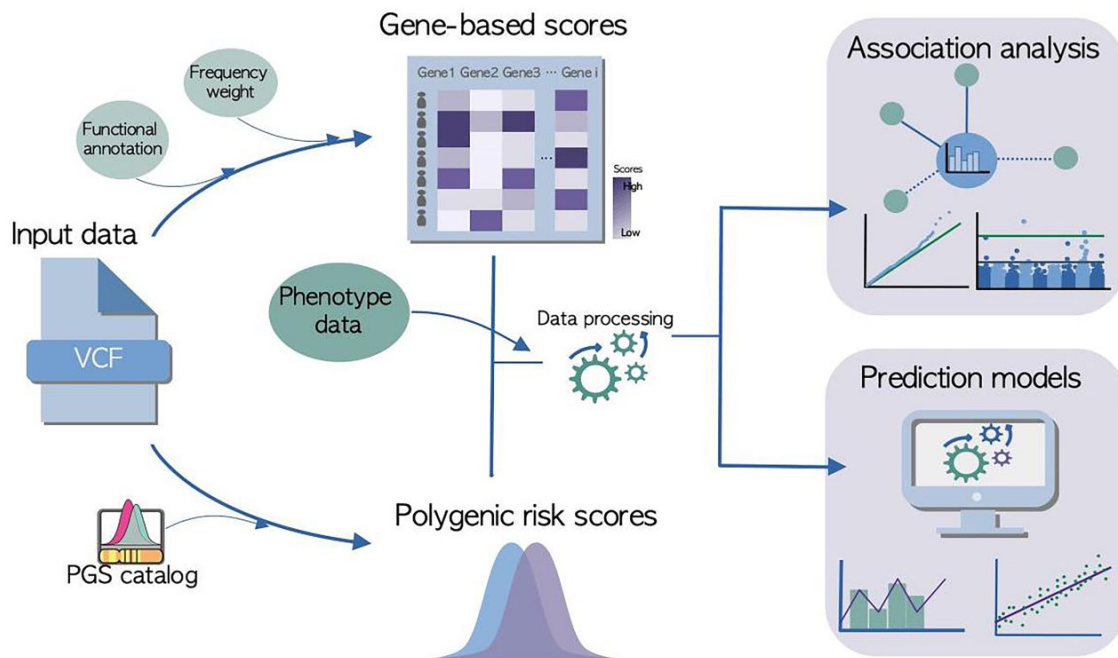


Fig. 1. GenRisk pipeline workflow. A VCF file with functional annotations and frequencies can be used to calculate gene-based scores, alternatively a VCF can be used to extract and calculate PRS. The scores can then be used with phenotypic data for association analysis or to develop prediction models

output of the previous module. In addition, each module can also be used independently with data provided by the user to increase flexibility of the tool for custom-analyses. Starting from a VCF, GenRisk computes gene scores based on variant annotations. Given a phenotype and potential covariates (possibly including PRS), the individual gene scores can be used to perform association analyses and to build phenotype prediction models. Furthermore, an interactive command implements PRS computation, the PRS model can be either provided by the user or available in pgscatalog (<https://www.pgscatalog.org/>).

The workflow of the pipeline is summarized in Figure 1. In the following sections the main features of GenRisk are described.

2.1 Gene-based scoring system

The gene scores are derived by the weighted sum of the variants in a gene. Each allele count is weighted according to the product of a deleteriousness score and a coefficient based on the allele frequency. Namely, a weighting function is applied to the variant frequency to potentially up-weight the biological importance of rare variants. Two weighting functions are implemented, $-\log_{10}$ as already applied in another gene-based score tool (Mossotto et al., 2019) and the beta density function, which contains two parameters α and β that can be adjusted for more flexible weight calculation as implemented in the sequence kernel association test (Lee et al., 2012). An adjustable threshold parameter for the minor allele frequency (MAF) can be also considered to filter only for rare variants.

2.2 Genetic risk scores analysis

According to the distribution of the scores, different statistical tests can be applied to analyze gene-phenotype associations starting from the derived individual-based gene scores. The association analysis results are generated as summary statistics and can be visualized via QQ-plots and Manhattan plots.

Prediction models are computed using the open-source Pycaret, a machine learning python library (Ali, 2020). The models can be generated for both quantitative and binary traits. The gene-based scores, as well as PRS and covariates, such as sex and age, can be used as features. The data given by the user can be divided into

training and testing sets (with flexible size). Cross-validation is applied on different models and the best performing model is selected, tuned and finalized. The model is then saved and can be further evaluated with external testing sets. Model evaluation reports and testing set labels are exported. Graphs like, feature importance, confusion matrix and prediction error, are also generated to visualize the model performance.

3 Usage case

We applied the pipeline on $\approx 160\,000$ samples from UK Biobank (application number 81202), the gene-based scores were calculated by applying the beta weighting function ($\alpha = 1$, $\beta = 2.5$) to up-weight rare variants while the CADD (Rentzsch et al., 2019) raw scores were used as deleteriousness weight and only variants with $MAF < 1\%$ were included. The derived scores were used for association test and prediction model with respect to alkaline phosphatase measurements (Field 30610) including also the first four genotyping principle components, sex, BMI and age as covariates. The association analysis based on a linear regression model detected significance in ALPL, GPLD1 and ASGR1 genes, all of which have been previously associated with alkaline phosphatase (Nioi et al., 2016; Yuan et al., 2008). In addition, a stochastic gradient boosted decision tree algorithm was identified as the best prediction model once both gene scores and PRS (from Sinnott-Armstrong et al., 2021) are taken into account and it showed an improved prediction performance compared with PRS-only model.

Detailed results, as well as comparisons with other methods, can be found in Supplementary Material.

4 Conclusion

GenRisk is a python package that processes input VCF files to generate both gene-based burden scores and PRS for association tests and development of prediction models. GenRisk provides a framework to model the effects of rare functional variants while considering the polygenic background. Thus, it is suitable for the analysis of phenotypes characterized by a complex genetic architecture.

Funding

C.M. and E.H. were supported by the BONFOR-program of the Medical Faculty, University of Bonn (O-147.0002).

Conflict of Interest: none declared.

Data availability

Genome-wide genotyping data, exome-sequencing data, and phenotypic data from the UK Biobank are available upon successful project application (<http://www.ukbiobank.ac.uk/about-biobank-uk/>). Restrictions apply to the availability of these data, which were used under license for the current study (Project ID: 81202).

References

- Ali, M. (2020) *PyCaret: An Open Source, Low-Code Machine Learning Library in Python. PyCaret Version 1.0.* <https://pycaret.gitbook.io/docs/#citation>.
- Bomba, L. *et al.* (2017) The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.*, **18**, 77.
- Choi, S.W. *et al.* (2020) Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.*, **15**, 2759–2772.
- Génin, E. (2020) Missing heritability of complex diseases: case solved? *Hum. Genet.*, **139**, 103–113.
- Havdahl, A. *et al.* (2021) Genetic contributions to autism spectrum disorder. *Psychol. Med.*, **51**, 2260.
- John, J. *et al.* (2019) Rare variant based evidence for oligogenic contribution of neurodevelopmental pathway genes to schizophrenia. *Schizophrenia Res.*, **210**, 296–298.
- Lee, S. *et al.*; NHLBI GO Exome Sequencing Project—ESP Lung Project Team. (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, **91**, 224–237.
- Lee, S. *et al.* (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.
- Mossotto, E. *et al.* (2019) GenePy – a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinformatics*, **20**, 254.
- Nioi, P. *et al.* (2016) VariantASGR1 associated with a reduced risk of coronary artery disease. *N. Engl. J. Med.*, **374**, 2131–2141.
- Rentzsch, P. *et al.* (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
- Russo, A. *et al.* (2018) Advances in the genetics of hypertension: the effect of rare variants. *Int. J. Mol. Sci.*, **19**, 688.
- Sinnott-Armstrong, N. *et al.*; FinnGen. (2021) Genetics of 35 blood and urine biomarkers in the UK biobank. *Nat. Genet.*, **53**, 185–194.
- Uffelmann, E. *et al.* (2021) Genome-wide association studies. *Nat. Rev. Methods Primers*, **1**, 59.
- Young, A.I. (2019) Solving the missing heritability problem. *PLOS Genet.*, **15**, e1008222.
- Yuan, X. *et al.* (2008) Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am. J. Hum. Genet.*, **83**, 520–528.

3.1.1 Publication 1 - Appendix A

This appendix contains the documentation of the package GenRisk, which is described in Publication 1. The documentation lists the functionalities of the GenRisk and shows use case example results. For a better visualization of the documentation, please visit the online version at: genrisk.readthedocs.io

Welcome to GenRisk's documentation!

GenRisk is a package that implements different gene-based scoring schemes to analyze and find significant genes within a phenotype in a population

Citation

Rana Aldisi, Emadeldin Hassanin, Sugirthan Sivalingam, Andreas Bunes, Hannah Klinkhammer, Andreas Mayr, Holger Fröhlich, Peter Krawitz, Carlo Maj, GenRisk: a tool for comprehensive genetic risk modeling, Bioinformatics, Volume 38, Issue 9, 1 May 2022, Pages 2651–2653, <https://doi.org/10.1093/bioinformatics/btac152>

Installation

Requirements

- [plink](#) >= 1.9
- R version >= 3.6.3
- python >= 3.

Package installation

Option 1: The latest release of `GenRisk` can be installed on python3+ with:

```
pip install genrisk
```

Option2: you can also install the package with the latest updates directly from [GitHub](#) with:

```
pip install git+https://github.com/AldisiRana/GenRisk.git
```

Indices and tables

- [Index](#)
- [Module Index](#)

Command Line Interface

Note

Detailed information about the functions can be found in the [pipeline](#).

The genrisk command line interface includes multiple commands which can be used as follows:

genrisk score-genes

Calculate the gene-based scores for a given dataset.

Example

```
$ genrisk score-genes -a /path/to/toy_vcf_info.vcf -o toy_genes_scores.tsv -t  
toy_vcf_scoring -v ID -f AF -g gene -l ALT -d RawScore
```

Parameters

annotation_file : str

an annotation file containing variant IDs, alt, AF and deleterious scores.

bfiles : str

the binary files for plink process.

plink : str

the location of plink, if not set in environment

beta_param : tuple

the parameters from beta weight function.

temp_dir : str

a temporary directory to save temporary files before merging.

output_file : str

the location and name of the final output scores matrix.

weight_func : str

the weighting function used on allele frequency in score calculation. [beta| log10]

variant_col : str

the column containing the variant IDs.

gene_col : str

the column containing gene names. If the genes are in the INFO column, use the identifier of the value (i.e gene=IF, identifier is 'gene')

af_col : str

the column containing allele frequency. If in INFO, follow previous example

del_col : str

the column containing deleteriousness score (functional annotation). If in INFO, follow previous example

alt_col : str

the column containing alternate base.

maf_threshold : float

the threshold for minor allele frequency.

Returns

DataFrame information

the final scores dataframe information the DataFrame is saved into the output path indicated in the arguments

```
genrisk score-genes [OPTIONS]
```

Options

-a, --annotation-file <annotation_file>

Required an annotation file containing variant IDs, alt, AF and deleterious scores.

-b, --bfiles <bfiles>

provide binary files that contain the samples info

--plink <plink>

the directory of plink, if not set in environment

-t, --temp-dir <temp_dir>

Required a temporary directory to save temporary files before merging.

-o, --output-file <output_file>

Required the final output path

-p, --beta-param <beta_param>

the parameters from beta weight function.

Default: `1.0, 25.0`

-w, --weight-func <weight_func>

the weighting function used in score calculation.

Default: `'beta'`

Options: `beta | log10`

-v, --variant-col <variant_col>

the column containing the variant IDs.

Default: `'SNP'`

-g, --gene-col <gene_col>

the column containing gene names.

Default: `'Gene.refGene'`

-f, --af-col <af_col>

the column containing allele frequency.

Default: `'MAF'`

-d, --del-col <del_col>

the column containing the deleteriousness score.

Default: `'CADD_raw'`

-l, --alt-col <alt_col>

the column containing the alternate base.

Default: `'Alt'`

-m, --maf-threshold <maf_threshold>

the threshold for minor allele frequency.

Default: `0.01`

-k, --keep

if flagged temporary files will not be deleted.

--vcf <vcf>

provide vcf that contain the samples info

genrisk normalize

Normalize/standarize data.

Example

```
$ genrisk normalize --data-file toy_example/toy_dataset_scores --method gene_length --
samples-col IID
--output-file toy_dataset_scores_normalized.tsv
```

Parameters

genes_info : str

the file containing genes names and length. if not provided ensembl database is used to retrieve data.

method : str

the method of normalizing data. [gene_length|zscore|minmax|maxabs|robust]

data_file : str

the file containg data to be normalized.

samples_col : str

the column containing sample ids.

genes_col : str

the column containing gene names. ignore if genes_info file is not provided.

lengths_col : str

the column containing gene lengths. ignore if genes_info file is not provided.

output_file : str

the name of the file for final output

Returns

DataFrame with normalized data.

```
genrisk normalize [OPTIONS]
```

Options

--method <method>

Required

Options: gene_length | zscore | minmax | maxabs | robust

--data-file <data_file>

Required

--genes-info <genes_info>

-m, --samples-col <samples_col>

the name of the column that contains the samples.

Default: `'IID'`

--genes-col <genes_col>

Default: `'HGNC symbol'`

--lengths-col <lengths_col>

Default: `'gene_length'`

-o, --output-file <output_file>

Required the final output path

genrisk find-association

Calculate the P-value between two given groups.

Example

```
$ genrisk find-association --scores-file toy_example/toy_dataset_scores --info-file
toy_example/toy.pheno --phenotype trait1,trait2 --samples-column IID --test logit
--covariates age,sex --adj-pval bonferroni
```

Parameters

scores_file : str

the file containing gene-based scores.

info_file : str

file containing the phenotype.

genes : str

a file that contains a list of genes to calculate p-values. if not, all genes in scoring file will be used.

phenotype : str

the name of the column with phenotypes. Phenotypes can be either binary or quantitative.

samples_col : str

the name of the column with sample IDs. All files need to have the same format.

test : str

the statistical test used for calculating p-values.

adj_pval : str, optional

the method used to adjust the p-values.

covariates : str, optional

the covariates used for calculation. Not all tests are able to include covariates. (e.g. Mann Whinteny U doesn't allow for covariates)

processes : int, optional

if more than 1 processor is selected, the function will be parallelized.

Returns

DataFrame information

the final dataframe information the DataFrame is saved into the output path indicated in the arguments

```
genrisk find-association [OPTIONS]
```

Options

-s, --scores-file <scores_file>

Required The scoring file of genes across a population.

-i, --info-file <info_file>

Required File containing information about the cohort.

-g, --genes <genes>

a file containing the genes to calculate. if not provided all genes will be used.

-t, --test <test>

Required statistical test for calculating P value.

Options: ttest_ind | mannwhitneyu | logit | linear

-c, --phenotype <phenotype>

Required the name of the column that contains the case/control or quantitative vals.

-m, --samples-col <samples_col>

the name of the column that contains the samples.

Default: IID

-a, --adj-pval <adj_pval>

Options: bonferroni | sidak | holm-sidak | holm | simes-hochberg | hommel | fdr_bh |
fdr_by | fdr_tsbh | fdr_tsbky

-v, --covariates <covariates>

the covariates used for calculation

-p, --processes <processes>

number of processes for parallelization

Default: 1

genrisk visualize

Visualize manhattan plot and qqplot for the data.

Example

```
$ genrisk visualize --pvals-file toy_example/toy_dataset_scores
--info-file annotated_toy_dataset.vcf
```

Parameters

pvals_file : str

the file containing the calculated p-values.

info_file : str

file containing variant/gene info.

genescol_1 : str

the name of the genes column in pvals file.

genescol_2 : str

the name of the genes column in info file.

pval_col : str

the name of the pvalues column.

chr_col : str

the name of chromosomes column.

pos_col : str

the name of the position/start column.

Returns

```
genrisk visualize [OPTIONS]
```

Options

-p, --pvals-file <pvals_file>

Required the file containing p-values.

-i, --info-file <info_file>

file containing variant/gene info.

--genescol-1 <genescol_1>

the name of the genes column in pvals file.

Default: `'genes'`

--genescol-2 <genescol_2>

the name of the genes column in info file.

Default: `'Gene.refGene'`

-v, --pval-col <pval_col>

the name of the pvalues column.

Default: `'p_value'`

-c, --chr-col <chr_col>

the name of the chromosomes column

Default: `'Chr'`

-s, --pos-col <pos_col>

the name of the position/start of the gene column

Default: `'Start'`

genrisk create-model

Create a prediction model with given dataset.

Example

```
$ genrisk create-model --data-file toy_example_regressor_features.tsv --model-type
regressor
--output-folder toy_regressor --test-size 0.25 --test --model-name toy_regressor
--target-col trait1 --imbalanced --normalize
```

Notes

The types of models available for training can be found `model_types`

Parameters

data_file : str

file containing features and target.

output_folder : str

a folder path to save all outputs.

test_size : float

the size of testing set.

test : bool

if True the dataset will be split into training and testing for extra evaluation after finalization.

model_name : str

the name of the model to be saved.

model_type : str

the type of model [regressor| classifier].

target_col : str

the name of the target column in data file.

imbalanced : bool

if true methods will be used to account for the imbalance.

normalize : bool

if true the data will be normalized before training

normalize_method : str

method used to normalize data. [zscore| minmax| maxab| robust]

folds : int

the number of folds used for cross validation

metric : str

the metric used to choose best model after training.

samples_col : str

the name of the column with samples IDs.

seed : int

random seed number to run the machine learning models.

include_models : str

list of specific models to compare. more information in the documentations

Returns

Final prediction model

```
genrisk create-model [OPTIONS]
```

Options

-d, --data-file <data_file>

Required file with all features and target for training model.

-o, --output-folder <output_folder>

Required path of folder that will contain all outputs.

-i, --test-size <test_size>

test size for cross validation and evaluation.

Default: `0.25`

-n, --model-name <model_name>

Required name of model file.

--model-type <model_type>

Required type of prediction model.

Options: regressor | classifier

-l, --target-col <target_col>

Required name of target column in data_file.

-b, --imbalanced

if flagged methods will be used to account for the imbalance.

--normalize

if flagged the data will be normalized before training.

--normalize-method <normalize_method>

features normalization method.

Default: `'zscore'`

Options: zscore | minmax | maxabs | robust

-f, --folds <folds>

number of cross-validation folds in training.

Default: `10`

--metric <metric>

the metric used to choose best model after training.

-m, --samples-col <samples_col>

the name of the column that contains the samples.

Default: `'IID'`

--seed <seed>

add number to create reproducible train_test splitting.

--include-models <include_models>

choose specific models to compare with comma in between. e.g lr,gbr,dt

--feature-selection

if selected feature selection will be implemented in training.

genrisk test-model

Evaluate a prediction model with a given dataset.

Example

```
$ genrisk test-model --model-path regressor_model.pkl --input-file
testing_dataset.tsv
--model-type regressor --labels-col target --samples-col IID
```

Parameters

model_path : str

the path to the ML model.

input_file : str

the testing (independent) dataset.

model_type : str

the type of model [classifier|regressor].

label_col : str

the labels/target column.

samples_col : str

the sample ids column.

output_file : str

the path to the dataframe with the prediction results.

Returns

DataFrame

dataframe with the prediction results.

```
genrisk test-model [OPTIONS]
```

Options

-t, --model-type <model_type>

Required type of prediction model.

Options: regressor | classifier

-i, --input-file <input_file>

Required testing dataset

-l, --label-col <label_col>

Required the target/phenotype/label column

-m, --model-path <model_path>

Required path to the trained model.

-s, --samples-col <samples_col>

the samples column.

Default: `'IID'`

-o, --output-file <output_file>

Required the final output path

genrisk get-prs

Calculate PRS. This command is interactive. This command gets a pgs file (provided by the user or downloaded) then calculates the PRS for dataset.

Example

This function is performed using commandline interface:

```
$ genrisk get-prs
```

Parameters

plink : str

provide plink path if not default in environment.

Returns

```
genrisk get-prs [OPTIONS]
```

Options

-p, --plink <plink>

Pipeline functions

Gene scoring

```
genrisk.pipeline.scoring_process(*, logger, annotation_file, temp_dir, beta_param,  
weight_func, del_col, maf_threshold, gene_col, variant_col, af_col, alt_col, bfiles, plink, output_file, vcf)
```

[\[source\]](#)

Calculate gene-based scores. This is calculated by a weighted sum of the variants in a gene.

- Parameters:**
- **logger** – an object that logs function outputs.
 - **annotation_file** (*str*) – an annotation file containing variant IDs, alt, and another info.
 - **temp_dir** (*str*) – a temporary directory to save temporary files before merging.
 - **beta_param** (*tuple*) – the parameters from beta weight function. ignore if log10 function is chosen.
 - **weight_func** (*str*) – the weighting function used in score calculation.
 - **del_col** (*str*) – the column containing deleteriousness score or functional annotation.
 - **maf_threshold** (*float*) – the threshold for minor allele frequency. between [0.0-1.0]
 - **gene_col** (*str*) – the column containing gene names.
 - **variant_col** (*str*) – the column containing variant IDs.
 - **af_col** (*str*) – the column containing allele frequency.
 - **alt_col** (*str*) – the column containing alternate allele.
 - **bfiles** (*str*) – the binary files for plink process. if a vcf is provided no binary files are needed.
 - **plink** (*str*) – the directory of plink, if not set in environment
 - **output_file** (*str*) – the path to save the final output scores matrix.
 - **vcf** (*str*) – the vcf file for plink process. if binary files are provided no vcf is needed.
- Returns:** final scores matrix
- Return type:** DataFrame

Gene-scoring equation

The gene scores are derived by the weighted sum of the variants in a gene.

$$G_{sg} = \sum_{i=1}^k (D_i \times A_i) C_i$$

D_i is the functional annotation (e.g CADD)

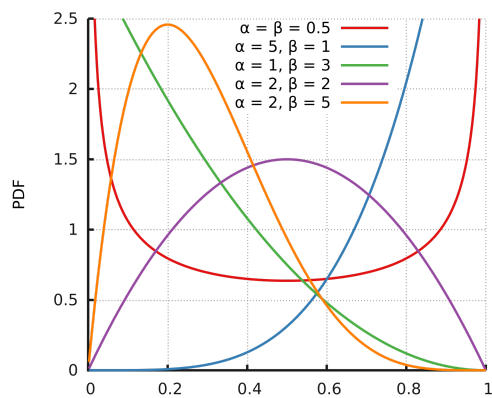
A_i is the weighted allele frequency

C_i is the allele count.

Weighting functions

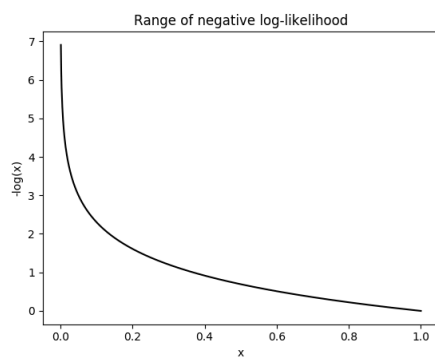
beta: this option uses two parameters α and β , to create beta distribution. Depending on the parameters chosen, the distribution can change its shape, giving more flexibility for the user to choose how to weight the variables.

The default for this function is [1,25] which are the same parameters used in SKAT-O.



[image source here](#)

log10: this option uses $-\log$ distribution to upweight rare variants. This has been applied previously in another [gene-based score tool](#)



[image source here](#)

Data normalization

genrisk.pipeline.normalize_data(*, *method='gene_length', genes_info=None, genes_col='HGNC symbol', length_col='gene_length', data_file, samples_col*) [\[source\]](#)

Normalize dataset using gene_length, minmax, maxabs, zscore or robust

Parameters:

- **method** (*str*) – the normalization method. [zscore, gene_length, minmax, maxabs, robust]
- **genes_info** (*str*) – file containing the genes and their lengths. if gene_length method chosen with no file, info will be retrieved from ensembl database.
- **genes_col** (*str*) – the column containing genes (if genes_info file is provided)
- **length_col** (*str*) – the columns containing genes length (if genes_info file is provided)
- **data_file** (*str*) – file containing dataset to be normalized.
- **samples_col** (*str*) – the column containing samples ids.

Returns: a df with the normalized dataset.

Return type: DataFrame

Normalization methods

Multiple methods have been implemented to normalize a dataset. Below is a brief description of each function.

gene_length: This method divides each gene-based score by the length of the gene. The genes lengths can be provided by the user, or retrieved from ensembl database. The gene length from ensembl database is calculated as such:
 $\text{gene length} = \text{gene end (bp)} - \text{gene start (bp)}$

minmax: This method rescales the values of each column to [0,1] by using the following formula $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$

maxabs: In this method, the values are normalized by the maximum absolute to [-1,1] using the following formula $x' = \frac{x}{\max(|x|)}$

zscore: This method uses the mean and standard deviation to normalize the values. Formula is $x' = \frac{x - \text{mean}(x)}{\text{std}}$

robust: Great choice for dataset with many outliers. In this method, the values are subtracted by the median then divided by the interquartile range (difference between the third and the first quartile). Formula $x' = \frac{x - \text{median}(x)}{Q3(x) - Q1(x)}$

Every normalization method has it's advantages and disadvantages, so choose the method that works best with your dataset. To learn more about the normalization methods, check out this helpful [article](#)

Association analysis

```
genrisk.pipeline.find_pvalue(*, scores_file, info_file, genes=None, phenotype, samples_column,  
test='mannwhitneyu', covariates=None, cases=None, controls=None, processes=1, logger,  
adj_pval=None) \[source\]
```

Calculate the significance of a gene in a population using different statistical analyses [mannwhitneyu, logit, linear, ttest_ind].

- Parameters:**
- **adj_pval** (*str*) – the method used to adjust the p-values.
 - **scores_file** (*str*) – dataframe containing the scores of genes across samples.
 - **info_file** (*str*) – a file containing the information of the sample. this includes target phenotype and covariates.
 - **genes** (*list*) – a list of the genes to calculate the significance. if None will calculate for all genes.
 - **phenotype** (*str*) – the name of the column containing phenotype information.
 - **samples_column** (*str*) – the name of the column containing samples IDs.
 - **test** (*str*) – the type of statistical test to use, choices are: ttest_ind, mannwhitneyu, linear, logit.
 - **covariates** (*str*) – the list of covariates used in the calculation.
 - **cases** (*str*) – the cases category. if binary phenotype.
 - **controls** (*str*) – the controls category. if binary phenotype.
 - **processes** (*int*) – number of processes used, for parallel computing.
 - **logger** – an object that logs function outputs.

Returns: dataframe with genes and their p_values

Return type: DataFrame

Beta regression function

```
genrisk.pipeline.betareg_pvalues(*, scores_file, pheno_file, samples_col, cases_col,  
output_path, covariates, processes, genes, logger) \[source\]
```

Calculate association significance using betareg. This function runs in Rscript.

- Parameters:**
- **scores_file** (*str*) – the path to the scores file.
 - **pheno_file** (*str*) – the path to the phenotypes and covariates file.
 - **samples_col** (*str*) – column containing samples ids.
 - **cases_col** (*str*) – column containing the phenotype.
 - **output_path** (*str*) – a path to save the summary statistics.
 - **covariates** (*str*) – the list of covariates used in the calculation.
 - **processes** (*int*) – number of processes used, for parallel computing.
 - **genes** (*str*) – a list of the genes to calculate the significance. if None will calculate for all genes.

- **logger** – an object that logs function outputs.

Prediction model generation

```
genrisk.pipeline.create_prediction_model(*, model_name='final_model',  
model_type='regressor', y_col, imbalanced=True, normalize=True, folds=10, training_set,  
testing_set=Empty DataFrame Columns: [] Index: [], metric=None, seed, include_models,  
normalize_method, feature_selection) \[source\]
```

Create a prediction model (classifier or regressor) using the provided dataset.

- Parameters:**
- **model_name** (*str*) – the name of the prediction model.
 - **model_type** (*str*) – type of model [regressor|classifier]
 - **y_col** (*str*) – the column containing the target (qualitative or quantitative).
 - **imbalanced** (*bool*) – True means data is imbalanced.
 - **normalize** (*bool*) – True if data needs normalization.
 - **folds** (*int*) – how many folds for cross-validation.
 - **training_set** (*pd.DataFrame*) – the training set for the model.
 - **testing_set** (*pd.DataFrame*) – if exists an extra evaluation step will be done using the testing set.
 - **test_size** (*float*) – the size to split the training set for cross-validation.
 - **metric** (*str*) – the metric to evaluate the best model.
 - **seed** (*int*) – the initialization state random number
 - **include_models** (*list*) – a list of models that the user wants to test. if None all models will be used.
 - **normalize_method** (*str*) – the method to normalize the data. Choices [zscore, minmax, maxabs, robust]

Return type: Final model

Utilities

```
genrisk.utils.draw_qqplot(*, pvals, qq_output) \[source\]
```

Generate QQ-plot for given data.

- Parameters:**
- **pvals** (*pd.Series*) – the list of p_values.
 - **qq_output** (*str*) – the path to output the QQplot image.

Return type: QQPlot

```
genrisk.utils.draw_manhattan(*, data, chr_col, pos_col, pvals_col, genes_col, manhattan_output) \[source\]
```

Generate manhattan plot from a given dataset.

- Parameters:**
- **data** (*pd.DataFrame*) – a dataframe with pvalues and gene information.
 - **chr_col** (*str*) – the column with the chromosomes.
 - **pos_col** (*str*) – the column containing the position/start.
 - **pvals_col** (*str*) – the column containing the position/start.
 - **genes_col** (*str*) – the column containing gene names.
 - **manhattan_output** (*str*) – the path to output the manhattan plot image.

Return type: Manhattan plot

Model types

Model types that can be computed. Lists taken from [Pycaret Documentation](#)

Regression models:

'lr' - Linear Regression

'lasso' - Lasso Regression

'ridge' - Ridge Regression

'en' - Elastic Net

'lar' - Least Angle Regression

'llar' - Lasso Least Angle Regression

'br' - Bayesian Ridge

'kr' - Kernel Ridge

'svm' - Support Vector Regression

'knn' - K Neighbors Regressor

'dt' - Decision Tree Regressor

'rf' - Random Forest Regressor

'et' - Extra Trees Regressor

'ada' - AdaBoost Regressor

'gbr' - Gradient Boosting Regressor

'xgboost' - Extreme Gradient Boosting

'lightgbm' - Light Gradient Boosting Machine

'catboost' - CatBoost Regressor

Classification models:

'kmeans' - K-Means Clustering

'ap' - Affinity Propagation

'meanshift' - Mean shift Clustering

'sc' - Spectral Clustering

'hclust' - Agglomerative Clustering

'dbscan' - Density-Based Spatial Clustering

'optics' - OPTICS Clustering

'birch' - Birch Clustering

'kmodes' - K-Modes Clustering

Example use case

The toy dataset is not real data. It contains real SNPs and gene info, but no real individuals. It was created for testing and evaluating the pipeline only. In this example use case, we use the toy dataset to generate gene-based scores, association analysis and machine learning models.

Annotations and PLINK files

([Click here](#) to download)

Annotations file

(filename: toy_vcf_data.tsv)

The annotated file contains information about the SNPs, gene, deleteriousness score and allele frequency. It should also contain samples genotypes. This information is important for the calculation of the gene-based scores.

Plink binary files

(filenames: toy_data.bed, toy_data.bim, toy_data.fam)

The binary files contain all genotype information for the cohort.

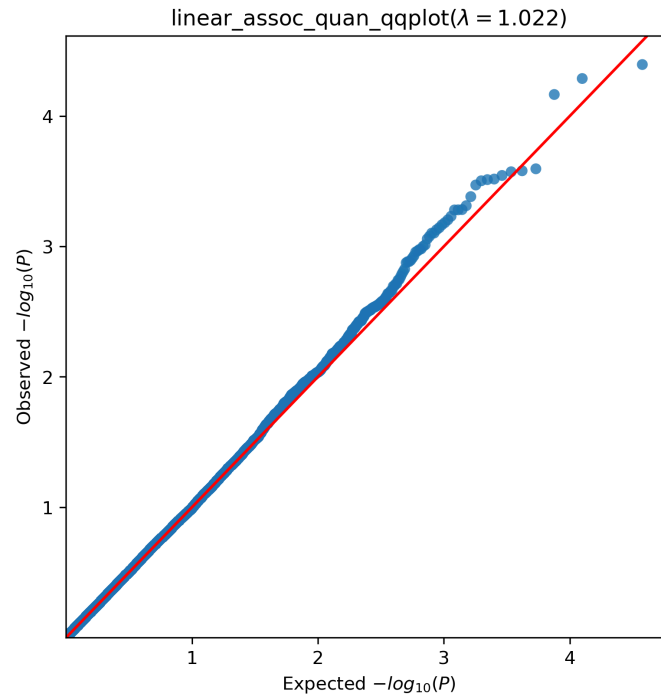
Gene-based scores

The gene-based scores can be found [here](#). These scores are used as input for the association analysis and as features for the machine learning models.

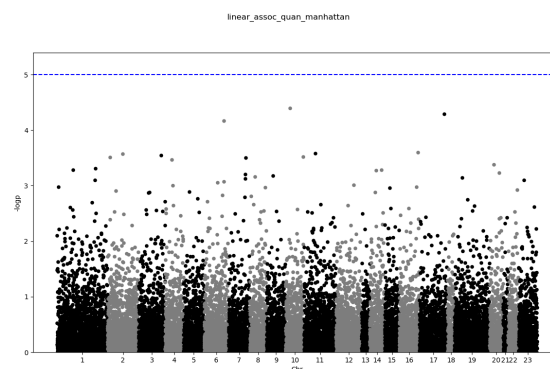
Association analysis

We performed linear regression on the quantitative phenotype and a logistic regression on the binary phenotype. Results can be found [here](#).

The QQ-plot of quantitative phenotype:



Manhattan plot of quantitative phenotype:



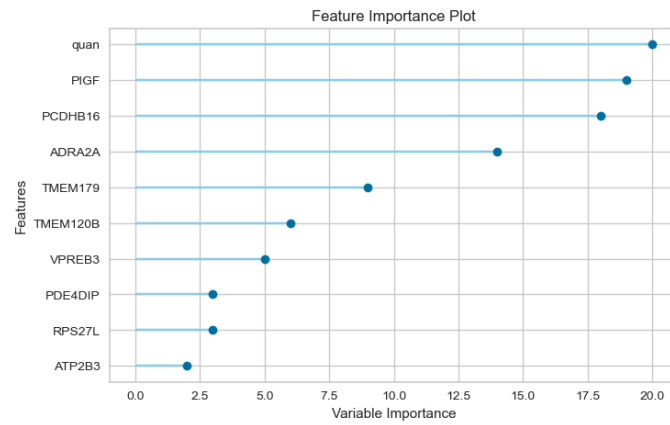
Machine learning models

We have created two types of models, regression and classification. Both models use the gene-based scores as features along with covariates. For the regression model, a quantitative trait was generated, while a binary trait was generated for the classification model. For each model, 10 fold cross validation was done on training dataset and an extra evaluation step was done on the testing set. The input for the model generation can be found [here](#).

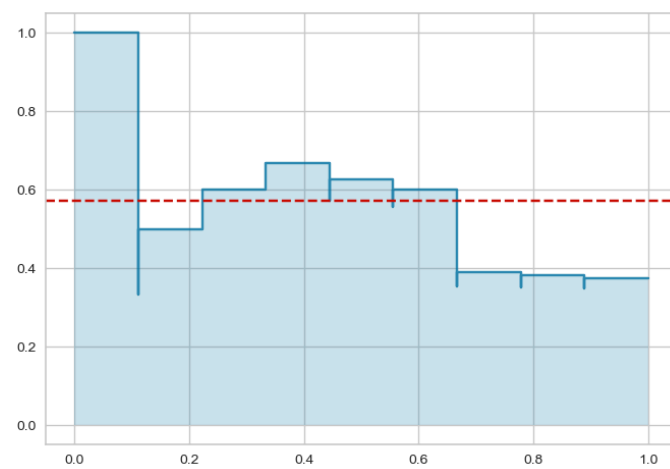
Classification model

Results of model training

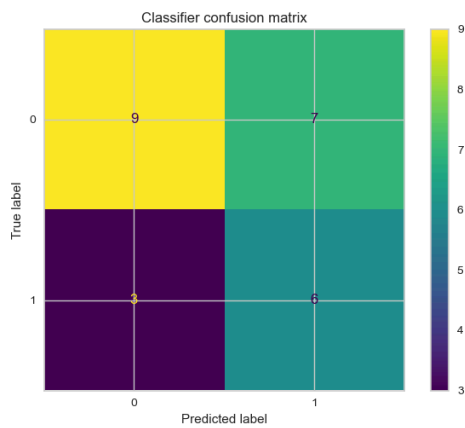
Feature importance:



Precision-recall curve:



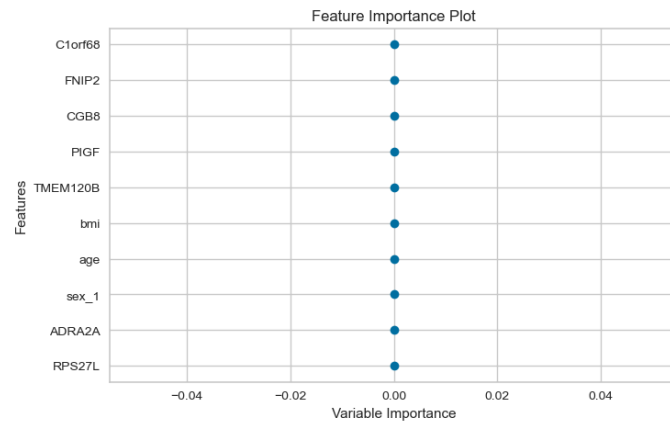
Confusion matrix:



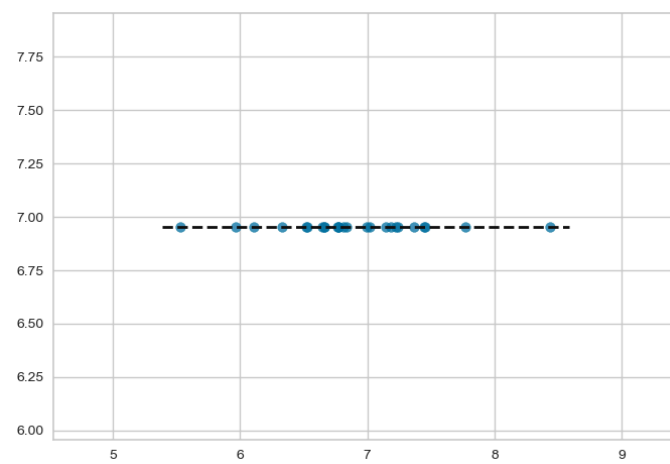
Regression model

Results for model training

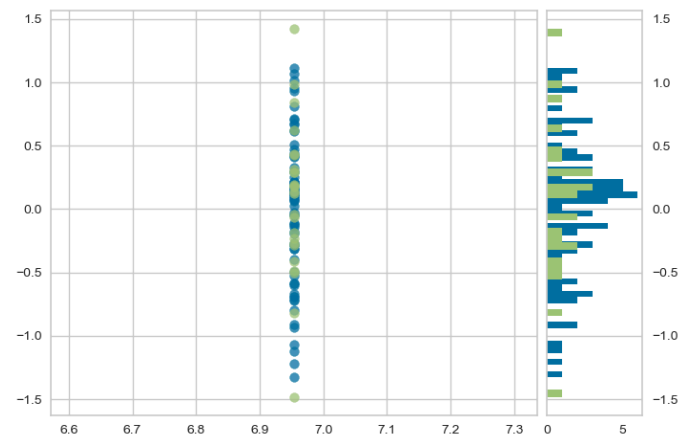
Feature importance:



Prediction error:



Residuals:



Real use case

Data

We ran the pipeline using about 200K samples from UKBiobank. We filtered for British to remove outliers (150K after filtering)

Gene scoring

We calculated the scores for the 150K samples.

MAF threshold:	1% MAF threshold.
Weighting function:	beta 1-25 weighting parameter.
Functional annotation:	CADD raw scores.

Association analysis

For each phenotype, we performed association analysis including all the samples (150K)

Method:	linear regression
Covariates:	age, sex, BMI and PC1-4

Prediction models

We also generated 3 models for each phenotype, a PRS prediction model, a gene-based prediction model and a combined model.

Feature Selection:	we used portion of the samples (50K) for feature selection with linear regression. The genes with p-values <0.05 were selected as features.
Covariates:	age, sex, BMI and PC1-4 were also included in the features.
model training:	Of the remaining 100K samples, 75% were used in training with 10-fold cross-validation.
model testing:	the 25% remaining of the dataset was used as external testing set for final model evaluation.
PRS calculation:	To avoid colinearity, we excluded variants that were included in the gene scores from the PRS calculation.

LDL Phenotype

We performed the analysis on the samples with LDL direct measurements as phenotype(quantitative).

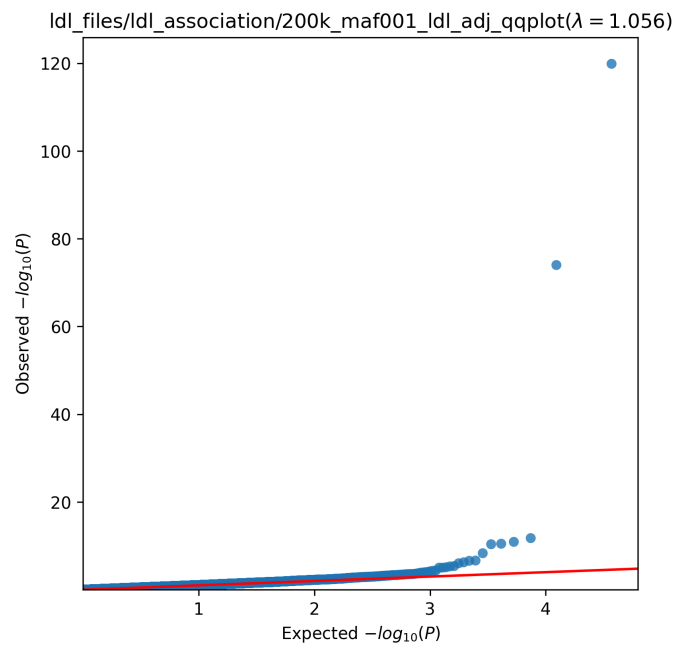
Note

For this phenotype we adjusted the values for individuals who take statin.

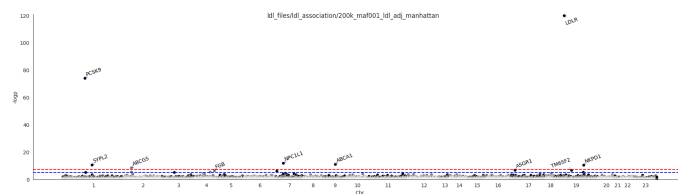
Association analysis

The association analysis highlighted PCSK9 and LDLR as significant genes, both are known to be associated with LDL. QQ-plot and Manhattan plot are presented below.

The QQ-plot:



The Manhattan plot:



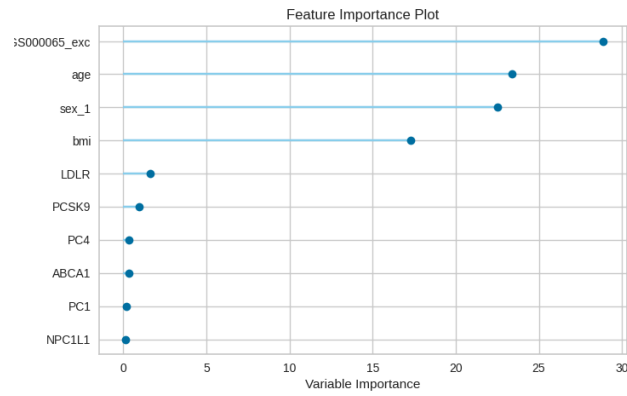
Regression model

For the prediction model, we used LDL direct measurements (adjusted for statin) as target. For features, we used the scores of 3 selected genes + BMI + age + sex + PC1-4. For the PRS and combined models we used the following PRS (PGS000688). The final prediction models was generated using gradient boosting regression, evaluation metric are shown in the table below.

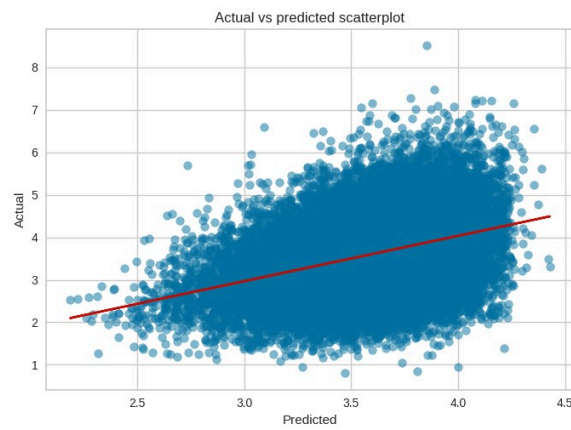
	Gene-based model	PRS model	Combined model
R ²	0.092	0.322	0.321

	Gene-based model	PRS model	Combined model
RMSE	0.849	0.729	0.725

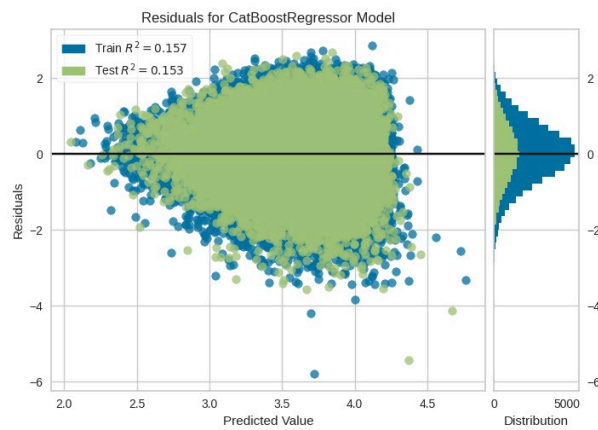
The images below are the output of the final combined model. Feature importance plot:



Actual vs Predicted:



Model residuals:



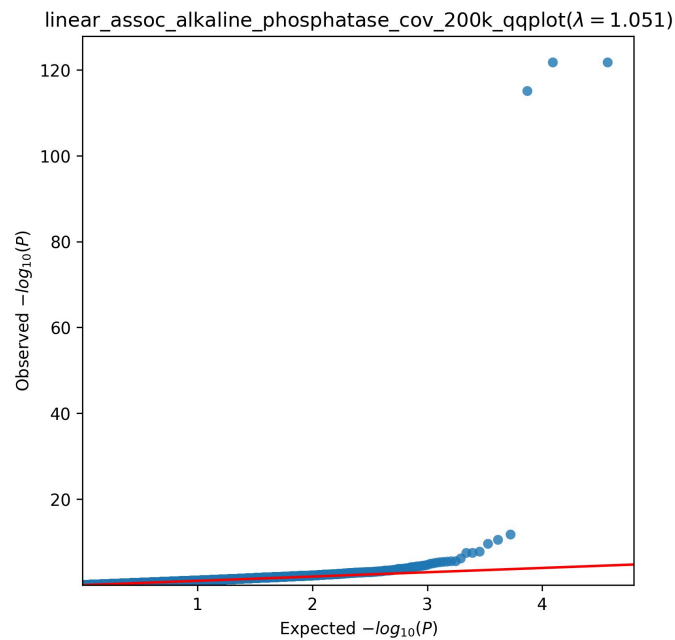
Alkaline phosphatase

We performed the analysis on the samples with ALP measurements as phenotype(quantitative).

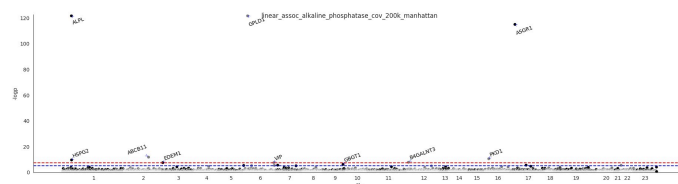
Association analysis

We used linear regression for the analysis and age, sex, BMI and PC1-4 were used as covaraites. The association analysis highlighted ALPL, GPLD1 and ASGR1 as significant genes, all of which are known to be associated with alkaline phosphatase. QQ-plot and Manhattan plot are presented below.

The QQ-plot:



The Manhattan plot:

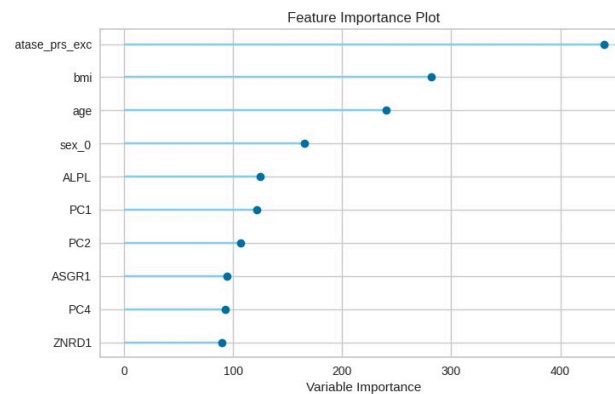


Regression model

For the prediction model, we used alkaline phosphatase measurements as target. For feature selection, For features we used 45 selected genes as features (45 genes) + BMI + age + sex + PC1-4. For the PRS and combined models we used the following PRS (PGS000670). The final prediction models was generated using gradient boosting regression, evaluation metric are shown in the table below.

	Gene-based model	PRS model	Combined model
R ²	0.084	0.255	0.281
RMSE	24.7	22.3	21.9

Feature importance plot for combined model:



Other phenotypes

Association analysis

biomarker	Top 3 significantly associated genes
Alanine aminotransferase	GPT, THRA, ACVR2B, More
Albumin	FCGRT, ALB, IQGAP2, More
Alkaline phosphatase	ALPL, GPLD1, ASGR1, More
Apolipoprotein A	ABCA1, LIPG, LCAT, More
Apolipoprotein B*	PCSK9, LDLR, NKPD, More
Aspartate aminotransferase	GOT1, GABRA5, THRA, More
Cholesterol*	PCSK9, LDLR, ABCA1, More
C reactive protein	CRP, PTGES3L, SLN, More
Creatinine (in serum)	NAA20, PRAMEF19, CLIC4, More
Gamma glutamyltransferase	GGT1, CCL1, RORC
Glucose	G6PC2, GCK, DYNLL1
Glycated Haemoglobin (HbA1C)	HBB, PIEZO1, GCK
Lipoprotein A	LPA, PLG, MRPL18
Triglycerides	APOA5I, APOC3, PLA2G12A

- values adjusted for statin

Note

summary statistics for biomarkers association analysis will be added soon.

Prediction models

Here we show a table of other phenotypes that we analyzed. For each phenotype we include the number of genes considered in the models as well as the R^2 of the gene-based model, PRS model and combined model.

	Number of genes	Gene-based model	PRS model	Comb
apolipoprotein a	6	0.227	0.413	0.403
apolipoprotein b*	5	0.059	0.267	0.269
aspartate aminotransferase	57	0.039	0.124	0.128
Cholesterol*	6	0.088	0.229	0.236
Creatinine	128	0.228	0.454	0.448
Hba1c	13	0.100	0.242	0.247
lipoprotein a	3	0.004	0.582	0.603
Triglyceride	5	0.143	0.316	0.315
urea	2	0.074	0.173	0.179
Urate	4	0.396	0.521	0.534

- values adjusted for statin

Methods Comparisons

We have compared GenRisk scores association test results with different burden test methods. In this example, we use LDL measurements (adjusted for statin) as phenotype and $\approx 160,000$ samples from UKbiobank.

We compared the results with SKATO, using the same input as GenRisk for LDL phenotype, and found that SKATO has detected many genes (58 genes after p-value adjustment) as significant, including PCSK9 and LDLR. However, the lambda of the p-values is inflated (1.325), as opposed to GenRisk which had a lambda of 1.056, which means there is a risk of having false-positives.

We, also performed two burden tests from rvtest (<https://github.com/zhanxw/rvtests>), CMC and Zeggini, and no significant genes were detected with CMC analysis, while Zeggini was able to detect PCSK9 but not LDLR.

Burden tests usually use genotypes only to score the genes, and can sometimes use filters like functional annotations and allele frequency, but this is just filtering (no values) and it has to be done as a pre-step before the actual analysis. An example of that is Genebass, where they applied SKATO on three different sets (Loss of function, missense and synonymous), and the results for LDL presented PCSK9 and LDLR on along with other genes, however the lambda for the p-values is inflated (e.g the lambda for LDL direct across all burden sets for the SKATO is 1.18 and for Burden test is 1.16), which means there is a risk of having false positives. https://genebass.org/gene/undefined/phenotype/continuous-30780-both_sexes-irnt?burdenSet=pLoF&resultIndex=gene-manhattan&resultLayout=full Another example is astrazeneca phewas portal (azphewas), here they have multiple models that filter for frequency, categorical annotations and/or deleteriousness score threshold. <https://azphewas.com/phenotypeView/f87604bb-7293-44e8-8e29-bf58d9872841/4b20a1ff-bded-4f1e-8301-f2922f0b8499/glr>

Computation information

It should be noted that the aim of our work is to provide a novel framework more comprehensive in terms of genetic risk assessment and currently is not yet optimized for computational performance. For all the computation below, we use a “standard” workstation (RAM=64GB with 6 CPU dual core).

Gene-based scoring and analysis

In the following table we show the computation time for the gene-core computation and association analysis with linear regression of the biggest chromosome (1,727,756 variants, MAF filtering = <1%) which includes also the higher number of genes (1,972 genes) by considering different numbers of individuals.

	1K samples	10K samples	100K samples
gene-scoring (in mins)	22	25	48
Find-association, linear regression (in sec)	8	28	134

While for prediction models the complete input matrix (i.e., samples and genes plus covariates) should be loaded in RAM, for gene-scoring we use the efficient score function implemented in PLINK v2 (). For gene-association GenRisk the memory usage depends on the size of the input matrix, the larger the matrix the more memory it uses.

	1K samples	10K samples	100K samples
Mem (in Gb)	3.1	3.4	9.4

Prediction models generation

GenRisk has “per-se” no limit in the number of features that can be used. However, there could be computational issues according to the dimensionality of the input, that is samples and features (genes, covariates, etc..). The tables below present the total run time (in seconds) and maximum memory usage (in GB) given different sample sizes with increasing number of features. Please note that it might be wise to run big data size (e.g 100K x 1000feats) using an HPC infrastructure. Another point to consider is that the time and memory usage also depends on the models included in the analysis and the best model fine-tuning and finalization. Some models, such as gradient boosting, might take more time than simpler models, like linear or lasso regression, to be finalized.

Total run time of prediction model generation in seconds

	1K samples	10K samples	100K samples
10 feats	14	19	1690
100 feats	24	678	41649
1000 feats	143	1034	432000(\approx 5days)

Maximum memory used in GB

	1K samples	10K samples	100K samples
10 feats	2.81	2.93	2.97
100 feats	2.93	2.95	3.29
1000 feats	3.51	3.82	8.29

Feature Selection

In general in the context of prediction models for big datasets we would suggest a feature selection using the “association” module and then generate prediction models. This is also in line with the expected genetic architecture of the majority of the traits in which only a small proportion of the genes plays a pivotal role. If instead we have a really highly polygenic phenotype the computation of genome-wide polygenic risk score (PRS) is probably the most appropriate approach, as PRS is a value per individual only a vector of scores would be generated and therefore the computational burden is limited.

3.2 Publication 2 - Gene-based burden scores identify rare variant associations for 28 blood biomarkers

In this paper we investigate the contribution of rare variants to the genetic landscape of 28 blood biomarkers from the UK Biobank cohort. Gene-based scores were calculated and association analyses were performed to detect significant genes in each biomarker. Furthermore, prediction models were generated with the gene-based scores and polygenic risk scores to explore their contributions to the genetic risk prediction. The identification of genes contributing to the blood biomarkers confirm that rare-variants play an important role in their genetic landscape. However, common variants might be a more informative method for predicting the genetic risk at a population level. Information about supplementary materials can be found in subsection 3.2.1 (Publication 2 - Appendix A). As the first author, I was directly involved in the planning of the work in this paper. I collected all the data needed and performed most the analyses, evaluation and interpretation of the results.

RESEARCH

Open Access



Gene-based burden scores identify rare variant associations for 28 blood biomarkers

Rana Aldisi^{1*}, Emadeldin Hassanin^{1,2}, Sugirthan Sivalingam^{1,3,4}, Andreas Bunes^{1,3,4}, Hannah Klinkhammer^{1,4}, Andreas Mayr⁴, Holger Fröhlich^{5,6}, Peter Krawitz¹ and Carlo Maj^{1,7}

Abstract

Background A relevant part of the genetic architecture of complex traits is still unknown; despite the discovery of many disease-associated common variants. Polygenic risk score (PRS) models are based on the evaluation of the additive effects attributable to common variants and have been successfully implemented to assess the genetic susceptibility for many phenotypes. In contrast, burden tests are often used to identify an enrichment of rare deleterious variants in specific genes. Both kinds of genetic contributions are typically analyzed independently. Many studies suggest that complex phenotypes are influenced by both low effect common variants and high effect rare deleterious variants. The aim of this paper is to integrate the effect of both common and rare functional variants for a more comprehensive genetic risk modeling.

Methods We developed a framework combining gene-based scores based on the enrichment of rare functionally relevant variants with genome-wide PRS based on common variants for association analysis and prediction models. We applied our framework on UK Biobank dataset with genotyping and exome data and considered 28 blood biomarkers levels as target phenotypes. For each biomarker, an association analysis was performed on full cohort using gene-based scores (GBS). The cohort was then split into 3 subsets for PRS construction and feature selection, predictive model training, and independent evaluation, respectively. Prediction models were generated including either PRS, GBS or both (combined).

Results Association analyses of the cohort were able to detect significant genes that were previously known to be associated with different biomarkers. Interestingly, the analyses also revealed heterogeneous effect sizes and directionality highlighting the complexity of the blood biomarkers regulation. However, the combined models for many biomarkers show little or no improvement in prediction accuracy compared to the PRS models.

Conclusion This study shows that rare variants play an important role in the genetic architecture of complex multifactorial traits such as blood biomarkers. However, while rare deleterious variants play a strong role at an individual level, our results indicate that classical common variant based PRS might be more informative to predict the genetic susceptibility at the population level.

Keywords Gene associations, Blood biomarkers, Genetic prediction, Rare variants, PRS, Complex phenotypes

*Correspondence:

Rana Aldisi

s0raaldi@uni-bonn.de

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

The genetic architecture of complex phenotypes has been studied extensively for over a century; however, a relevant part of the genetics still elude us. That is because, essentially, many factors are involved in the development of such traits, both biological and environmental, which makes it harder to discover causative effects for any complex phenotype or disease [1]. Genome-wide association studies (GWAS) investigate the associations of low-effect single nucleotide polymorphisms (SNPs) with specific phenotypes. For the last decade, GWAS have been used to identify many common variants that are associated with diseases and other phenotypes such as cancer [2], autism [3] and cholesterol [4]. About 90% of the variants identified by GWAS are located in the non-coding regions of the genome. This gives insight to the mechanisms behind development and progress of complex phenotypes by exploring regulatory elements that could have an effect on disease related genes [5]. However, the narrow sense of heritability estimated from the GWAS, also known as $SNP-h^2$, is typically lower than the broad sense of heritability H^2 estimate from twins and family studies, this is known as the missing heritability [6]. Different hypotheses have been suggested to resolve the difference between observed and measured heritability, such as non-linear effects, epigenetics and rare variants [6]. It has also been hypothesized that family studies or twin studies might have overestimated the heritability and that the shared environment plays a significant role in these traits [7]. On the other hand, many studies suggest that more genetic variations need to be included in the analysis of complex traits to account for the unexplained heritability, such as small to moderate effect low-frequency (MAF 1–5%) variants, and potentially highly damaging rare variants (MAF < 1%) [8]. In fact, it has been observed that rare variants contribute to the genetic landscape of complex phenotypes such as inflammatory bowel disease [9], hypertension [10] and autism [11].

Common and rare variants are typically analyzed independently. Common variants' effects on a certain phenotype are analyzed using polygenic risk scores (PRS), these scores are usually derived from large-scale GWAS and are used to assess an individual's genetic liability for a certain trait or disease [12]. However, current PRSs explain only a small part of the heritability of complex traits [13]. On the other hand, multiple methods have been developed to find phenotype associations with rare variants. A widely known category is burden test, which collapses all information in a genetic region (e.g. gene) into one genetic burden score that can be used for association analysis. The association is then analyzed between the burden score and a certain phenotype. However, burden tests assume that all rare variants are causal and

have the same directional effect on the trait tested [14]. Another class of methods was developed to avoid these limitations, which is known as the variance-component tests. These tests analyze associations by looking at joint genetic effect for variants in a genetic region. For example, sequence kernel association test (SKAT), aggregates score statistics of multiple variants then evaluates the distribution [15]. While this class has dealt with the limitations of burden tests, it might not perform well when a large proportion of the variants have strong effects in the same direction [14]. For this purpose, methods combining burden tests and variance-component tests have been proposed. One of these methods is SKAT-O, an extension of SKAT which can incorporate both common and rare variants in the analysis [16]. While all these different approaches have their advantages, one of their disadvantages is that they do not provide individual-level data, therefore, other methods based on functional annotations and frequency weight have been developed, such as Genepy [17] and GenRisk [18]. These approaches are more general and allow gene-based scores at individuals levels to be derived which can be used subsequently for multiple analyses.

For both common and rare variants, well-established methods exist to perform genotype-phenotype association and prediction analysis; however, their combined contributions have not been fully studied. Our paper aims to analyze the contribution of both rare and common variants to complex phenotypes. We achieve this by integrating gene-based scores for rare variants and PRS for common variants in genetic risk modeling.

Results

We used gene-based scores, calculated based on the burden of rare functional variants and allele frequency, to analyze gene associations with 28 quantitative biomarkers. We further integrated the gene-based scores with the PRS models, aiming to enhance the risk prediction.

Identification of phenotype-associated genes

To identify genes associated with different biomarkers, we performed association analysis, using linear regression, on the UK biobank cohort with 28 blood biomarkers extracted as phenotypes. Furthermore, we calculated the effect size (z-score) of each gene on each biomarker phenotype using the beta coefficient and standard error extracted from the association analysis. Figure 1 displays the distribution of the effect sizes of genes with P -value < 0.05 after Bonferroni correction for each phenotype with highlight on the highest and lowest effect size genes, with effect sizes ranging between -49.6 (ALPL in alkaline phosphatase) and 23.4 (LDLR in LDL direct

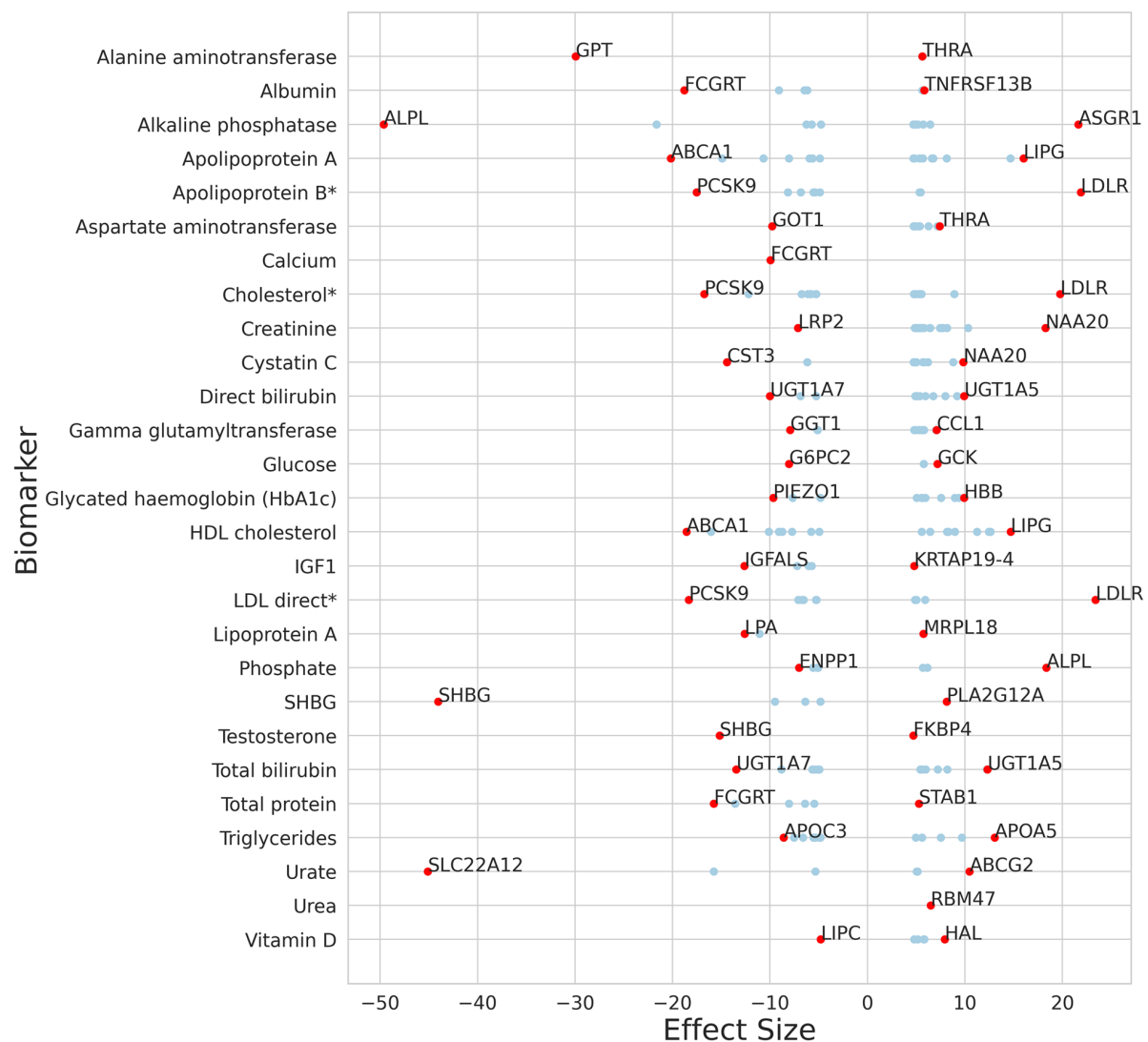


Fig. 1 Distribution of effect sizes of genes with P -value < 0.05 after Bonferroni correction, the highest and lowest genes' effect sizes are labeled for every biomarker

measurement). The number of genes with positive and negative effects for each biomarker is shown in Table 1.

Rare and common variants integrated risk prediction models

In order to assess the contribution of rare and common variants on complex phenotypes, we generated prediction models for each biomarker. These models were generated using GenRisk pipeline, which evaluates different regression models and outputs the model with the best performance as a final output, we then calculated the R^2 for each model using an independent testing set. Four different models for each biomarker were generated:

based on polygenic risk scores for common variant effect (PRS model); based on selected gene-based scores for rare variant effect (GBS model); combining both rare and common variant effects (PRS+GBS combined model); a only covariates-model (in order to assess the incremental performance due to the genetic factors). Table 2 presents the R^2 for the covariates models and the incremental R^2 for all other models in comparison.

Discussion

In this study, we evaluated the association of rare genetic variants with 28 blood biomarkers. In addition, we explore the genetic contribution of these variants to the

Table 1 Number of significantly associated genes with negative and positive effect sizes

Biomarker	Negative effect	Positive effect
Alanine aminotransferase	1	1
Albumin	4	2
Alkaline phosphatase	5	6
Apolipoprotein A	7	9
Apolipoprotein B ^a	6	3
Aspartate aminotransferase	1	8
Cholesterol ^a	8	7
Creatinine	1	27
Cystatin C	2	9
Direct bilirubin	3	10
Gamma glutamyltransferase	3	9
Glucose	1	2
Glycated haemoglobin (HbA1c)	3	8
HDL cholesterol	8	9
IGF1	4	1
LDL direct ^a	7	4
Lipoprotein A	2	1
Phosphate	4	3
SHBG	4	1
Testosterone	1	1
Total bilirubin	7	7
Total protein	5	1
Triglycerides	8	5
Urate	3	3
Vitamin D	1	5

^a Values adjusted for statins

regulation of the biomarkers levels using samples from the UK Biobank. The association analysis, based on gene-scores derived from the burden of rare functional variants, revealed several interesting gene candidates associated with different blood biomarkers, showing both positive (increasing) and negative (decreasing) effect sizes. Some of these candidate genes have clear known associations with their respective biomarker; for example, ALPL gene was identified in association with alkaline phosphatase biomarker levels, and SHBG gene was associated with both sex hormone binding globulin (SHBG) and testosterone biomarkers' levels. In addition, the negative effect direction of those associations indicates that the presence of rare functional, possibly damaging, variants, as measured by the gene-based scores, decreases the biomarkers' levels. This is consistent with the fact that ALPL and SHBG are the protein-coding genes for the alkaline phosphatase and SHBG biomarkers, respectively. Consequently, the presence of damaging variants in these genes could lead to a decrease in the production of their corresponding biomarkers. Additionally, since

SHBG regulates testosterone levels in the body, a reduction in SHBG levels may also result in a reduction of testosterone levels [19].

Another clear example for rare variant associations is LDL (low-density lipoprotein), which showed association and positive effect direction with LDLR and negative effect direction with PCSK9. In this case, damaging mutations in LDLR, the gene for the LDL receptor, result in an increase in LDL levels in plasma. This finding is not surprising, as it has been previously suggested that mutations in LDLR are often responsible for familial hypercholesterolemia [20]. Instead, PCSK9 is a regulatory protein that degrades LDLR and thus leads to an increase in LDL plasma levels. In fact, PCSK9 inhibitors have been used as a treatment for hypercholesterolemia [21].

To confirm and validate our result, we also compared our findings with two different approaches that try to find gene-phenotype associations using rare variants and are performed on UK biobank samples, genebass [22] and AstraZeneca PheWAS [23]. Genebass uses SAIGE-GENE [24] to perform gene-based burden test and SKAT-O, while AstraZeneca PheWAS analysis was performed using Fisher's exact test on different models each with their own variant functional and allele frequency filtering criteria. In general, the different methods share many similar associations, however, our method has shown to have less inflated lambda in comparison to genebass. Typically, the lambda values are expected to be near 1, a lambda lower than 1 (deflation) could mean under-powered analysis and a lambda higher than 1 (inflation) could mean high false positive rate. Table 3 presents the lambdas as calculated from the three different approaches, since genebass and Astrazeneca PheWAS used different models to find associations, the average of these models is reported. Lambdas for all models' values can be found in the supplementary material (Table S2).

All approaches identified genes that are previously known to be associated with the respective biomarker (P -value < 0.05 after Bonferroni correction), for example PCSK9, LDLR, NPC1L1 and ABCG5 association with LDL levels [25–27]. However, our approach was able to identify potential novel associations that were not found with the other methods, such as, SNX8 for LDL and cholesterol, which is a part of the sorting nexin family and have been previously associated with the distribution of neuronal cholesterol [28]. Another example of shared association among all approaches is the association of GOT1, also known as AST1, with aspartate aminotransferase (AST), which is the gene encoding AST. GenRisk further identified THRA, also known as thyroid hormone receptor alpha. AST is a liver enzyme that is used as a biomarker to indicate liver damage or disease and in fact, the liver plays an important role in the activation,

Table 2 The R^2 of prediction models for blood biomarkers, with calculated incremental R^2 values between covariates only model and the rest of the models

Biomarker	Gene predictors	Covariates Model R^2	Incremental R^2		
			Genes	PRS	Combined
Alanine aminotransferase	4	0.137	0.003	0.011	0.014
Albumin	5	0.059	0.005	0.027	0.032
Alkaline Phosphatase	8	0.071	0.026	0.088	0.103
Apolipoprotein A	11	0.208	0.009	0.075	0.083
Apolipoprotein B ^a	5	0.088	0.007	0.157	0.162
Aspartate Aminotransferase	10	0.040	0.000	0.009	0.009
Calcium	2	0.028	0.002	0.017	0.018
Cholesterol ^a	6	0.089	0.006	0.096	0.099
C-reactive protein	5	0.066	0.004	0.009	0.011
Creatinine	41	0.248	-0.006	0.011	0.005
Cystatin C	11	0.177	-0.001	0.043	0.043
Direct bilirubin	14	0.045	0.011	0.272	0.272
Gamma glutamyltransferase	11	0.053	0.001	0.015	0.015
Glucose	8	0.030	0.001	0.003	0.003
Glycated haemoglobin (HbA1c)	16	0.098	0.001	0.020	0.022
HDL cholesterol	14	0.274	0.011	0.113	0.120
IGF1	5	0.091	0.003	0.067	0.070
LDL direct ^a	5	0.077	0.006	0.109	0.113
Lipoprotein A	3	0.000	0.003	0.567	0.591
Phosphate	3	0.067	0.003	0.020	0.023
SHBG	5	0.309	0.017	0.053	0.065
Testosterone	1	0.828	0.001	0.006	0.008
Total bilirubin	11	0.064	0.012	0.399	0.400
Total protein	4	0.003	0.005	0.039	0.042
Triglycerides	7	0.139	0.003	0.058	0.061
Urate	4	0.387	0.013	0.065	0.077
Urea	2	0.070	0.000	0.009	0.010
Vitamin D	2	0.040	0.001	0.015	0.015

^a Values adjusted for statins

metabolism and transport of thyroid hormone, while thyroid hormones are said to affect hepatic cells metabolism [29]. Notably, THRA was also identified by GenRisk as significant, for alanine aminotransferase, another liver biomarker. Figures 2, 3 and 4 display the association analysis results along with venn diagram representing the number of significant associations identified from each approach mentioned above for LDL, aspartate aminotransferase and alanine aminotransferase, respectively. Similar figures for the rest of the biomarkers are provided in the supplementary material (Figs. S2–S26). The summary statistics for the association analysis performed by GenRisk for each biomarker are also provided in the supplementary material (Tables S3–S31).

In addition, in order to assess the contribution of rare-variants in the 28 blood biomarkers, we compared risk prediction models using four different modalities

(see [Methods](#) for details). Our prediction model results suggest that the effect of rare variants on complex phenotypes differs depending on the distinct genetic architecture of the phenotypes. Furthermore, even though most of the biomarkers predictions show improvements when combining rare (GBS) and common (PRS) variants, these improvements are marginal in many cases which suggest that the added predictive value of rare variants in risk prediction is limited. Interestingly, gradient boosting regressor was selected by our pipeline as best performing model for most biomarkers. In gradient boosting machines, weak performing models, e.g decision trees, are combined together to generate a more powerful predictive model [30]. In fact, it has been shown that gradient boosting and other machine learning models perform better than traditional linear models in complex phenotypes when non-additive effects might be involved [31].

Table 3 The lambdas of the three different approaches, averaged in case of multiple values. Full and detailed table with all values can be found in [Supplementary material](#)

Biomarker	GenRisk	GeneBass Burden Average	GeneBass SKATO Average	AstraZeneca PheWAS Average
Alanine aminotransferase	1.016	1.139 ± 0.081	1.1967 ± 0.278	1.046 ± 0.016
Albumin	1.069	1.136 ± 0.132	1.231 ± 0.243	1.050 ± 0.018
Alkaline phosphatase	1.078	1.322 ± 0.259	1.739 ± 1.130	1.084 ± 0.023
Apolipoprotein A	1.068	1.207 ± 0.170	1.340 ± 0.317	1.070 ± 0.020
Apolipoprotein B	1.105	1.149 ± 0.094	1.247 ± 0.289	1.053 ± 0.018
Aspartate aminotransferase	0.951	1.174 ± 0.104	1.253 ± 0.285	1.064 ± 0.027
Calcium	1.063	1.099 ± 0.045	1.162 ± 0.140	1.053 ± 0.020
Cholesterol	1.085	1.158 ± 0.117	1.199 ± 0.266	1.050 ± 0.014
C-reactive protein	0.995	1.228 ± 0.178	1.505 ± 0.786	1.082 ± 0.0201
Creatinine	0.861	1.201 ± 0.158	1.328 ± 0.418	1.097 ± 0.027
Cystatin C	0.995	1.221 ± 0.173	1.376 ± 0.371	1.093 ± 0.030
Direct bilirubin	0.993	1.168 ± 0.146	1.411 ± 0.613	1.036 ± 0.001
Gamma glutamyltransferase	0.965	1.207 ± 0.078	1.384 ± 0.289	1.065 ± 0.030
Glucose	0.998	1.081 ± 0.081	1.082 ± 0.111	1.019 ± 0.013
Glycated haemoglobin HbA1c	1.018	1.224 ± 0.125	1.391 ± 0.387	1.090 ± 0.026
HDL Cholesterol	1.076	1.231 ± 0.175	1.417 ± 0.474	1.075 ± 0.026
IGF1	1.084	1.212 ± 0.145	1.352 ± 0.396	1.096 ± 0.019
LDL direct	1.092	1.132 ± 0.119	1.179 ± 0.245	1.039 ± 0.016
Lipoprotein A	0.992	1.156 ± 0.152	1.354 ± 0.534	1.020 ± 0.008
Phosphate	1.065	1.041 ± 0.028	0.976 ± 0.060	1.054 ± 0.020
SHBG	1.07	1.194 ± 0.076	1.353 ± 0.336	1.065 ± 0.025
Testosterone	1.005	1.088 ± 0.105	1.072 ± 0.205	1.016 ± 0.014
Total bilirubin	1.028	1.264 ± 0.193	1.648 ± 0.911	1.030 ± 0.013
Total protein	1.059	1.194 ± 0.183	1.286 ± 0.323	1.078 ± 0.021
Triglycerides	1.066	1.197 ± 0.187	1.279 ± 0.362	1.071 ± 0.011
Urate	1.076	1.227 ± 0.054	1.429 ± 0.335	1.058 ± 0.018
Urea	1.036	1.116 ± 0.091	1.157 ± 0.226	1.049 ± 0.012
Vitamin D	1.034	1.089 ± 0.016	1.133 ± 0.135	1.049 ± 0.019

It is noteworthy to mention that some risk prediction models were mostly influenced by other factors, like sex for testosterone and creatinine, as seen in Fig. 5, which was identified as the variable with the highest influence in these models with the other features playing only a minor role in the prediction. This is to be expected, since testosterone is a sex-specific hormone and creatinine levels vary depending on the individual's size and muscle mass, which is usually higher in men [32]. The true vs. predicted value plot and the top features figures for all the biomarkers' models can be found in the supplementary materials (Figs. S27–S54).

Conclusion

In this study, we investigate the contribution of rare functional variants in blood biomarkers. We performed association analysis on gene-based burden scores and built genetic risk models using rare and common variant

effects. The results suggest that gene-based score is a powerful instrument to identify gene-phenotype associations between rare-variants and complex phenotypes. While some of the associations were replicated by other methods, our tool has the advantage of producing individual-level scores that can be used for multiple subsequent analyses. Although gene-based scores have proven to be useful on the individual-level, traditional PRS provides more information for risk prediction purposes on the population-level scale. It is important to mention that these results are limited to the effects of rare and common variants at gene-based level. Even though we included non-linear models in the analysis to potentially detect gene-gene interactions, they cannot capture effects that happen at variant level. Furthermore, other potential factors influencing the genetic susceptibility (i.e., epigenetics, gene-environment) are not considered in our current work.

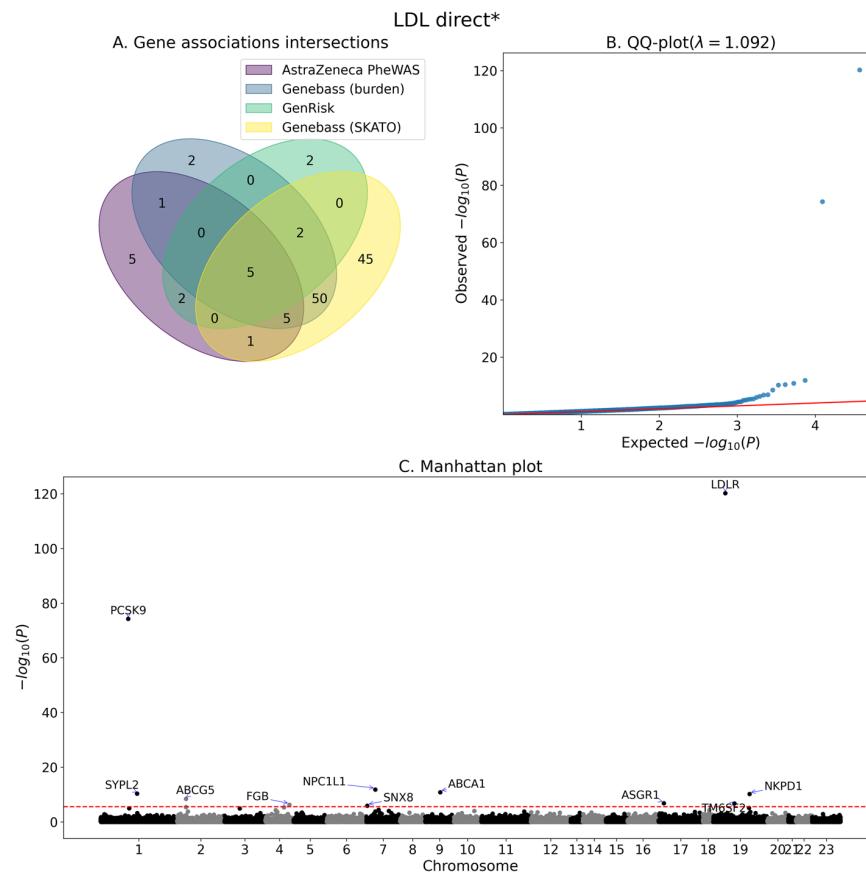


Fig. 2 Association analysis summary for LDL direct*. **A** Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS (all models) and genebass (Burden and SKATO). **B** QQ-plot of the P -values of GenRisk pipeline results. **C** Manhattan plot of GenRisk pipeline results. *statin adjusted values

Methods

Cohort and data processing

All analyses were performed on the UK biobank cohort, which is a large-scale population-based biomedical database that contains data for half a million participants. Data include questionnaires, biomarkers, imaging and genetic data. For our analysis, we used imputed genotype data, whole exome sequencing data, biometric data (age, sex, BMI) and all blood biomarker measurements except for rheumatoid factor and estradiol, which were excluded because of low sample size. The UK biobank field identifiers used can be found in supplementary material (Table S1). Variants were annotated with genes using NCBI's gene and reference sequences [33], gnomad allele frequency and CADD v1.6 raw scores [34]. We filtered the cohort to include participants with white British ancestry that have whole exome sequencing data and genotype data, resulting in $n=145,464$ samples. For individuals using the cholesterol lowering statins as medication, cholesterol, LDL and apolipoprotein B levels were

adjusted by using previously estimated factors of 0.684, 0.749, and 0.719, respectively [35]. For risk prediction modeling, the cohort was split into three subsets: 60% ($n=87,278$) for constructing the PRS and feature selection, 30% ($n=43,639$) for training the prediction models, and 10% ($n=14,547$) for model testing. The number of samples per phenotype varied depending on the availability of measurements. Distribution and number of samples per biomarker can be found in the supplementary material (Fig. S1).

Polygenic risk score (PRS)

To generate the PRS for each biomarker, we applied snpnet pipeline [36] on the the imputed genotyping samples of the construction dataset. This pipeline uses batch screening iterative lasso framework to select effect variants and generate polygenic score which can be used to calculate PRS for a cohort. We used the default parameters defined in snpnet pipeline for polgenic score derivation and excluded SNPs with $MAF < 0.01$. After polygenic

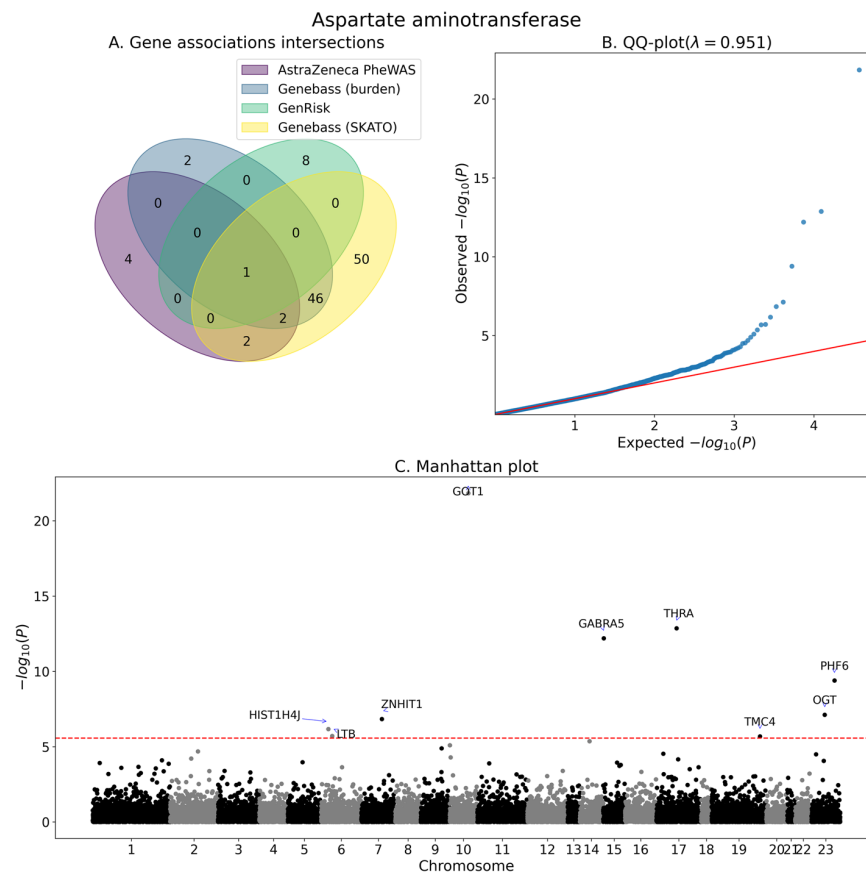


Fig. 3 Association analysis summary for aspartate aminotransferase. **A** Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS (all models) and geneBass (Burden and SKATO). **B** QQ-plot of the P -values of GenRisk pipeline results. **C** Manhattan plot of GenRisk pipeline results

score construction, we calculated the PRS for the remaining cohort to be included in the prediction model training and testing subsets.

Rare variants analysis

We used GenRisk, a python package that implements a gene-based scoring system, association analysis, risk scores calculations and machine learning models generation [18]. The gene-based scoring system depends on frequency and functional annotations, with up-weighting function for rare variants. Gene-based scores (GBS) were derived from whole exome data for all individuals in the cohort, using default settings (MAF threshold < 0.01, beta weighting function with parameters 1 and 25), and associations were assessed for the 28 biomarkers with quantitative values. For association analysis, linear regression was applied to the gene-based scores of the whole cohort with BMI, age, sex and the first four genetic principal components (PCs) as covariates. The number of PCs was chosen based on the variance explained in UK

biobank European cohort [37]. Manhattan and QQ plots were generated to visualize the results, and the lambda statistic, representing the inflation of P -values in comparison to the expected distribution of P , was also calculated. To account for multiple testing, Bonferroni correction was applied to adjust the P -values. Thus, the genome-wide significance threshold level was calculated based on the number of tested genes ($0.05/18556 = 2.69E-07$).

Feature selection

To reduce the numbers of input variables in prediction models, feature selection was applied on the GBS matrix to select genes that are associated with the respective biomarker. Association analysis was performed using linear regression with the same previously stated covariates on the GBS of the construction subset for each of the biomarker and genes with P -value < 0.05 after Bonferroni correction were selected as gene predictors. Number of gene predictors per biomarker can be found in Table 2.

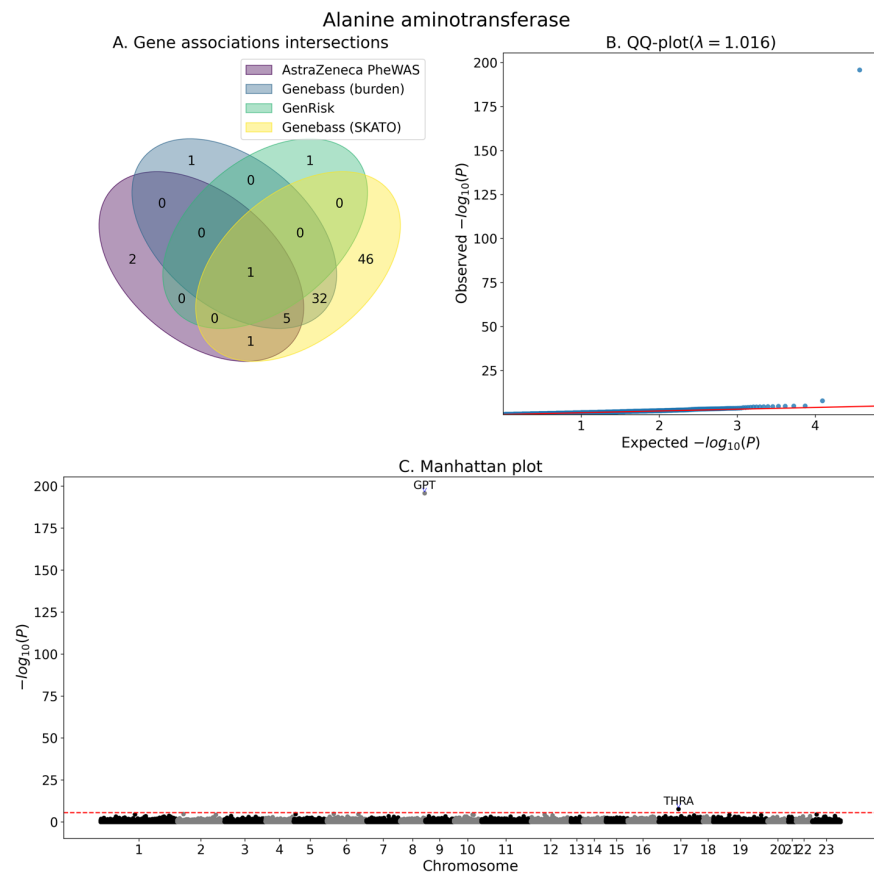


Fig. 4 Association analysis summary for alanine aminotransferase. **A** Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS (all models) and genebase (Burden and SKATO). **B** QQ-plot of the P -values of GenRisk pipeline results. **C** Manhattan plot of GenRisk pipeline results

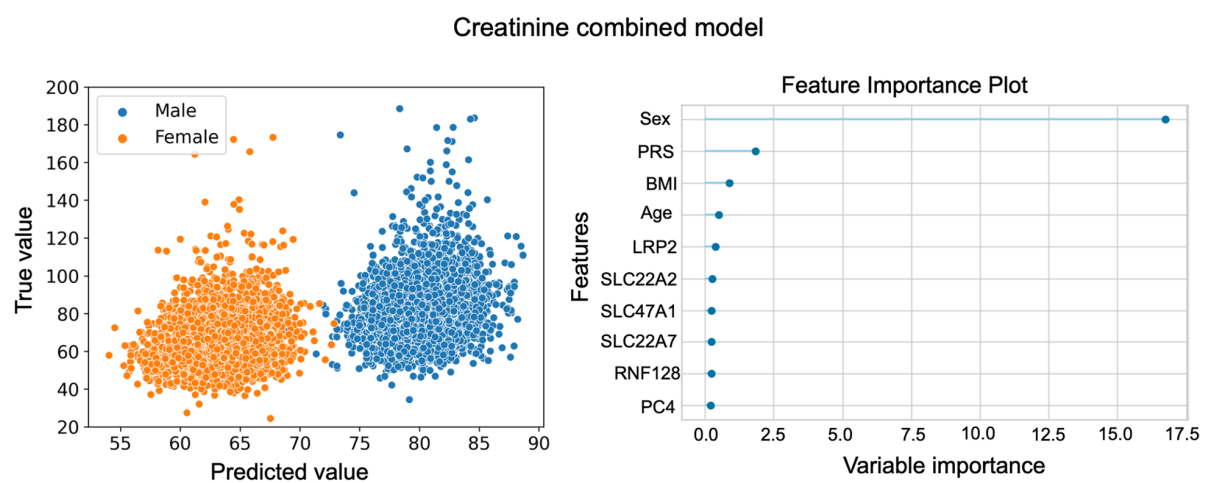


Fig. 5 True vs. Predicted value plot (left) and top 10 features (right) for creatinine combined model. Values that are a 3 standard deviations away from the mean were eliminated for a better visualization

Risk prediction modeling

For each biomarker, four different prediction models were generated using the machine learning model training subset.

- Covariates model: biomarker = sex + age + BMI + PC1 + PC2 + PC3 + PC4
- GBS model: biomarker = covariates + GBS
- PRS model: biomarker = covariates + PRS
- Combined model: biomarker = covariates + GBS + PRS

Our tool, GenRisk, uses PyCaret as underlying framework for prediction model generation. PyCaret is a python library that implements different machine learning models and can be used for training and testing, selecting, fine tuning and finalizing models¹. Different models (n=17) including linear, such as ridge, elastic net and lasso regression, and non-linear models, like gradient boosting and random forest regression, are tested. A list of all models can be found in the GenRisk documentation². For the GBS, only the gene predictors that were selected in the feature selection step for each biomarker were included. All features were normalized by calculating the z-score. The training step was performed on the training set, with the corresponding biomarker as target, using 10 fold cross-validation and the best performing model for each biomarker is then finalized considering the complete training cohort and applied on the independent test set.

Abbreviations

AST	Aspartate aminotransferase
GBS	Gene-based scores
GWAS	Genome-wide association studies
LDL	Low-density lipoprotein
MAF	Minor allele frequency
PRS	Polygenic risk scores
SHBG	Sex hormone binding globulin
SNPs	Single nucleotide polymorphisms
SKAT	Sequence kernel association test

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12863-023-01155-0>.

Additional file 1.

Additional file 2.

Acknowledgements

Not applicable.

Authors' contributions

C.M. and P.K. conceived the idea. H.K. and E.H. acquired, processed, managed and prepared data for further analysis. R.A. developed and designed the methodology. R.A., S.S. and A.B. performed all analyses, with support from H.K., R.A. and C.M. investigated and interpreted the results. C.M., P.K., A.M. and H.F. supervised the project and provided support and feedback throughout the work. R.A. and C.M. wrote the first version of the manuscript. P.K., A.M., H.F., E.H., S.S., A.B. and H.K. provided feedback, substantively revised the manuscript and contributed to the final version of the work.

Funding

Open Access funding enabled and organized by Projekt DEAL. R.A. Received funds from the German Research Foundation (www.dfg.de). Project ID: 428902522. The funding body played no roles in the design of the study and collection, analysis, and interpretation of data or in writing the manuscript.

Availability of data and materials

UK Biobank is a large-scale biomedical database and research resource. Data from UK Biobank (Genotyping data, exome data, and phenotypic data) are available upon application (<http://www.ukbiobank.ac.uk/about-biobank-uk/>). Restrictions apply to the availability of these data, which were used under license for the current study (Project ID: 81202).

Declarations

Ethics approval and consent to participate

Ethical approval for the UK Biobank study has been granted by the National Information Governance Board for Health and Social Care and the NHS North West Multicentre Research Ethics Committee (11/NW/0382). This approval means that researchers do not require separate ethical clearance and can operate under the RTB approval. Informed consent was obtained from all participants. More information about ethical approval can be found here: <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics>.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Institute of Genomic Statistic and Bioinformatics, University Hospital Bonn, Bonn, Germany. ²Luxembourg Center for Systems Biomedicine, University of Luxembourg, Esch-Sur-Alzette, Luxembourg. ³Core Unit for Bioinformatics Analysis, University Hospital Bonn, Bonn, Germany. ⁴Institute of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany. ⁵Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany. ⁶Bonn-Aachen International Center for IT (b-it), University of Bonn, Bonn, Germany. ⁷Centre for Human Genetics, University of Marburg, Marburg, Germany.

Received: 14 November 2022 Accepted: 28 August 2023

Published online: 04 September 2023

References

1. Hindorf LA, Gillanders EM, Manolio TA. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis*. 2011;32(7):945–54. <https://doi.org/10.1093/carcin/bgr056>.
2. Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer*. 2017;17(11):692–704. <https://doi.org/10.1038/nrc.2017.82>.
3. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet*. 2019;51(3):431–44. <https://doi.org/10.1038/s41588-019-0344-8>.
4. Ma L, Yang J, Runesha HB, Tanaka T, Ferrucci L, Bandinelli S, et al. Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham Heart Study data. *BMC Med Genet*. 2010;11(1). <https://doi.org/10.1186/1471-2350-11-55>.

¹ <https://pycaret.gitbook.io/docs/>

² <https://genrisk.readthedocs.io/en/latest/>

5. Lee PH, Lee C, Li X, Wee B, Dwivedi T, Daly M. Principles and methods of in-silico prioritization of non-coding regulatory variants. *Hum Genet*. 2017;137(1):15–30. <https://doi.org/10.1007/s00439-017-1861-0>.
6. Young AI. Solving the missing heritability problem. *PLoS Genet*. 2019;15(6):e1008222. <https://doi.org/10.1371/journal.pgen.1008222>.
7. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci*. 2012;109(4):1193–8. <https://doi.org/10.1073/pnas.1119675109>.
8. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol*. 2017;18(1). <https://doi.org/10.1186/s13059-017-1212-4>.
9. Venkataraman GR, Rivas MA. Rare and common variant discovery in complex disease: the IBD case study. *Hum Mol Genet*. 2019;28(R2):R162–9. <https://doi.org/10.1093/hmg/ddz189>.
10. Russo A, Gaetano CD, Cugliari G, Matullo G. Advances in the Genetics of Hypertension: The Effect of Rare Variants. *Int J Mol Sci*. 2018;19(3):688. <https://doi.org/10.3390/ijms19030688>.
11. Havdahl A, Niarchou M, Starnawska A, Uddin M, van der Merwe C, Warrier V. Genetic contributions to autism spectrum disorder. *Psychol Med*. 2021;1–14. <https://doi.org/10.1017/S0033291721000192>.
12. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219–24. <https://doi.org/10.1038/s41588-018-0183-z>.
13. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759–72. <https://doi.org/10.1038/s41596-020-0353-1>.
14. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet*. 2014;95(1):5–23. <https://doi.org/10.1016/j.ajhg.2014.06.009>.
15. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am J Hum Genet*. 2011;89(1):82–93. <https://doi.org/10.1016/j.ajhg.2011.05.029>.
16. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *Am J Hum Genet*. 2012;91(2):224–37. <https://doi.org/10.1016/j.ajhg.2012.06.007>.
17. Mossotto E, Ashton JJ, O'Gorman L, Pengelly RJ, Beattie RM, MacArthur BD, et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinformatics*. 2019;20(1). <https://doi.org/10.1186/s12859-019-2877-3>.
18. Aldisi R, Hassanin E, Sivalingam S, Buness A, Klinkhammer H, Mayr A, et al. GenRisk: a tool for comprehensive genetic risk modeling. *Bioinformatics*. 2022. <https://doi.org/10.1093/bioinformatics/btac152>.
19. Winters SJ. SHBG and total testosterone levels in men with adult onset hypogonadism: what are we overlooking? *Clin Diabetes Endocrinol*. 2020;6(1). <https://doi.org/10.1186/s40842-020-00106-3>.
20. Cuchel M, Bruckert E, Ginsberg HN, Raal FJ, Santos RD, Hegele RA, et al. Homozygous familial hypercholesterolaemia: new insights and guidance for clinicians to improve detection and clinical management. A position paper from the Consensus Panel on Familial Hypercholesterolaemia of the European Atherosclerosis Society. *Eur Heart J*. 2014;35(32):2146–57. <https://doi.org/10.1093/eurheartj/ehu274>.
21. Reiss AB, Shah N, Muhieddine D, Zhen J, Yudkevich J, Kasselmann LJ, et al. PCSK9 in cholesterol metabolism: from bench to bedside. *Clin Sci*. 2018;132(11):1135–53. <https://doi.org/10.1042/cs20180190>.
22. Karczewski KJ, Solomonson M, Chao KR, Goodrich JK, Tiao G, Lu W, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics*. 2022;100168. <https://doi.org/10.1016/j.xgen.2022.100168>.
23. Wang Q, Dhindsa RS, Carrs K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*. 2021;597(7877):527–32. <https://doi.org/10.1038/s41586-021-03855-y>.
24. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Taliun SAG, et al. Scalable generalized linear mixed model tests for region-based association tests in large biobanks and cohorts. *Nat Genet*. 2020;52(6):634–9. <https://doi.org/10.1038/s41588-020-0621-6>.
25. Sabatine MS, Giugliano RP, Keech AC, Honarpour N, Wiviott SD, Murphy SA, et al. Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *N Engl J Med*. 2017;376(18):1713–22. <https://doi.org/10.1056/nejmoa1615664>.
26. Liao J, Yang L, Zhou L, Zhao H, Qi X, Cui Y, et al. The NPC1L1 Gene Exerts a Notable Impact on the Reduction of Low-Density Lipoprotein Cholesterol in Response to Hyeztimibe: A Factorial-Designed Clinical Trial. *Front Pharmacol*. 2022;13. <https://doi.org/10.3389/fphar.2022.755469>.
27. Zein AA, Kaur R, Hussein TOK, Graf GA, Lee JY. ABCG5/G8: a structural view to pathophysiology of the hepatobiliary cholesterol secretion. *Biochem Soc Trans*. 2019;47(5):1259–68. <https://doi.org/10.1042/bst20190130>.
28. Yang J, Villar VAM, Rozyyev S, Jose PA, Zeng C. The emerging role of sorting nexins in cardiovascular diseases. *Clin Sci*. 2019;133(5):723–37. <https://doi.org/10.1042/cs20190034>.
29. Piantanida E, Ippolito S, Gallo D, Masiello E, Premoli P, Cusini C, et al. The interplay between thyroid and liver: implications for clinical practice. *J Endocrinol Investig*. 2020;43(7):885–99. <https://doi.org/10.1007/s40618-020-01208-6>.
30. Hastie T, Friedman J, Tibshirani R. Boosting and Additive Trees. In: *The Elements of Statistical Learning*. New York: Springer New York; 2001. p. 299–345. <https://doi.org/10.1007/978-0-387-21606-5>.
31. Perez BC, Bink MCAM, Svenson KL, Churchill GA, Calus MPL. Prediction performance of linear models and gradient boosting machine on complex phenotypes in outbred mice. *G3 Genes Genomes Genet*. 2022;12(4). <https://doi.org/10.1093/g3journal/jkac039>.
32. Schott HC, Walldridge B, Bayly WM. Disorders of the Urinary System. In: *Equine Internal Medicine*. Philadelphia: Saunders; 2018. p. 888–990. <https://doi.org/10.1016/b978-0-323-44329-6.00014-0>.
33. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*. 2013;42(D1):D756–63. <https://doi.org/10.1093/nar/gkt114>.
34. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med*. 2021;13(1). <https://doi.org/10.1186/s13073-021-00835-9>.
35. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet*. 2021;53(2):185–94. <https://doi.org/10.1038/s41588-020-00757-z>.
36. Qian J, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLOS Genet*. 2020;16(10):e1009141. <https://doi.org/10.1371/journal.pgen.1009141>.
37. Constantinescu AE, Mitchell RE, Zheng J, Bull CJ, Timpson NJ, Amulic B, et al. A framework for research into continental ancestry groups of the UK Biobank. *Human Genomics*. 2022;16(1). <https://doi.org/10.1186/s40246-022-00380-5>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



3.2.1 Publication 2 - Appendix A

This appendix contains information about the supplementary materials for publication 2. The tables and figures can be downloaded directly from the paper: <https://doi.org/10.1186/s12863-023-01155-0>. Because of size issues, the tables cannot be included directly in this thesis.

Supplementary material

Table of figures

Figure S1 Distribution of samples across biomarkers.....	3
Figure S2 Association analysis summary for albumin.....	4
Figure S3 Association analysis summary for alkaline phosphatase.....	5
Figure S4 Association analysis summary for apolipoprotein A.....	6
Figure S5 Association analysis summary for apolipoprotein B*.....	7
Figure S6 Association analysis summary for C-reactive protein.....	8
Figure S7 Association analysis summary for calcium.....	9
Figure S8 Association analysis summary for cholesterol*.....	10
Figure S9 Association analysis summary for creatinine.....	11
Figure S10 Association analysis summary for cystatin C.....	12
Figure S11 Association analysis summary for direct bilirubin.....	13
Figure S12 Association analysis summary for gamma glutamyltransferase.....	14
Figure S13 Association analysis summary for glucose.....	15
Figure S14 Association analysis summary for glycated haemoglobin (HbA1c).....	16
Figure S15 Association analysis summary for HDL cholesterol.....	17
Figure S16 Association analysis summary for IGF1.....	18
Figure S17 Association analysis summary for lipoprotein A.....	19
Figure S18 Association analysis summary for phosphate.....	20
Figure S19 Association analysis summary for SHBG.....	21
Figure S20 Association analysis summary for testosterone.....	22
Figure S21 Association analysis summary for total bilirubin.....	23
Figure S22 Association analysis summary for total protein.....	24
Figure S23 Association analysis summary for triglycerides.....	25
Figure S24 Association analysis summary for urate.....	27
Figure S25 Association analysis summary for urea.....	28
Figure S26 Association analysis summary for vitamin D.....	28
Figure S27 True vs. Predicted value plot (left) and top 10 features (right) for alanine aminotransferase.....	29
Figure S28 True vs. Predicted value plot (left) and top 10 features (right) for albumin.....	30
Figure S29 True vs. Predicted value plot (left) and top 10 features (right) for alkaline phosphatase.....	31
Figure S30 True vs. Predicted value plot (left) and top 10 features (right) for apolipoprotein A.....	32
Figure S31 True vs. Predicted value plot (left) and top 10 features (right) for apolipoprotein B*.....	33
Figure S32 True vs. Predicted value plot (left) and top 10 features (right) for aspartate aminotransferase.....	34
Figure S33 True vs. Predicted value plot (left) and top 10 features (right) for calcium.....	35
Figure S34 True vs. Predicted value plot (left) and top 10 features (right) for cholesterol*.....	36
Figure S35 True vs. Predicted value plot (left) and top 10 features (right) for C reactive protein.....	37
Figure S36 True vs. Predicted value plot (left) and top 10 features (right) for creatinine.....	38
Figure S37 True vs. Predicted value plot (left) and top 10 features (right) for cystatin C.....	39
Figure S38 True vs. Predicted value plot (left) and top 10 features (right) for direct bilirubin.....	40
Figure S39 True vs. Predicted value plot (left) and top 10 features (right) for gamma glutamyltransferase.....	41
Figure S40 True vs. Predicted value plot (left) and top 10 features (right) for glucose.....	42
Figure S41 True vs. Predicted value plot (left) and top 10 features (right) for glycated haemoglobin (HbA1c).....	43
Figure S42 True vs. Predicted value plot (left) and top 10 features (right) for HDL cholesterol.....	44
Figure S43 True vs. Predicted value plot (left) and top 10 features (right) for IGF1.....	45
Figure S44 True vs. Predicted value plot (left) and top 10 features (right) for LDL direct*.....	46
Figure S45 True vs. Predicted value plot (left) and top 10 features (right) for lipoprotein A.....	47
Figure S46 True vs. Predicted value plot (left) and top 10 features (right) for phosphate.....	48
Figure S47 True vs. Predicted value plot (left) and top 10 features (right) for SHBG.....	49
Figure S48 True vs. Predicted value plot (left) and top 10 features (right) for testosterone.....	50
Figure S49 True vs. Predicted value plot (left) and top 10 features (right) for total bilirubin.....	51
Figure S50 True vs. Predicted value plot (left) and top 10 features (right) for total protein.....	52
Figure S51 True vs. Predicted value plot (left) and top 10 features (right) for triglycerides.....	53

<i>Figure S52 True vs. Predicted value plot (left) and top 10 features (right) for urate.....</i>	<i>54</i>
<i>Figure S53 True vs. Predicted value plot (left) and top 10 features (right) for urea</i>	<i>55</i>
<i>Figure S54 True vs. Predicted value plot (left) and top 10 features (right) for vitamin D.....</i>	<i>56</i>

Biomarkers distribution

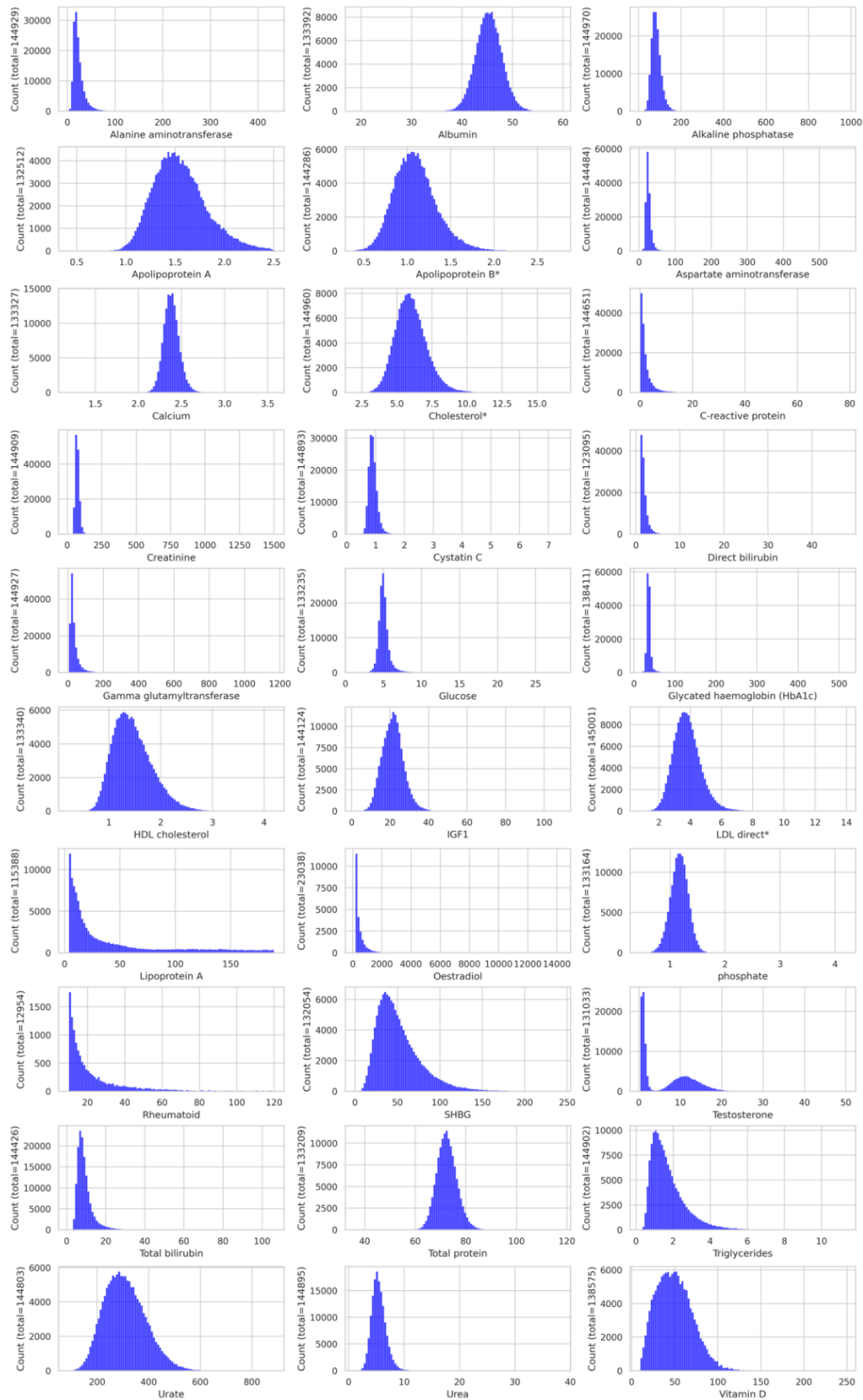


Figure S1 Distribution of samples across biomarkers.

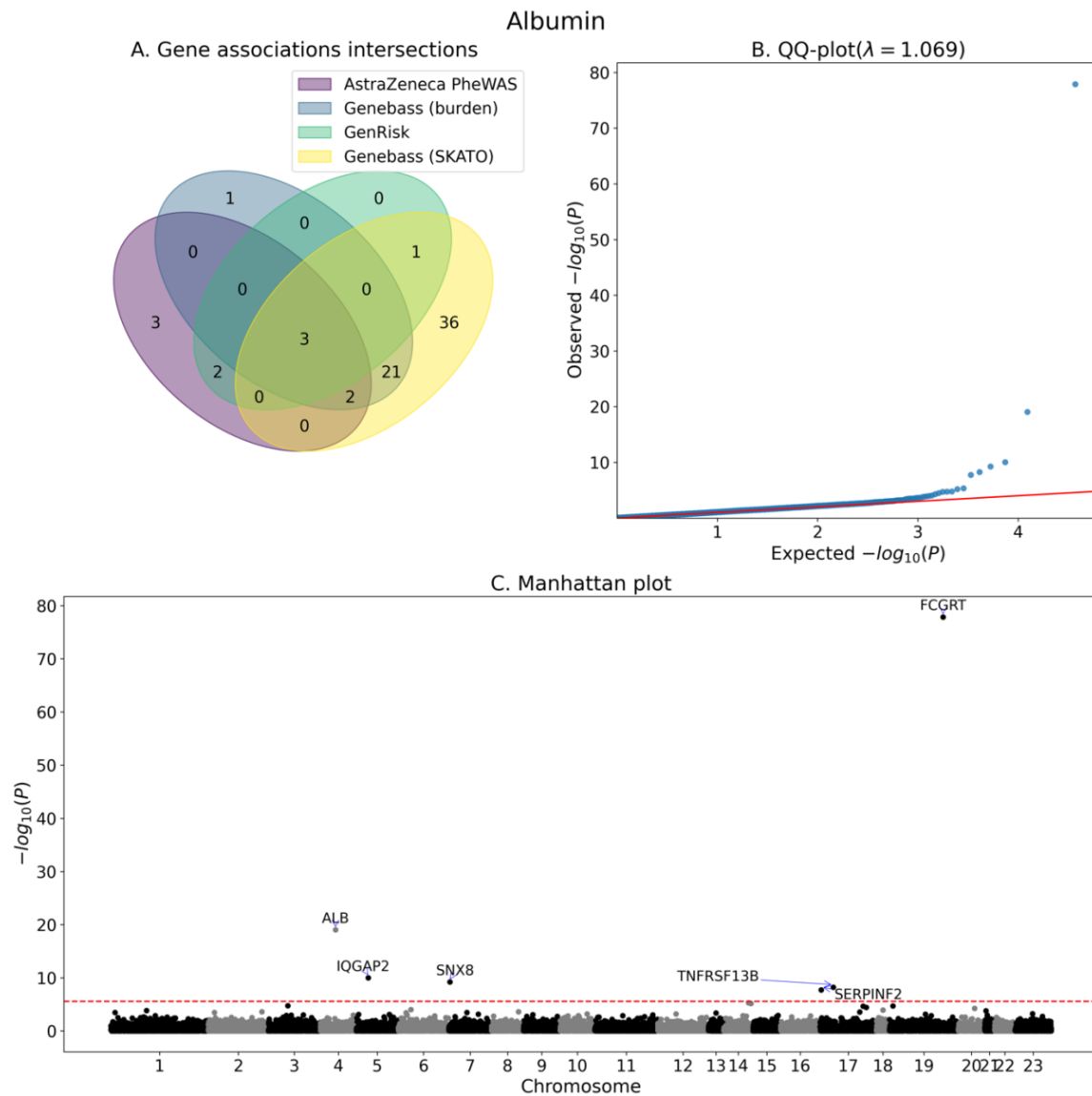


Figure S2 Association analysis summary for albumin.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

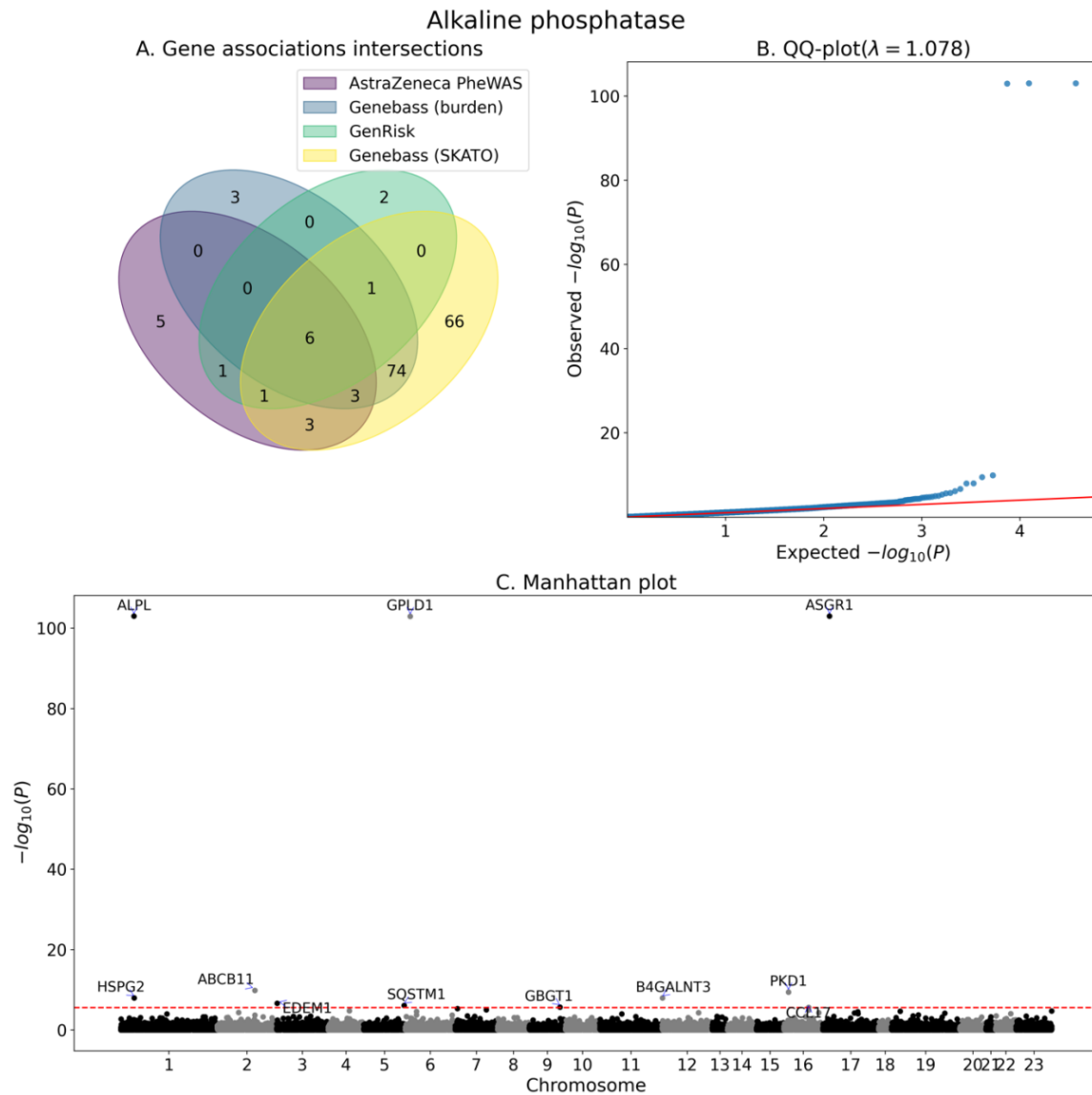


Figure S3 Association analysis summary for alkaline phosphatase.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

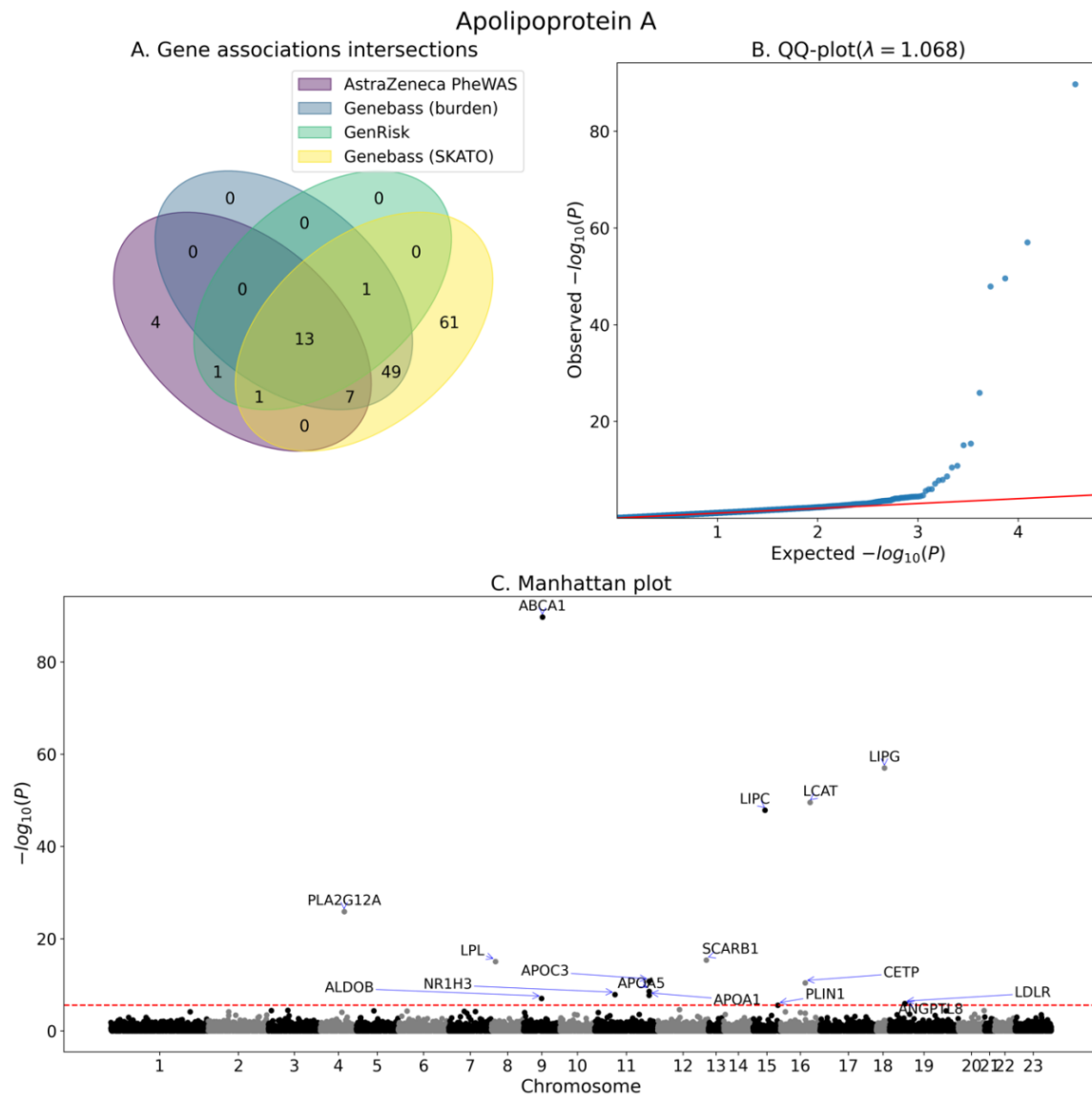


Figure S4 Association analysis summary for apolipoprotein A.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

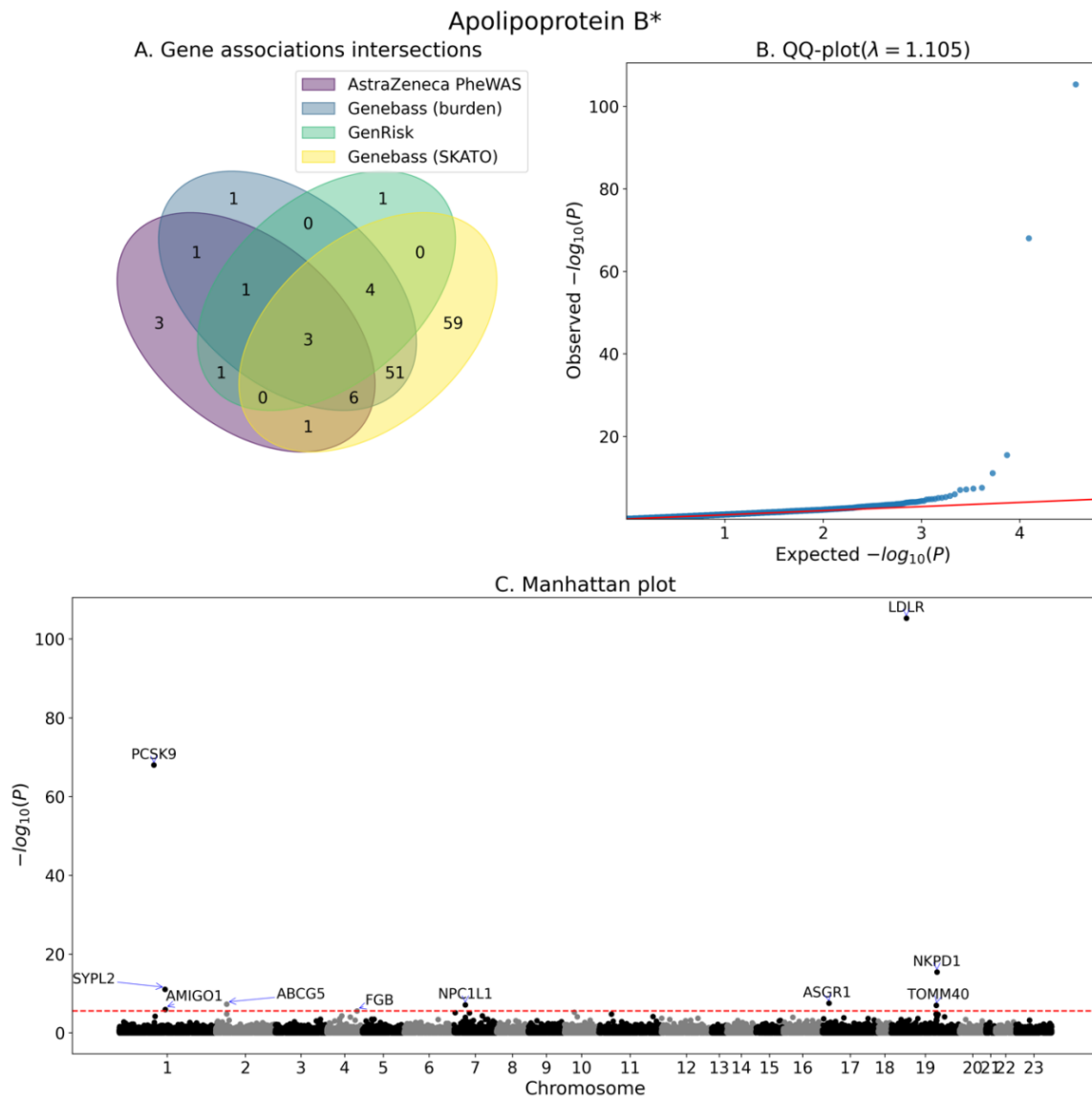


Figure S5 Association analysis summary for apolipoprotein B*.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

* statin adjusted values

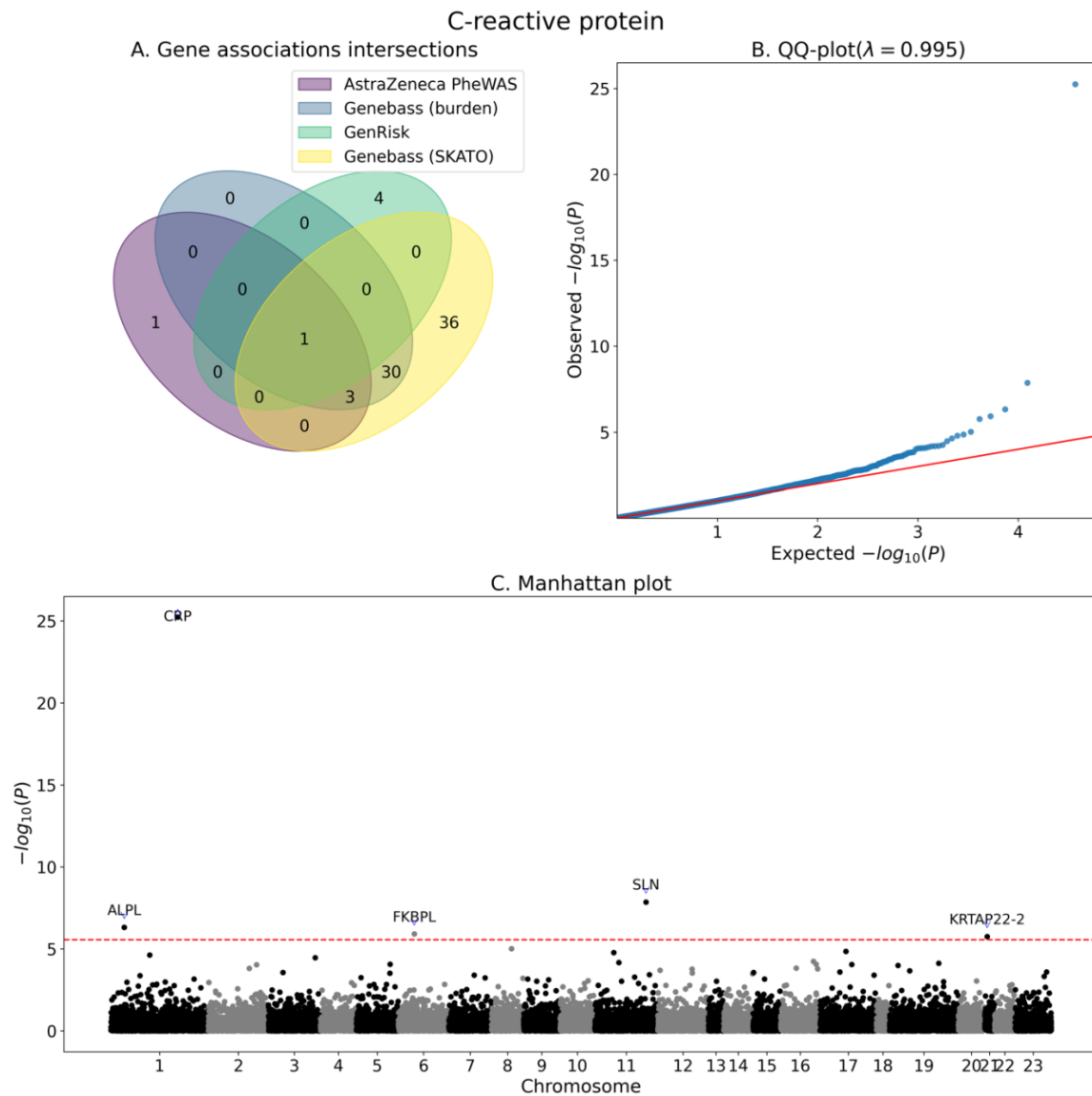


Figure S6 Association analysis summary for C-reactive protein.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

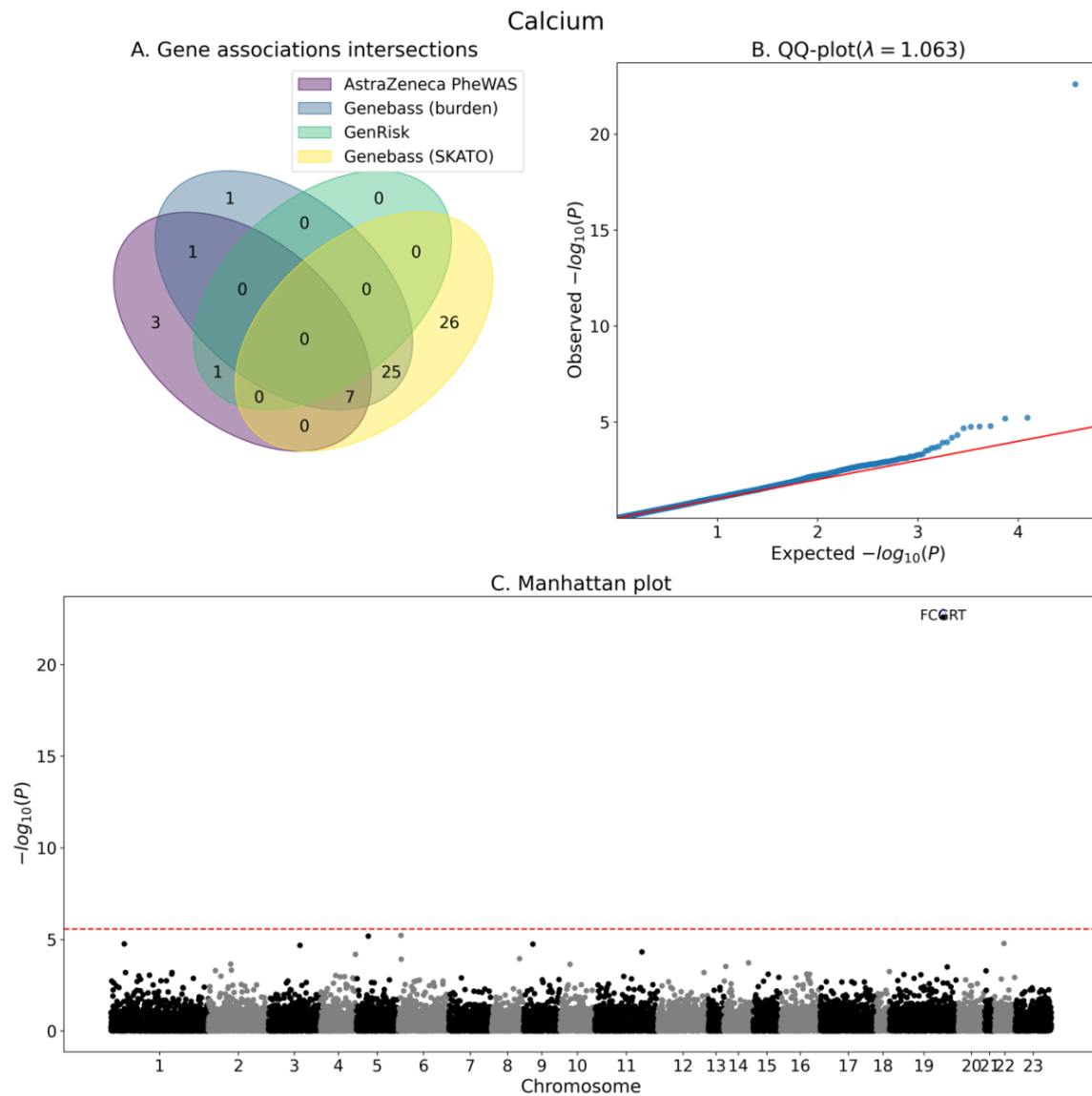


Figure S7 Association analysis summary for calcium.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

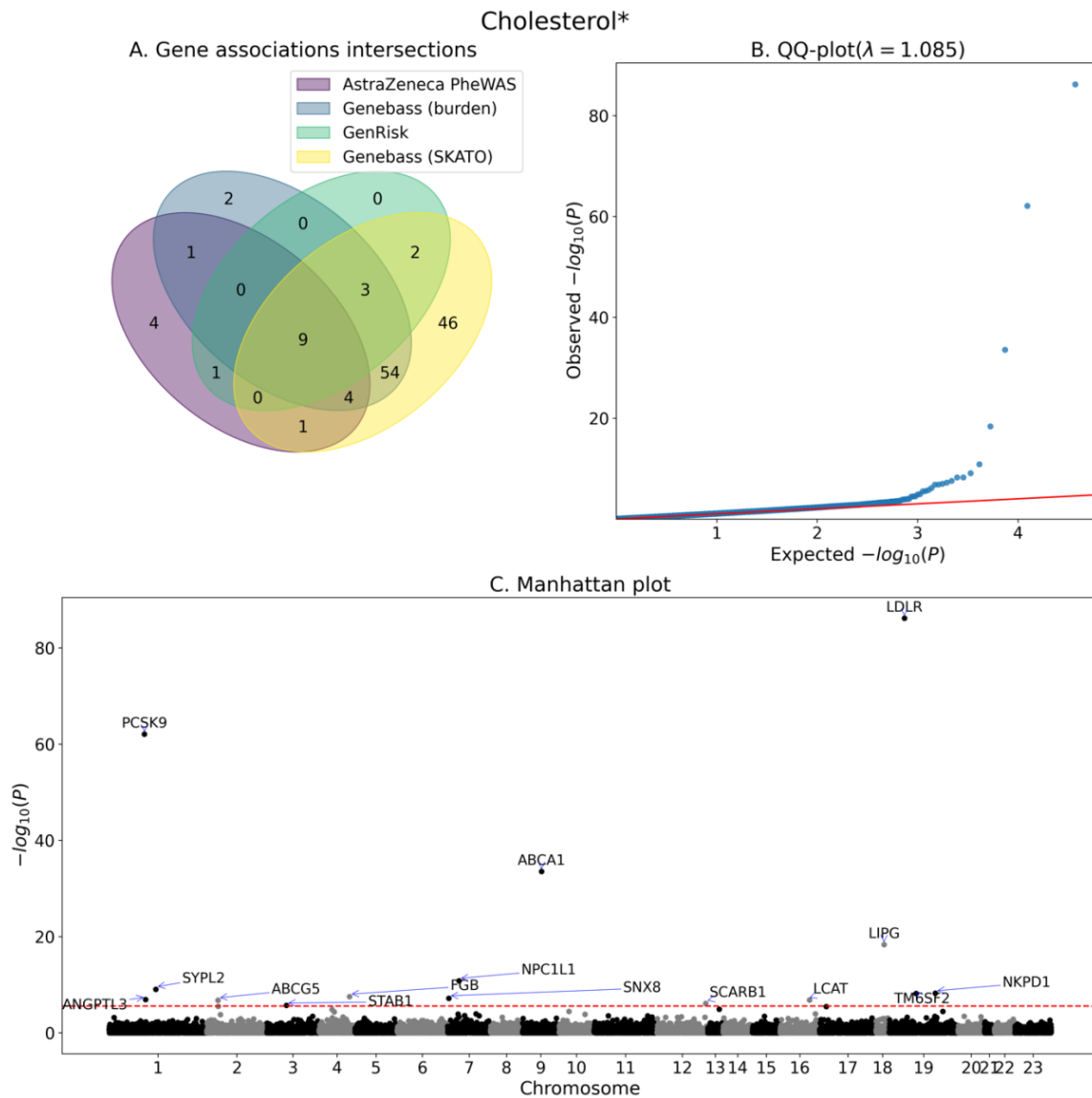


Figure S8 Association analysis summary for cholesterol*.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

* statin adjusted values

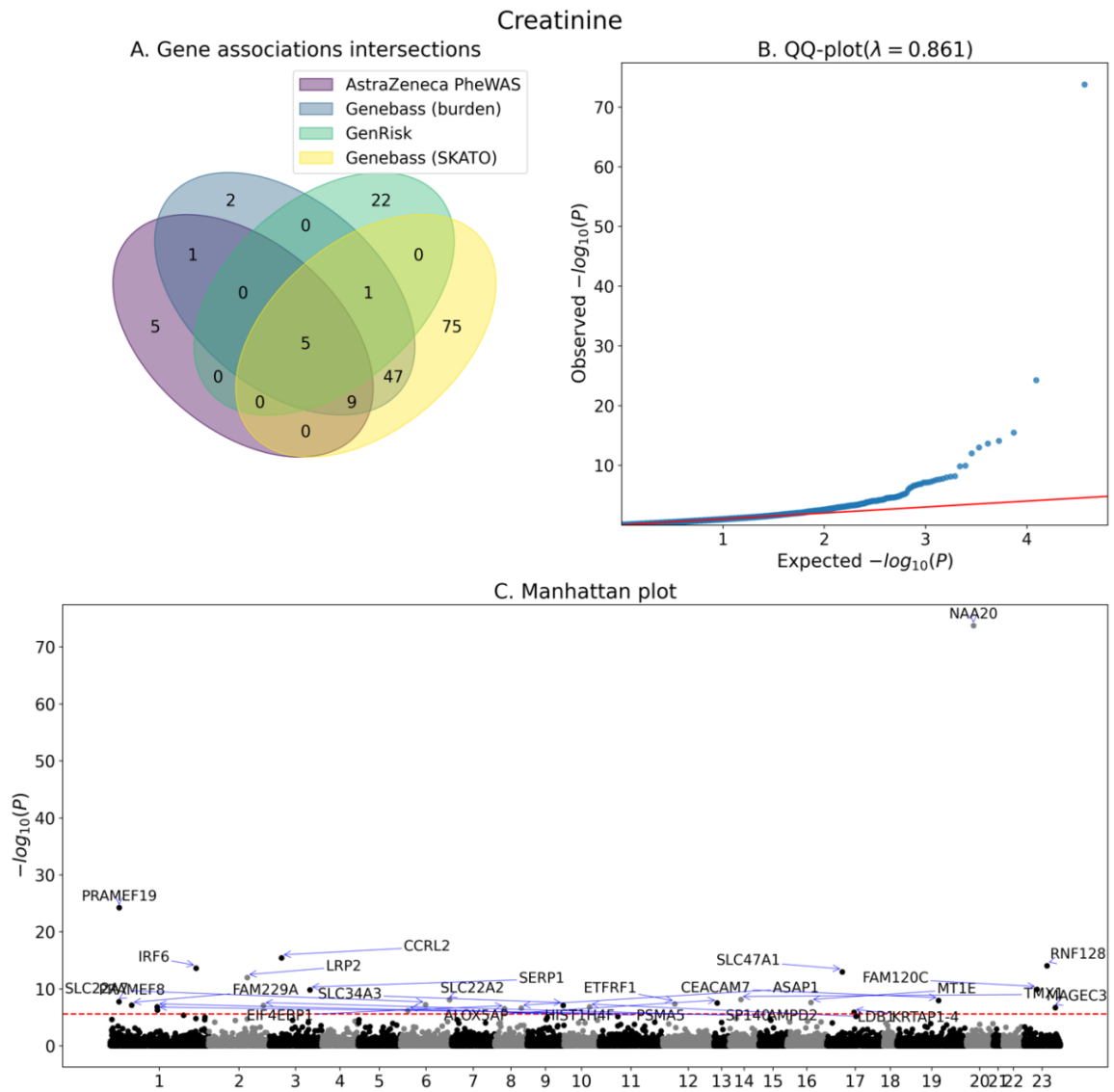


Figure S9 Association analysis summary for creatinine.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

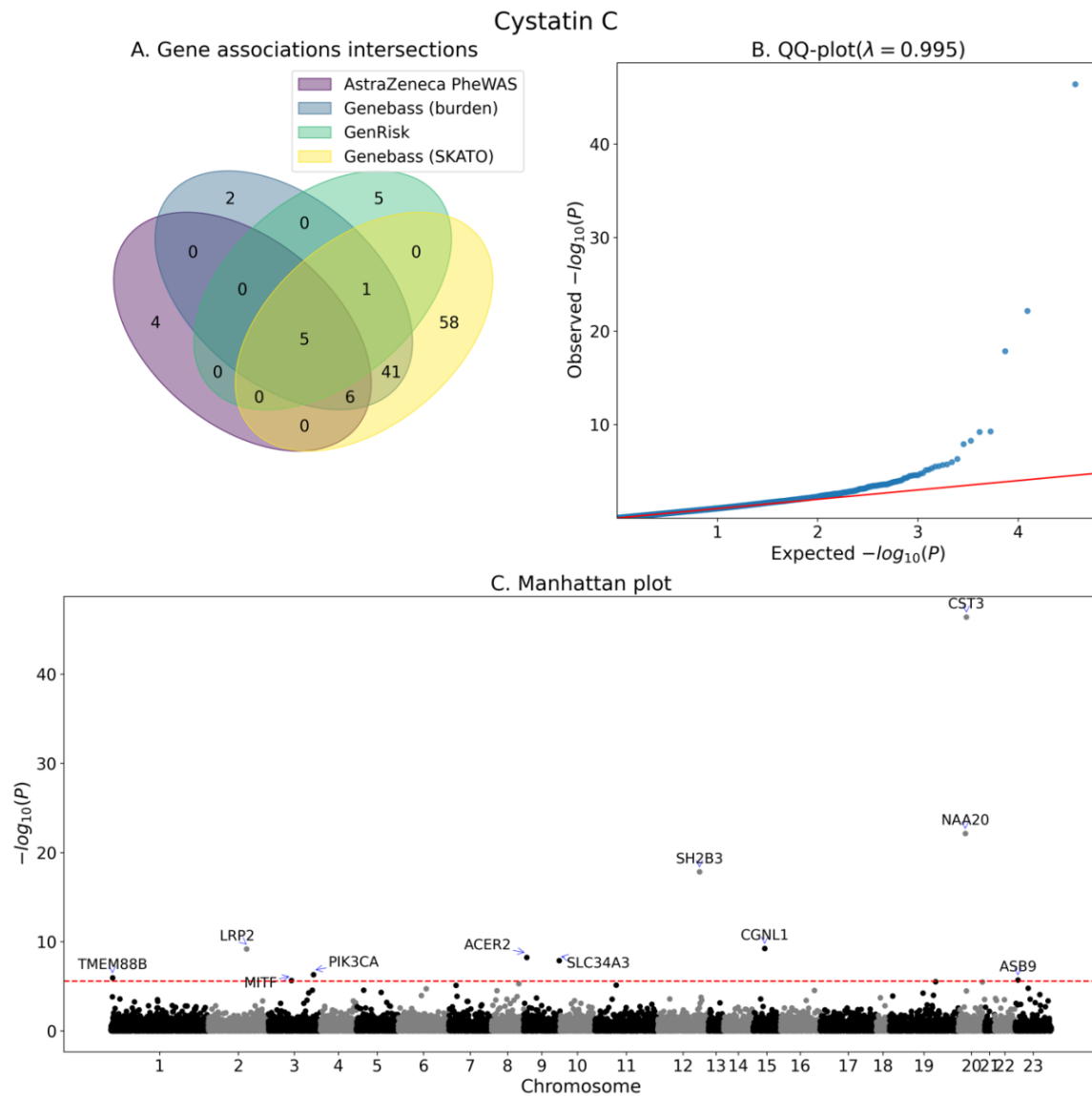


Figure S10 Association analysis summary for cystatin C.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

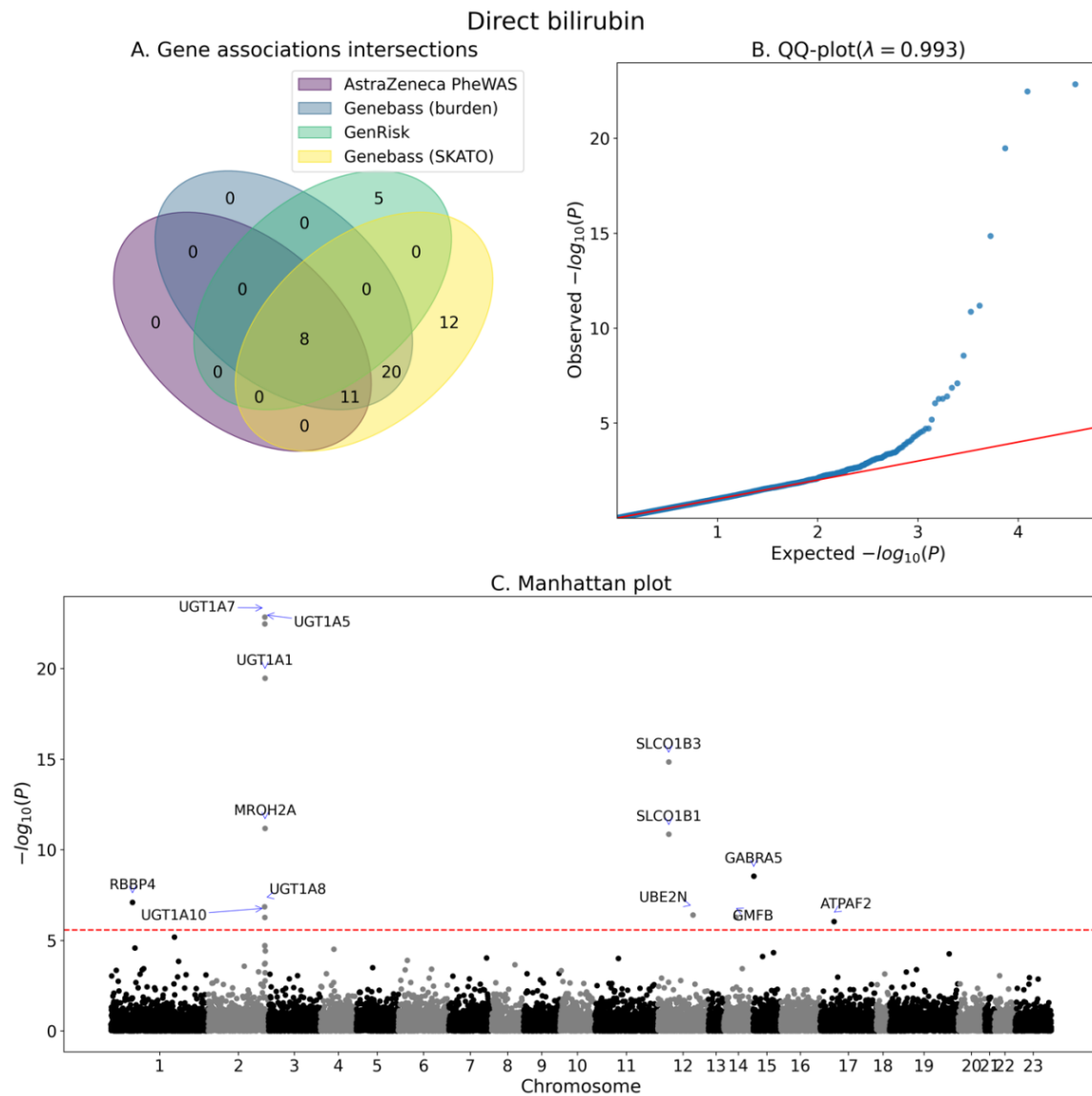


Figure S11 Association analysis summary for direct bilirubin.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

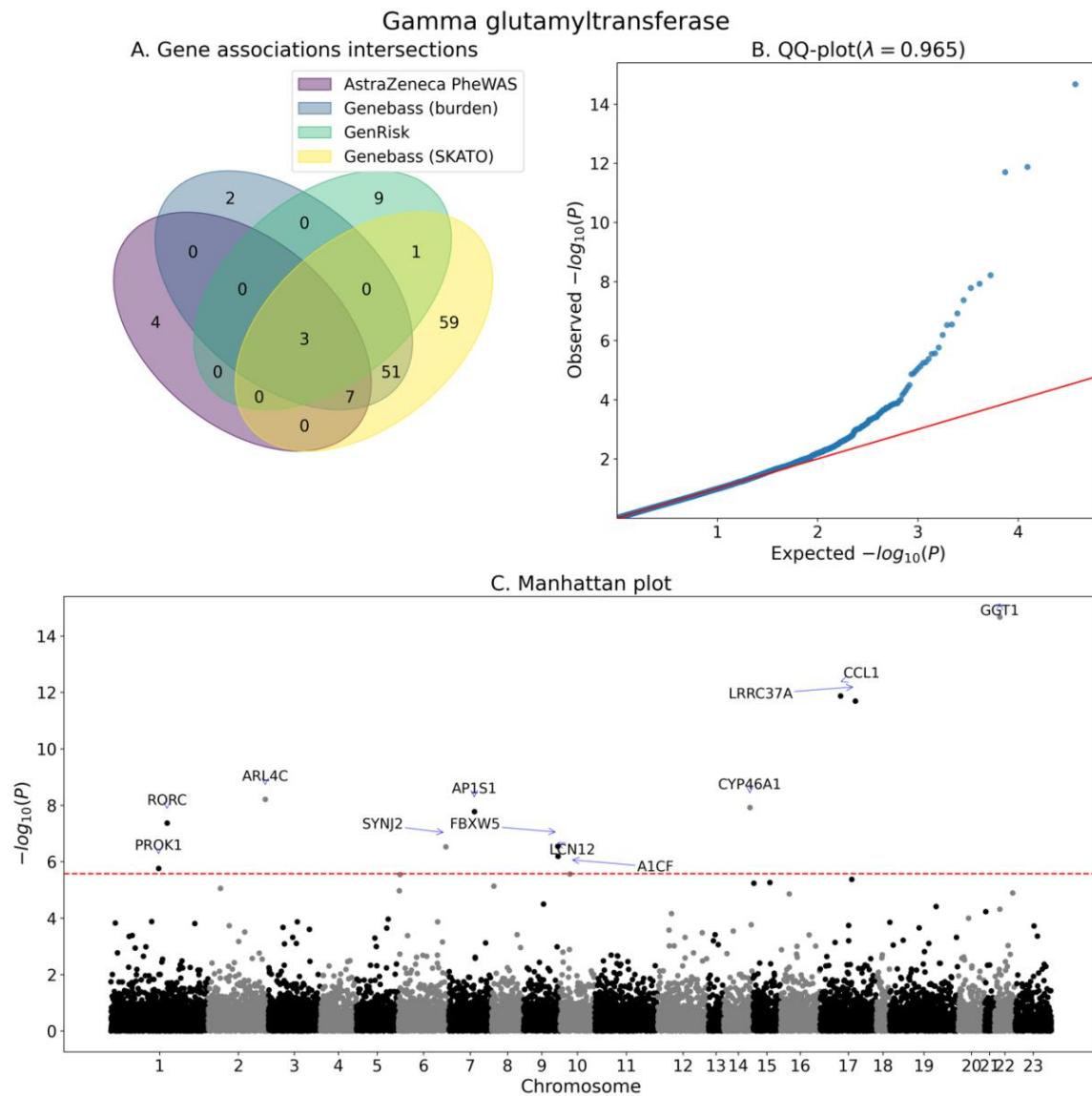


Figure S12 Association analysis summary for gamma glutamyltransferase.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

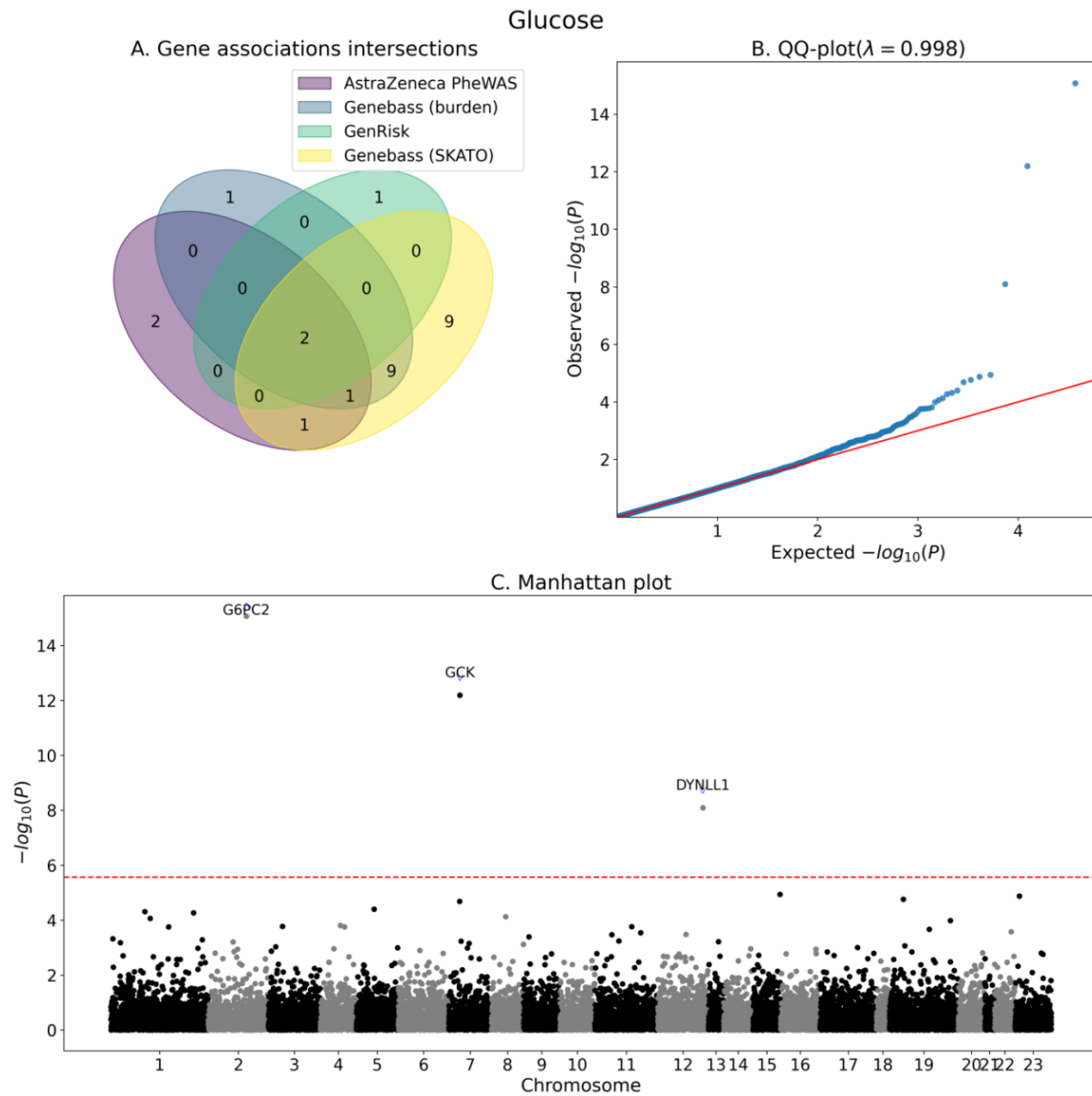


Figure S13 Association analysis summary for glucose.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

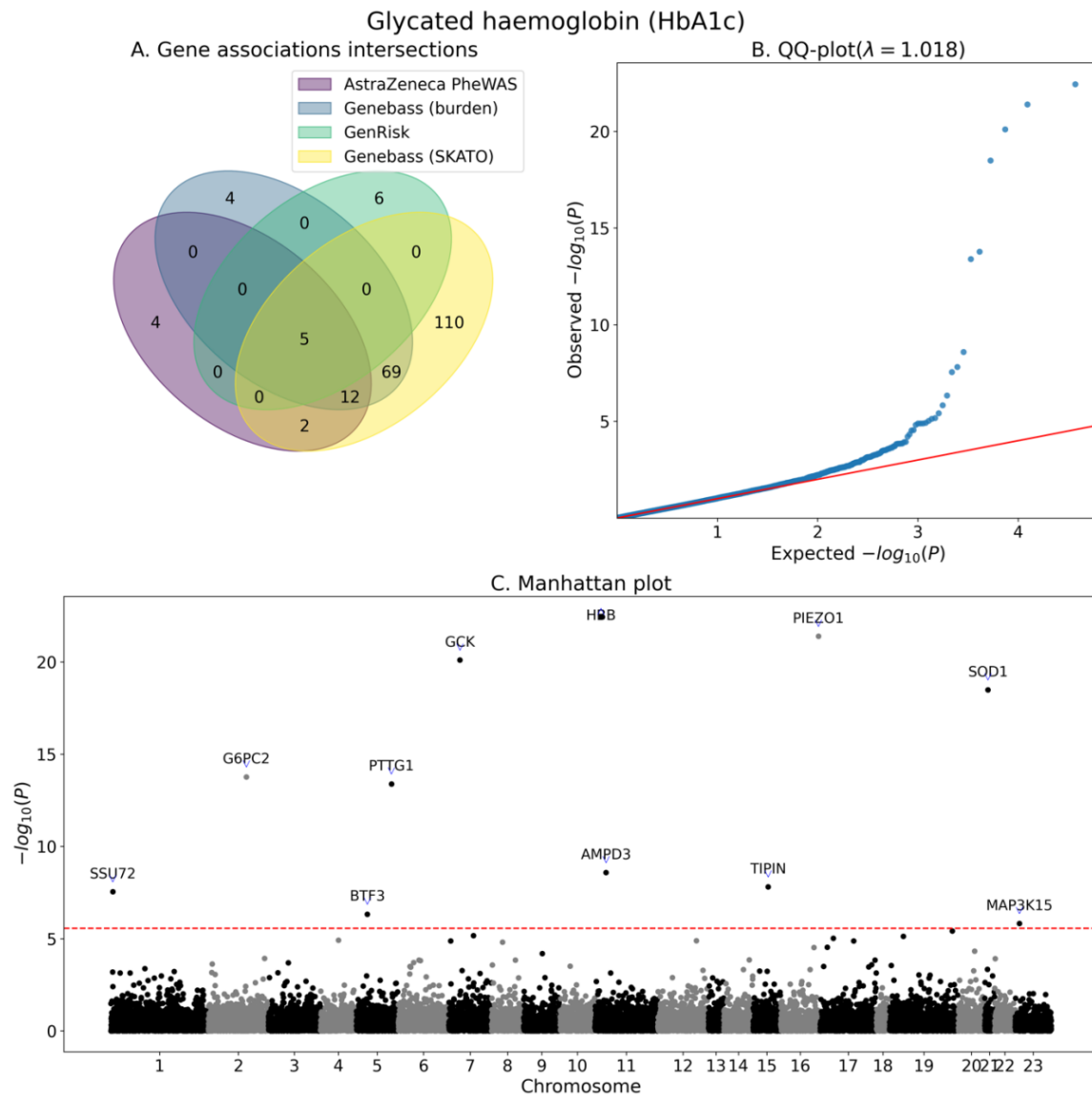


Figure S14 Association analysis summary for glycated haemoglobin (HbA1c).

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

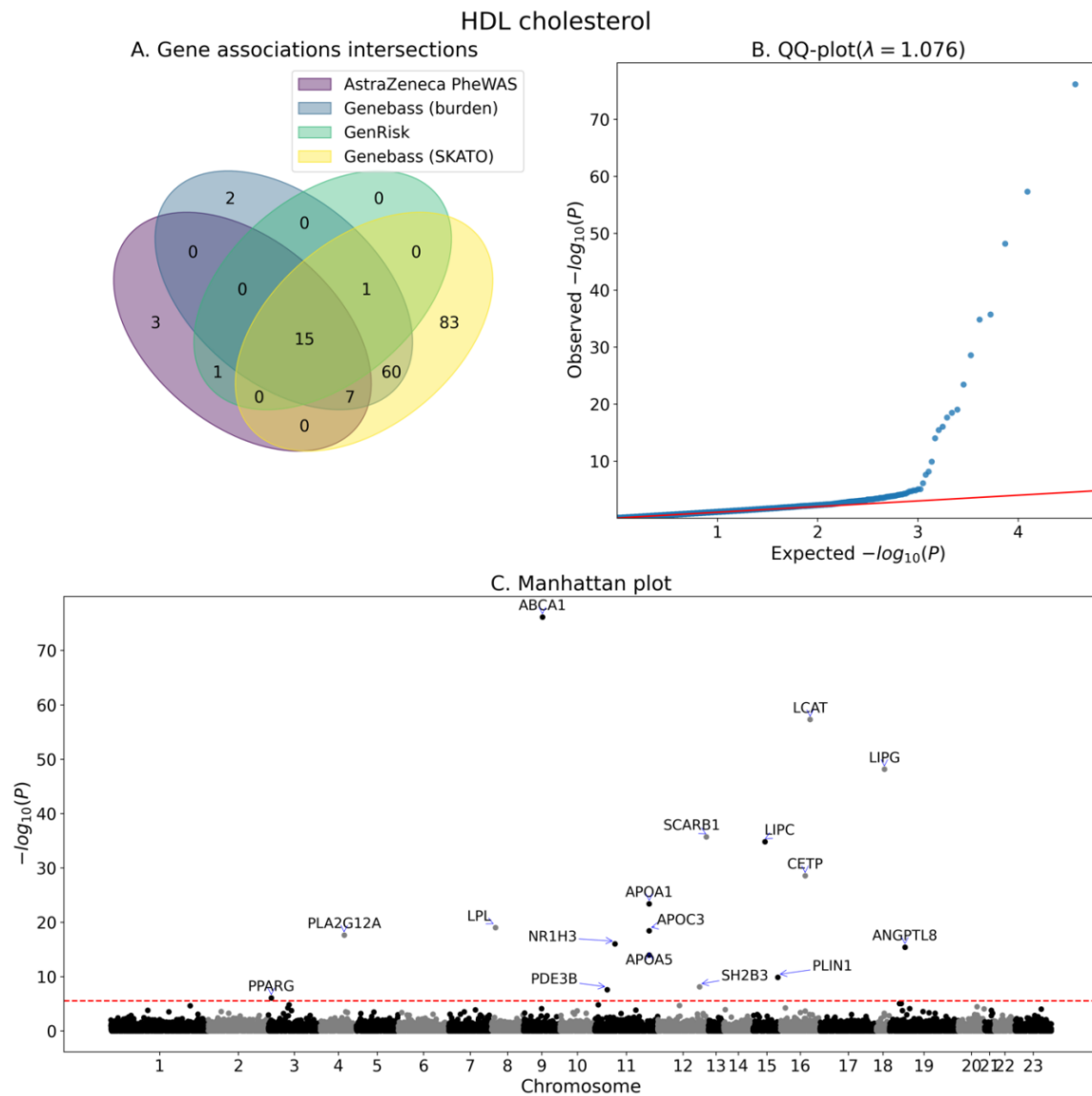


Figure S15 Association analysis summary for HDL cholesterol.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

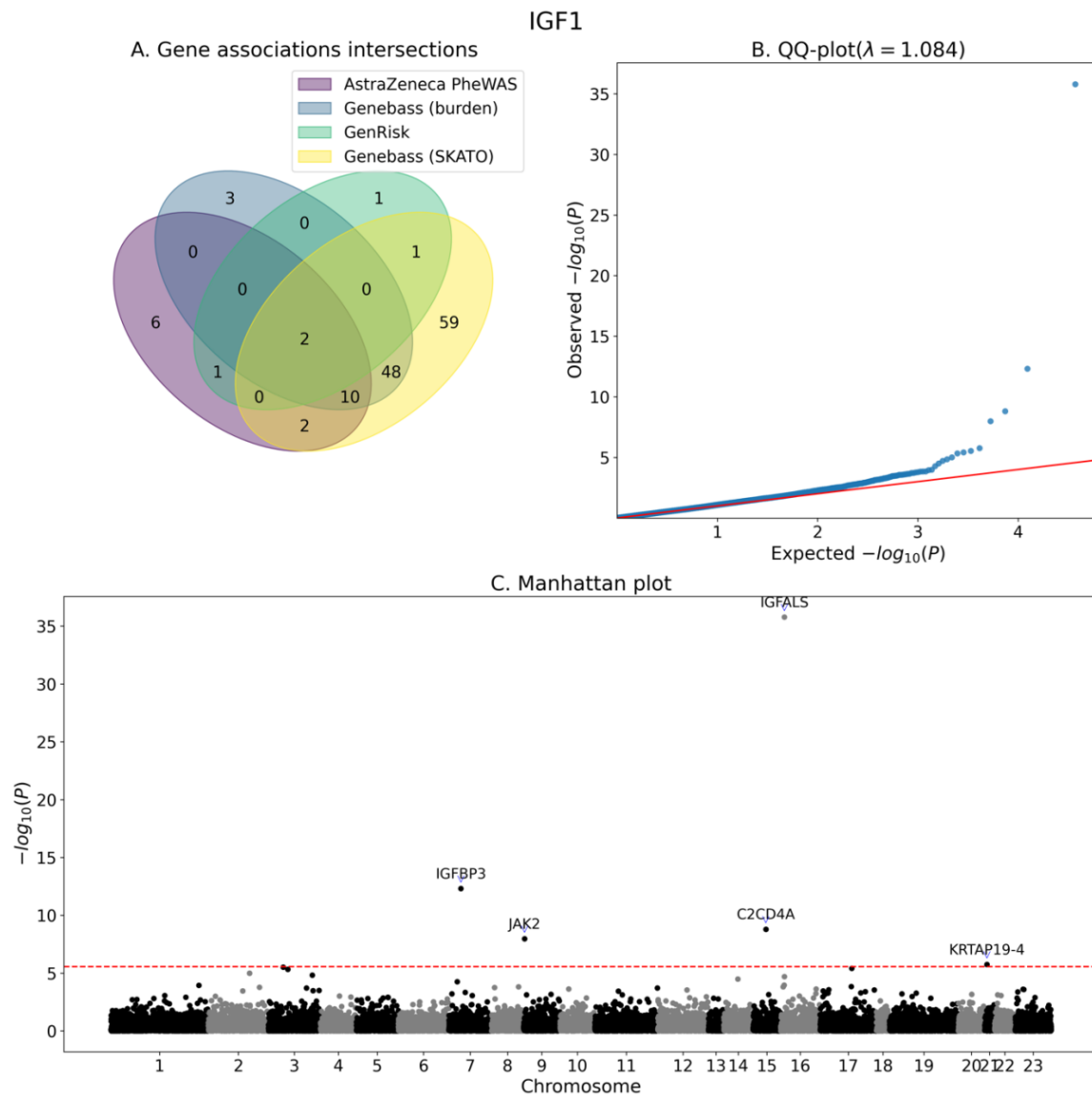


Figure S16 Association analysis summary for IGF1.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

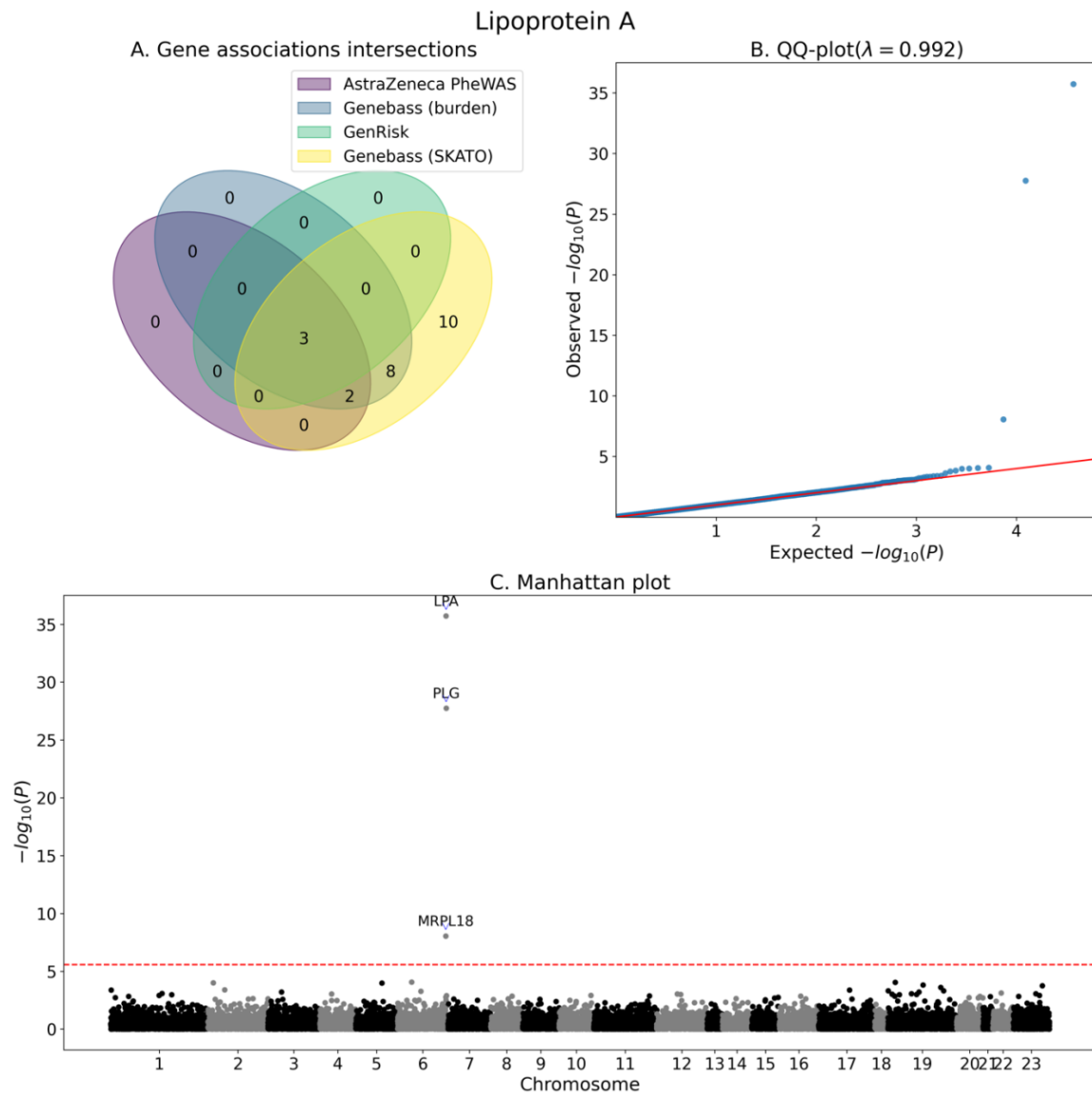


Figure S17 Association analysis summary for lipoprotein A.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

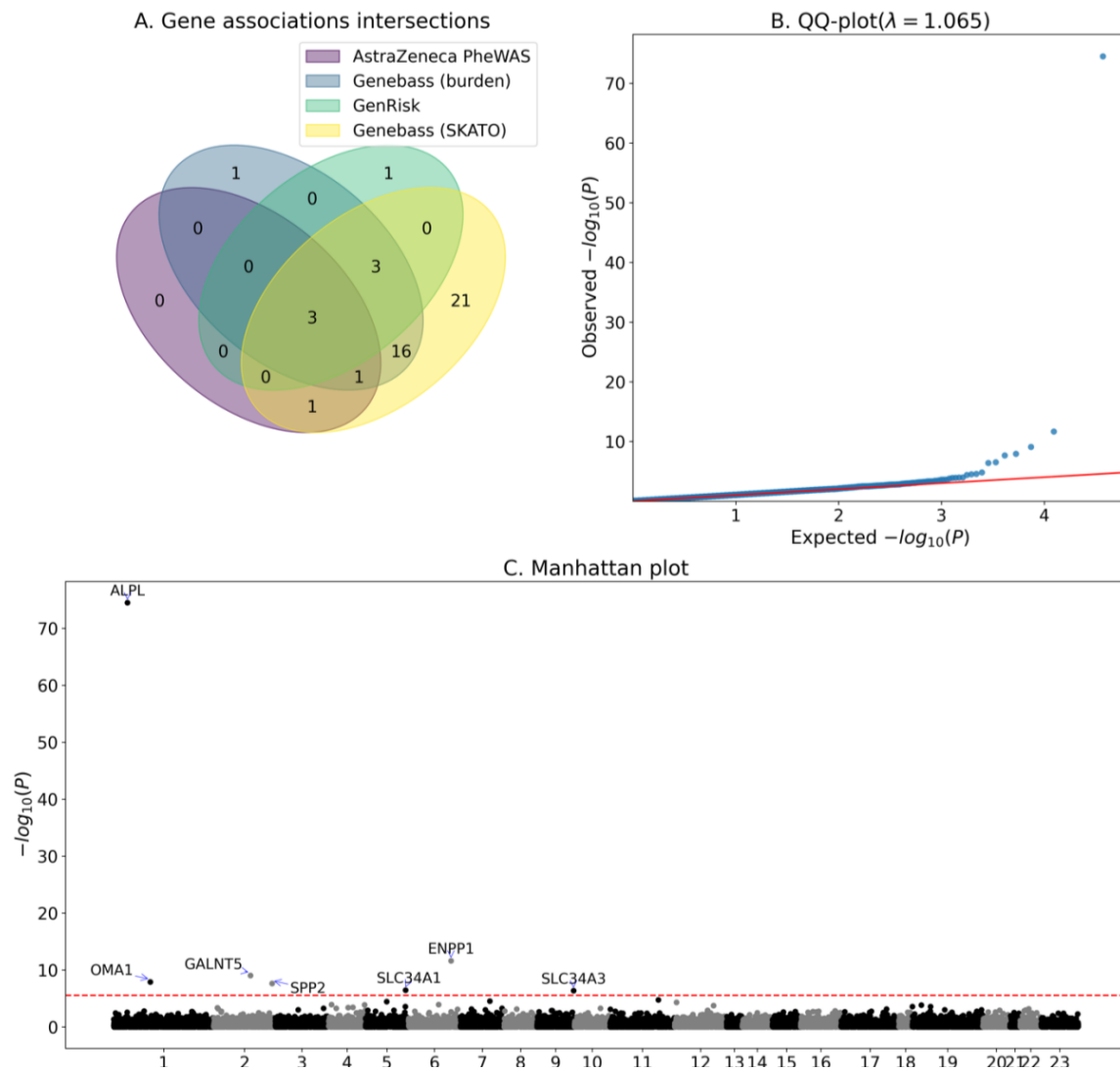


Figure S18 Association analysis summary for phosphate.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

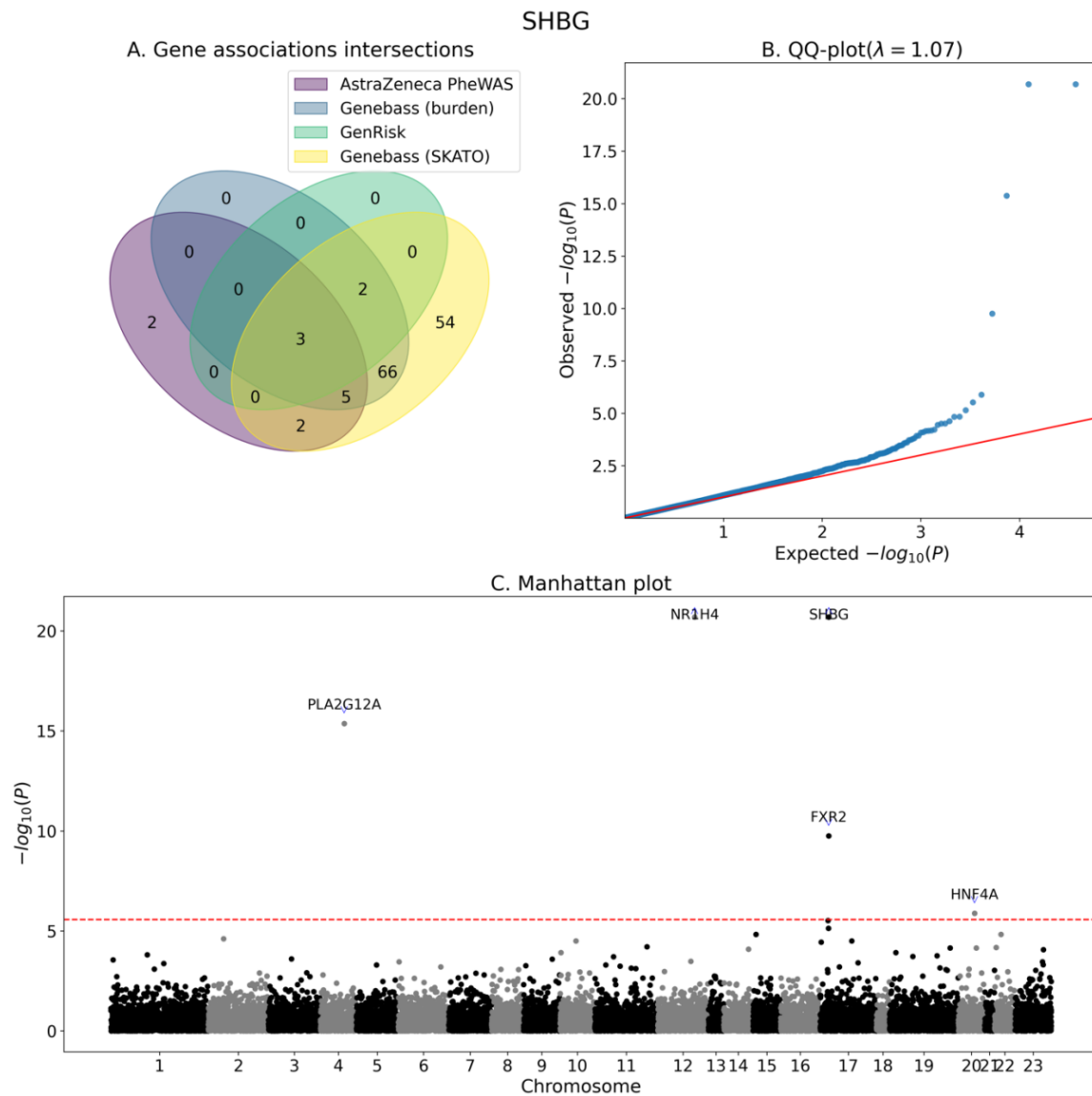


Figure S19 Association analysis summary for SHBG.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

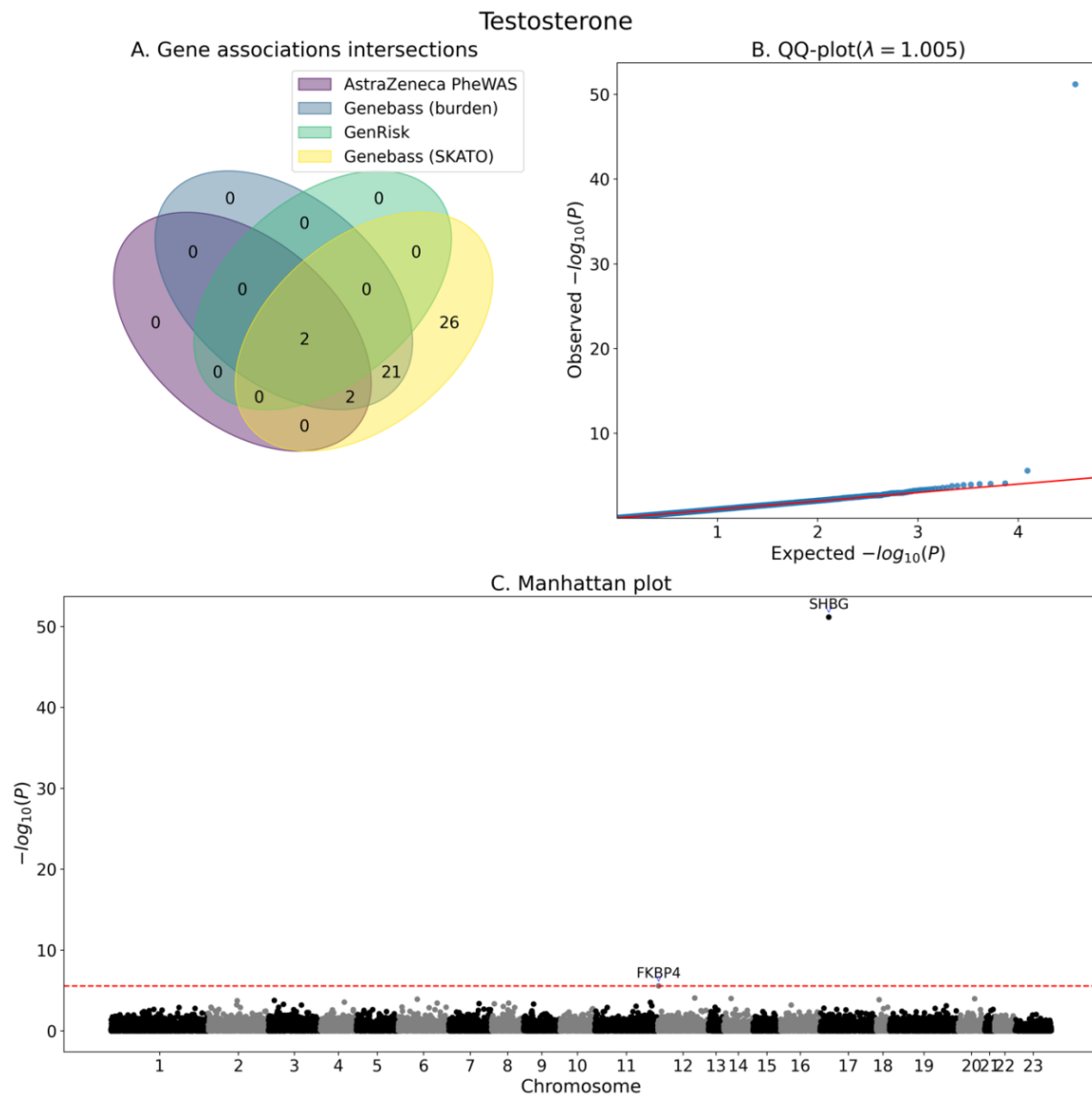


Figure S20 Association analysis summary for testosterone.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

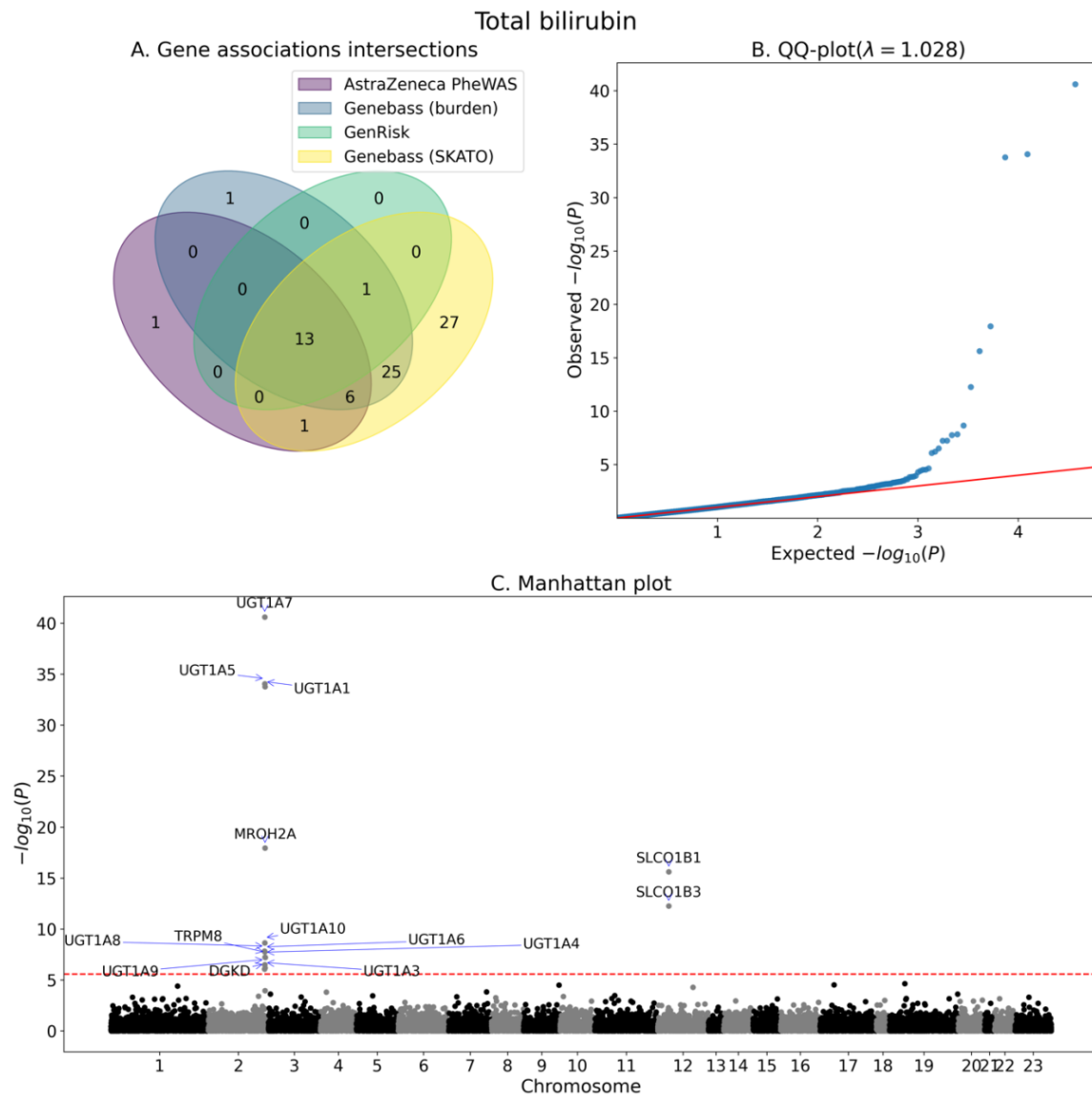


Figure S21 Association analysis summary for total bilirubin.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

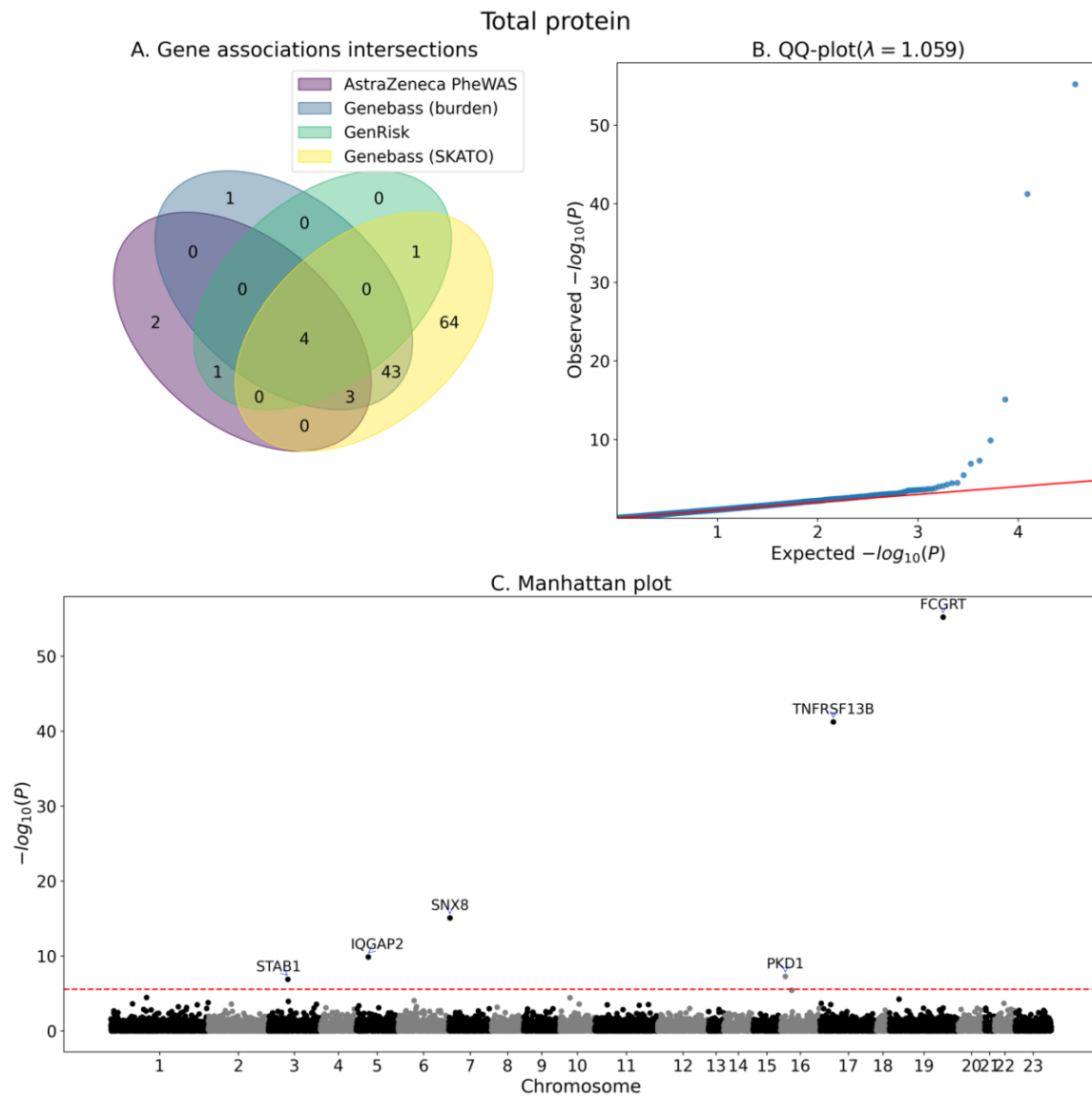


Figure S22 Association analysis summary for total protein.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

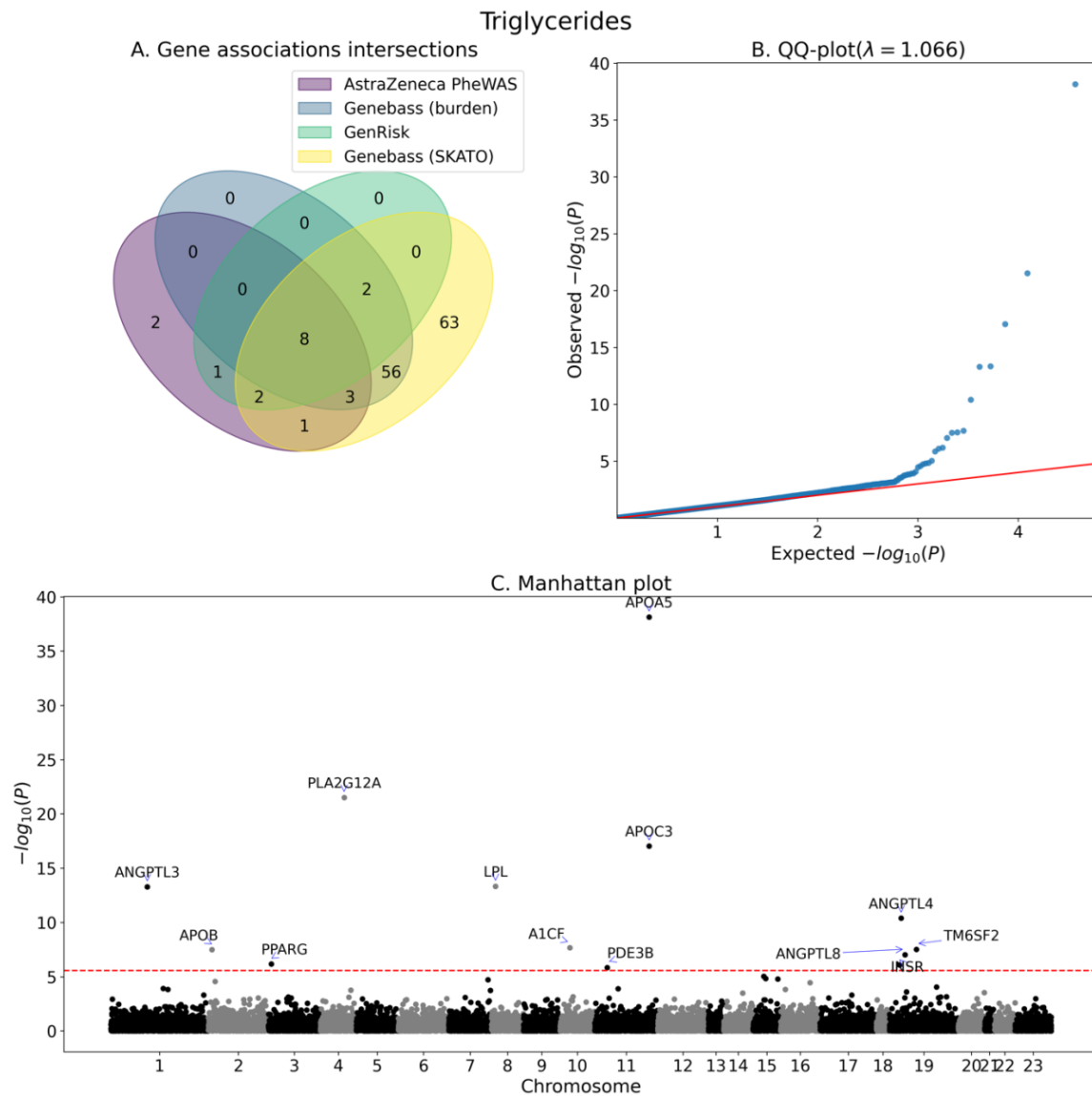


Figure S23 Association analysis summary for triglycerides.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

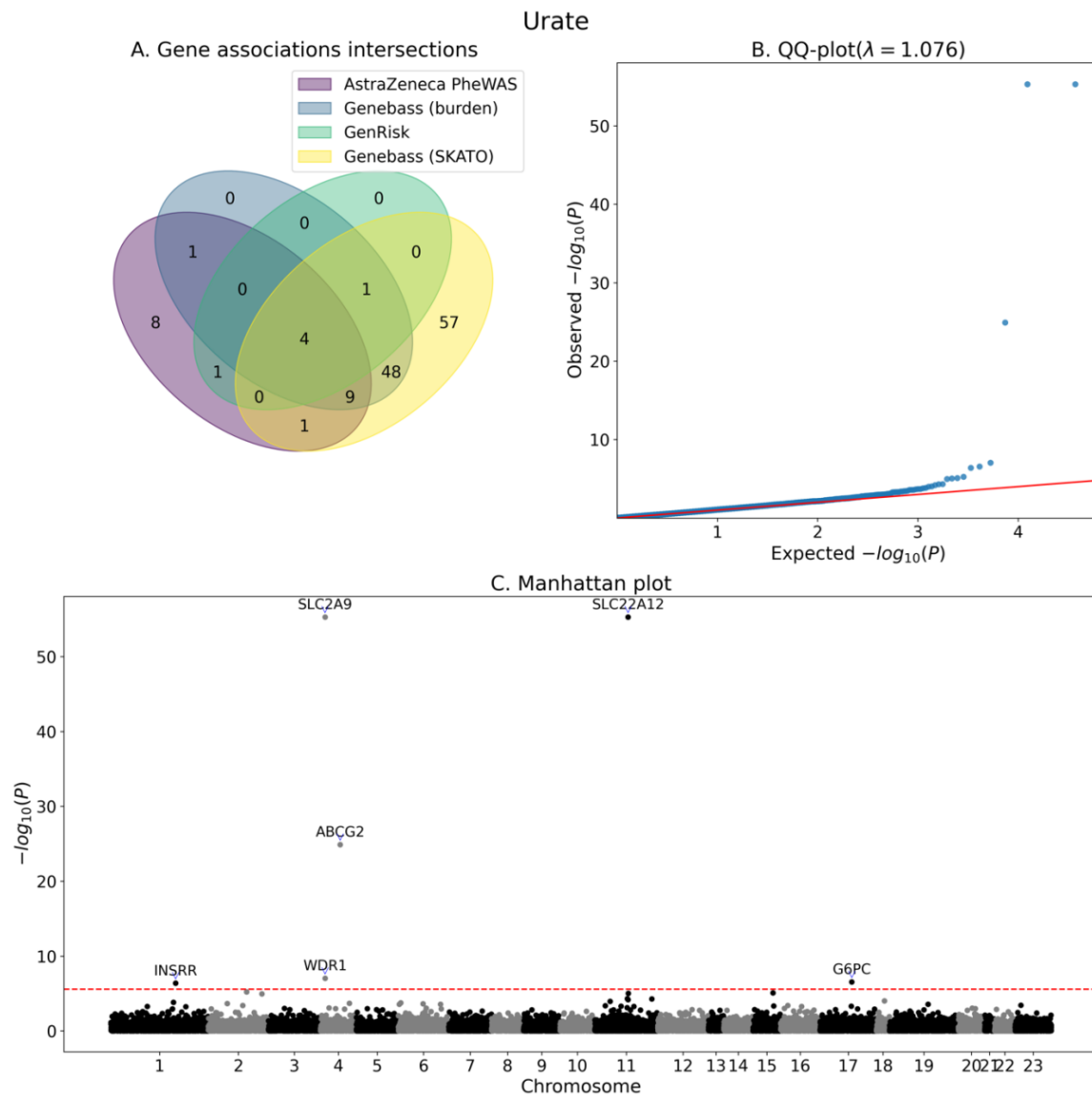


Figure S24 Association analysis summary for urate.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results

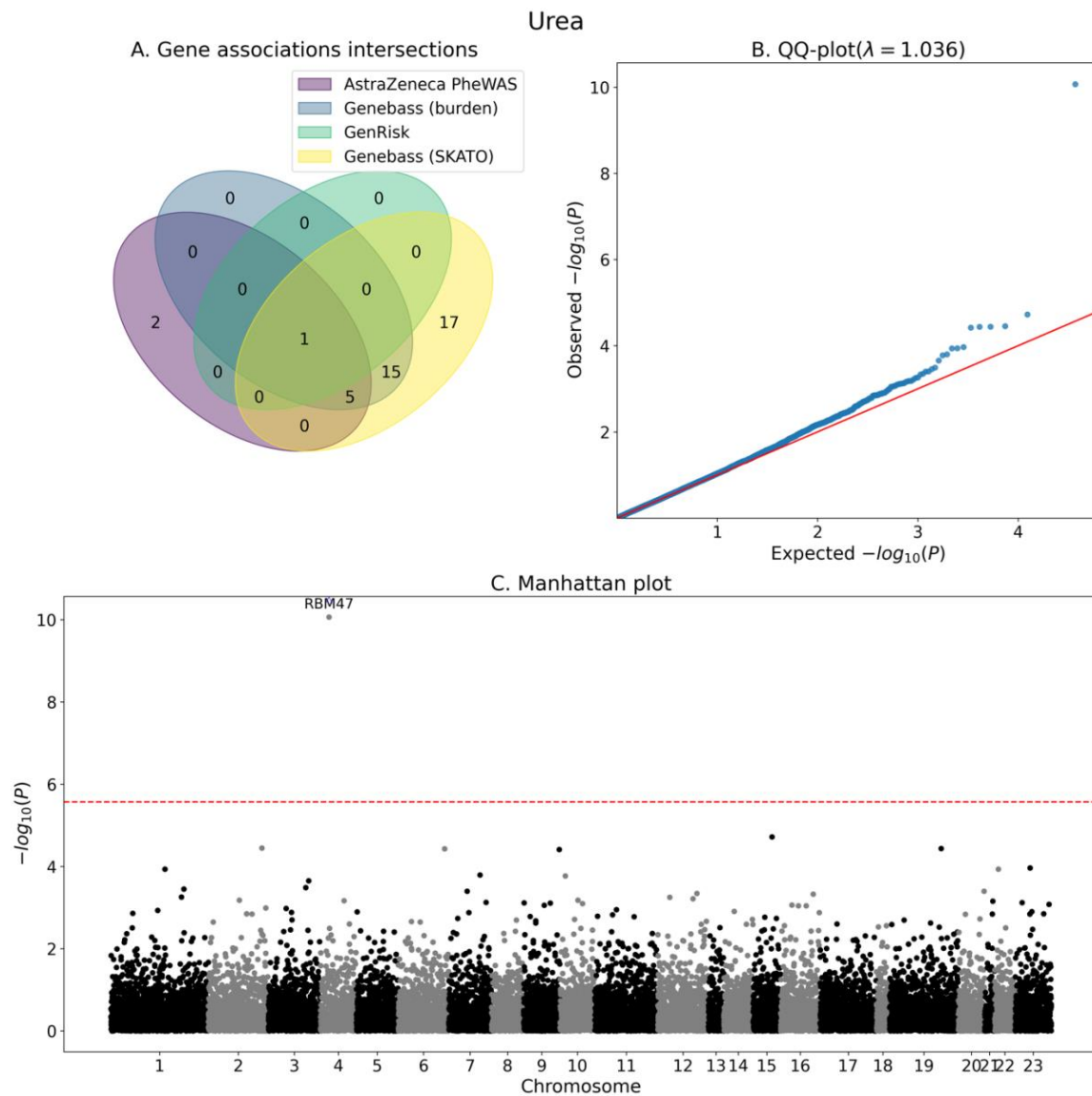


Figure S25 Association analysis summary for urea.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results

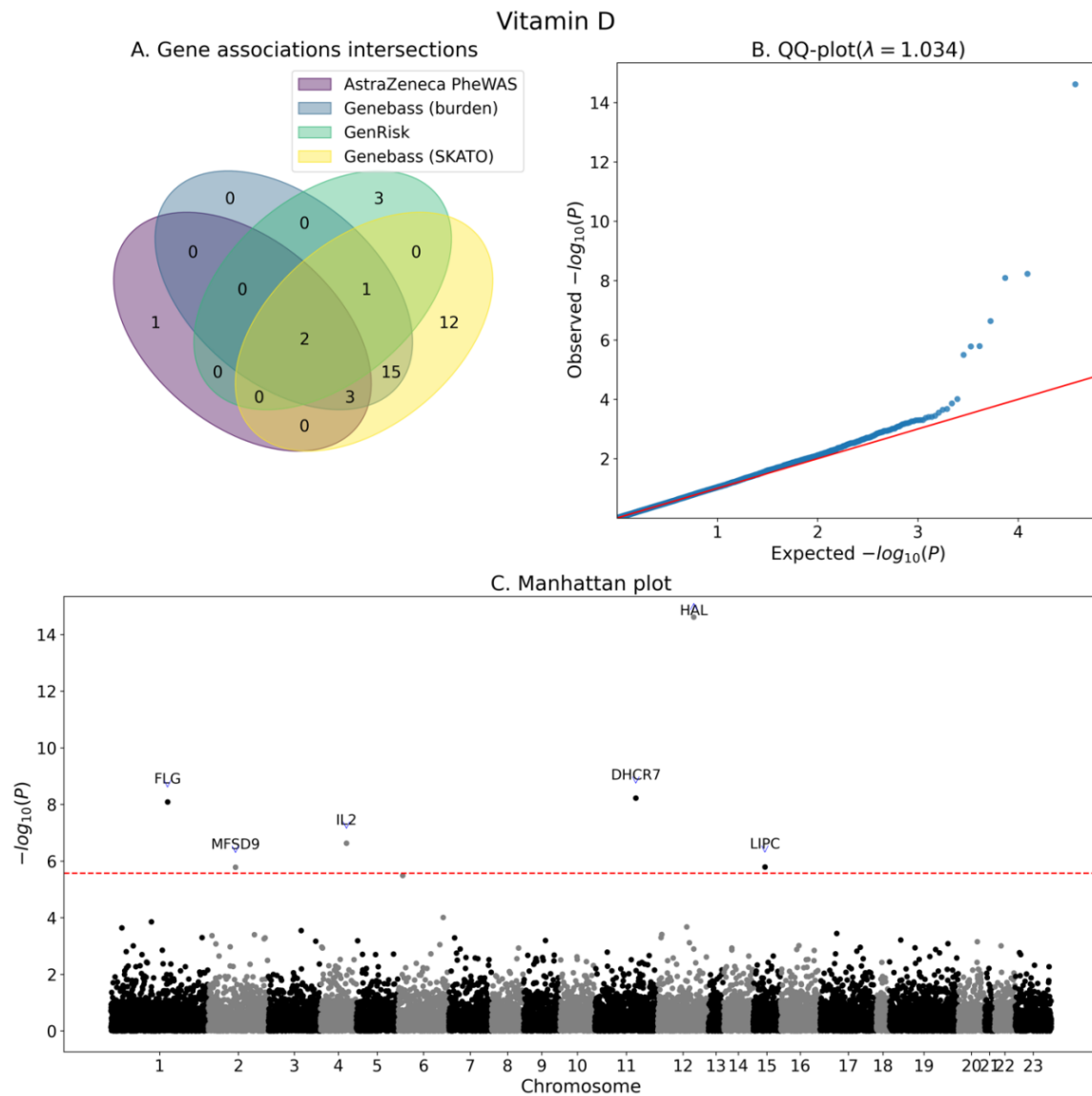
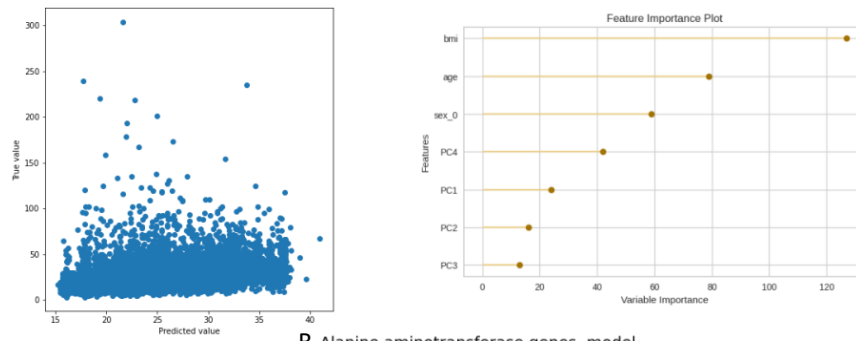


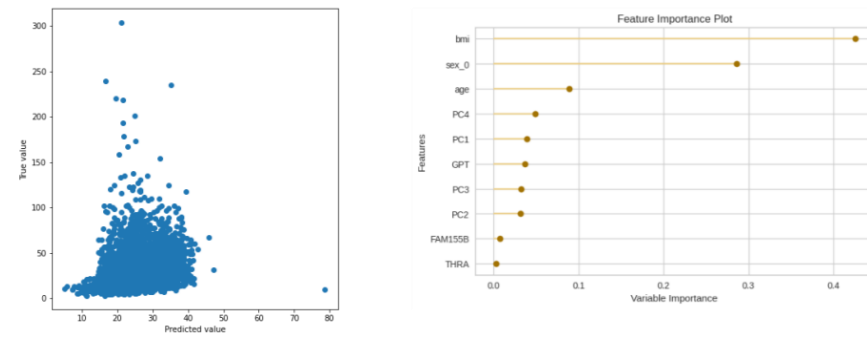
Figure S26 Association analysis summary for vitamin D.

A. Venn diagram of the number significantly associated genes as identified by GenRisk, AstraZeneca PheWAS and genebass. B. QQ-plot of the P-values of GenRisk pipeline results. C. Manhattan plot of GenRisk pipeline results.

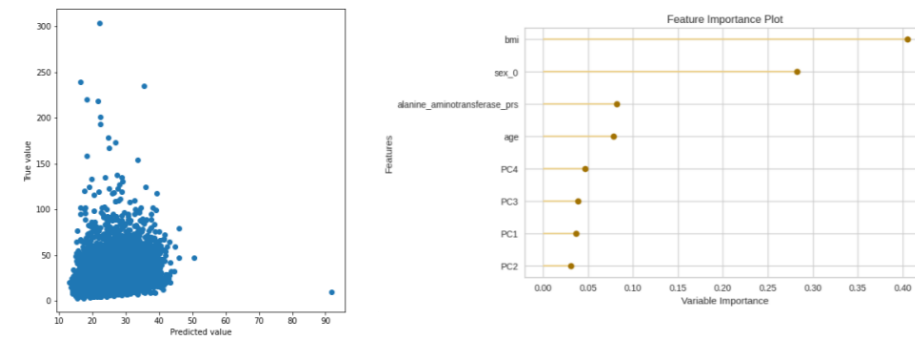
A. Alanine aminotransferase cov model



B. Alanine aminotransferase genes model



C. Alanine aminotransferase prs model



D. Alanine aminotransferase genes prs model

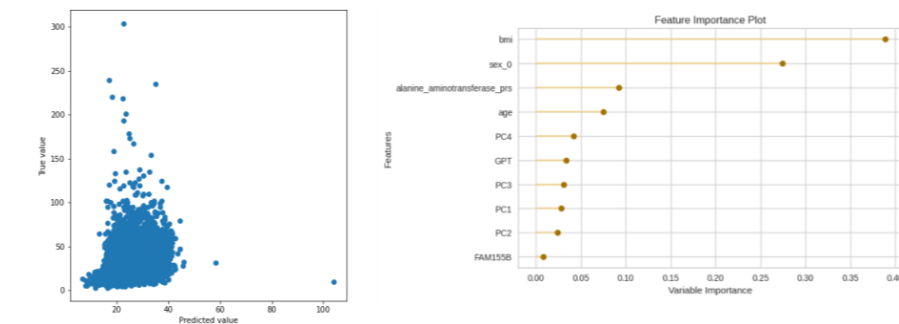
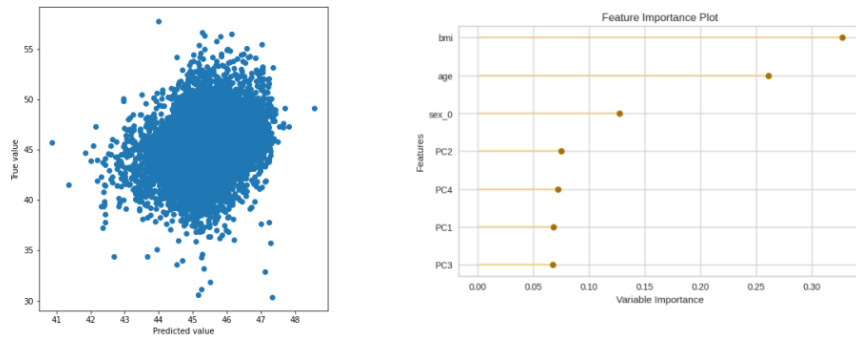
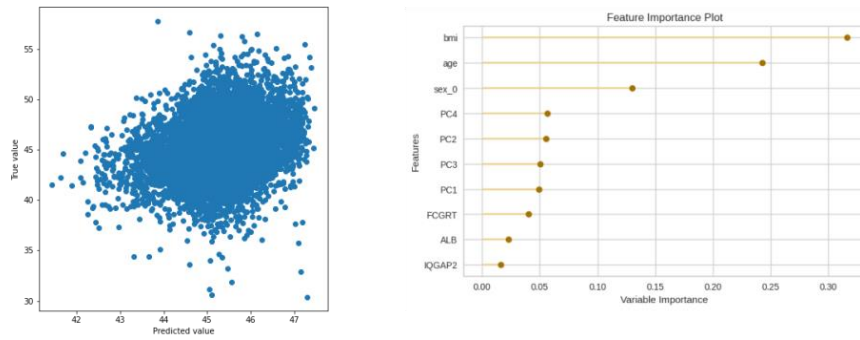


Figure S27 True vs. Predicted value plot (left) and top 10 features (right) for alanine aminotransferase A. covariates model B. genes model C. PRS model and D. combined model.

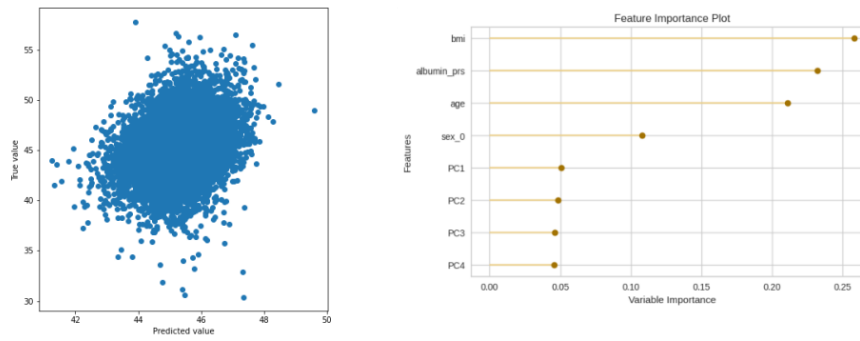
A. Albumin cov model



B. Albumin genes model



C. Albumin prs model



D. Albumin genes prs model

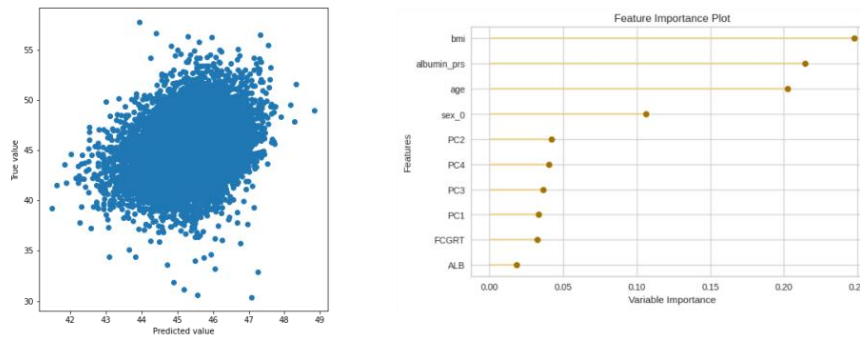
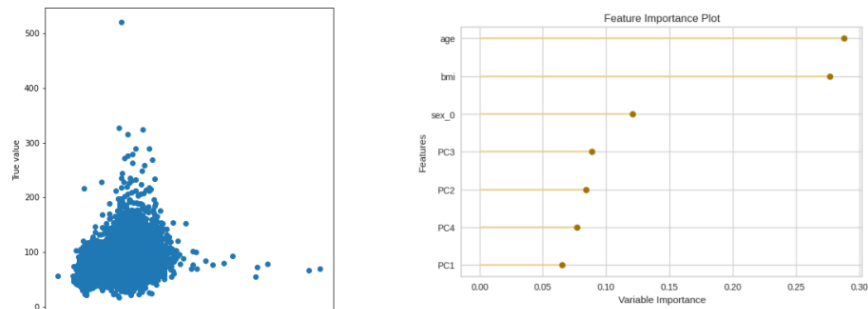
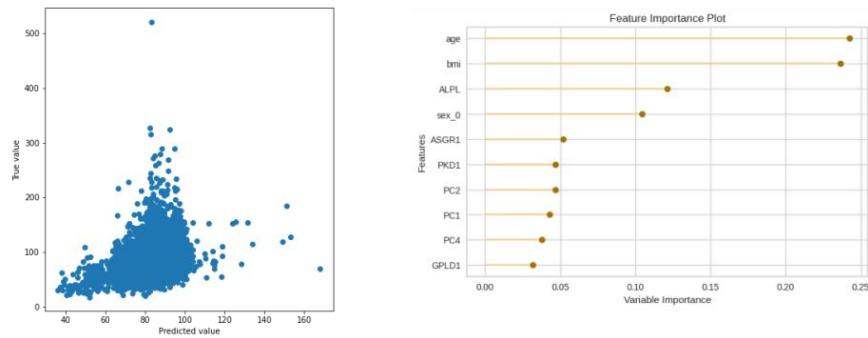


Figure S28 True vs. Predicted value plot (left) and top 10 features (right) for albumin A. covariates model B. genes model C. PRS model and D. combined model.

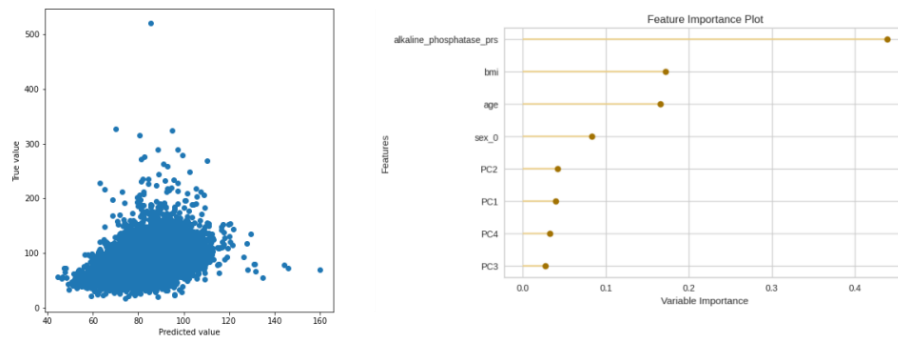
A. Alkaline phosphatase cov model



B. Alkaline phosphatase genes model



C. Alkaline phosphatase prs model



D. Alkaline phosphatase genes prs model

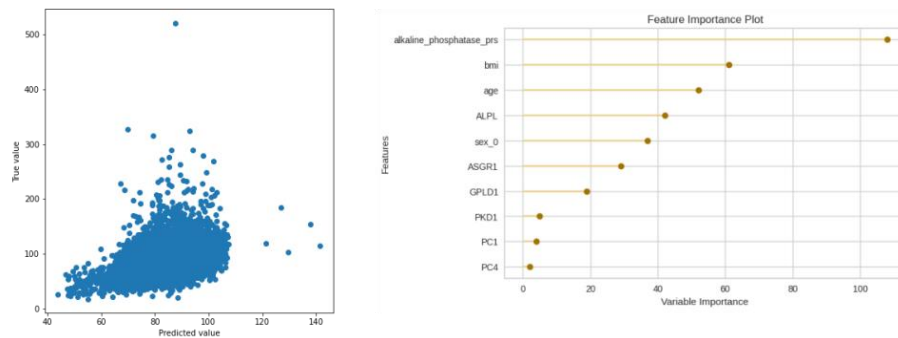
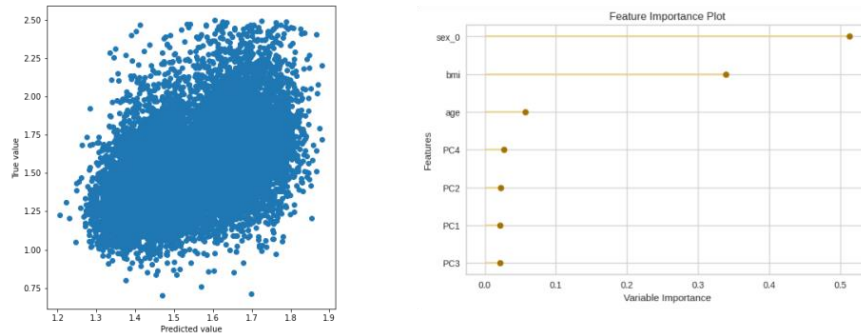
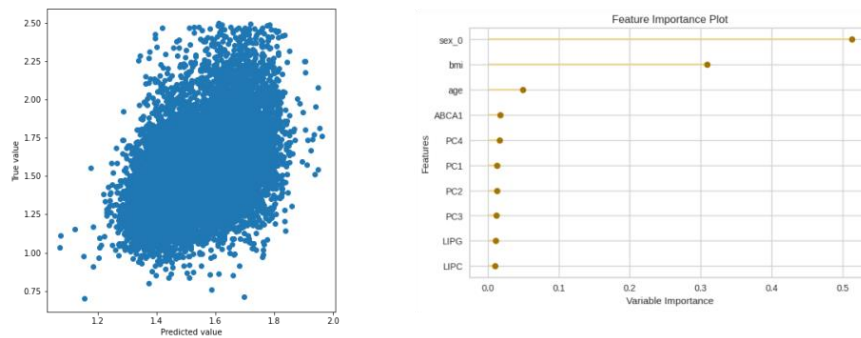


Figure S29 True vs. Predicted value plot (left) and top 10 features (right) for alkaline phosphatase A. covariates model B. genes model C. PRS model and D. combined model.

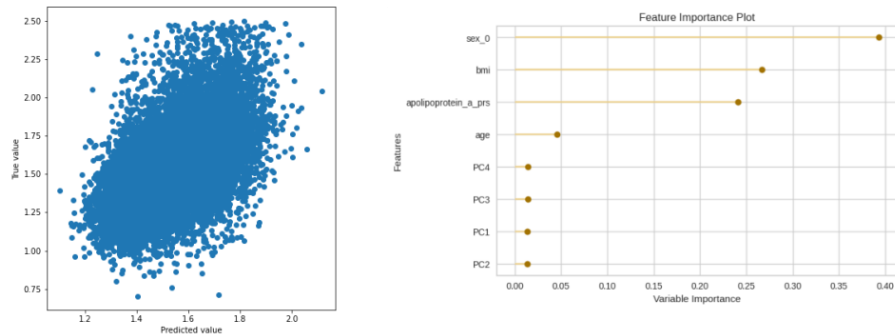
A. Apolipoprotein a cov model



B. Apolipoprotein a genes model



C. Apolipoprotein a prs model



D. Apolipoprotein a genes prs model

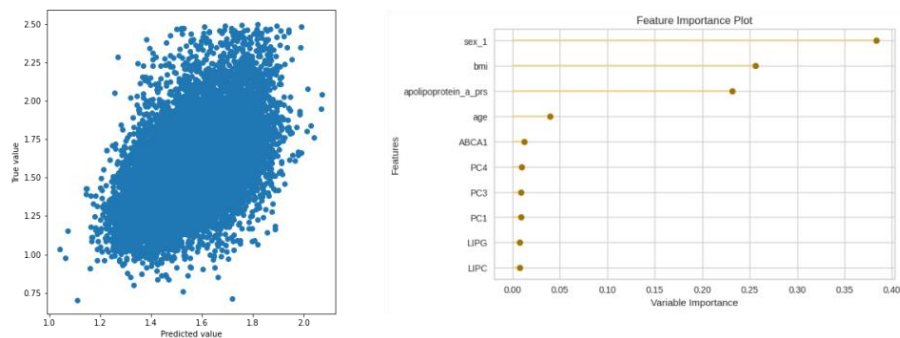
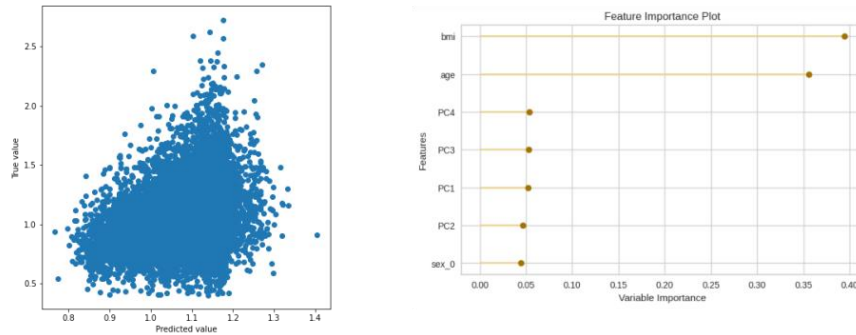
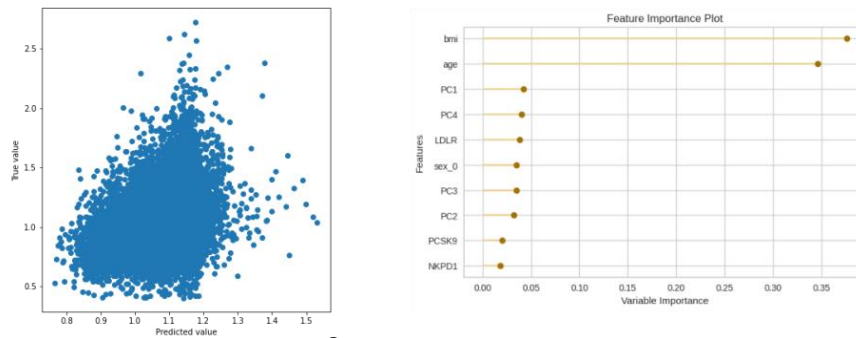


Figure S30 True vs. Predicted value plot (left) and top 10 features (right) for apolipoprotein A A. covariates model B. genes model C. PRS model and D. combined model.

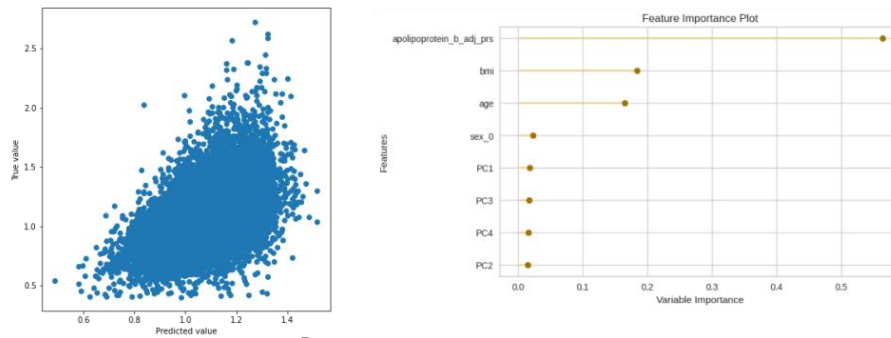
A. Apolipoprotein b * cov model



B. Apolipoprotein b * genes model



C. Apolipoprotein b * prs model



D. Apolipoprotein b * genes prs model

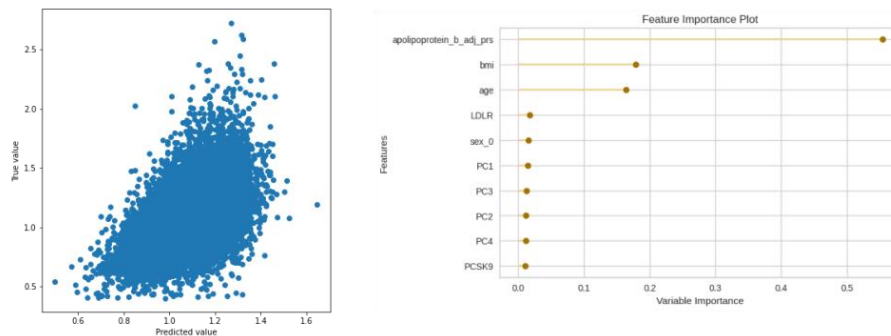
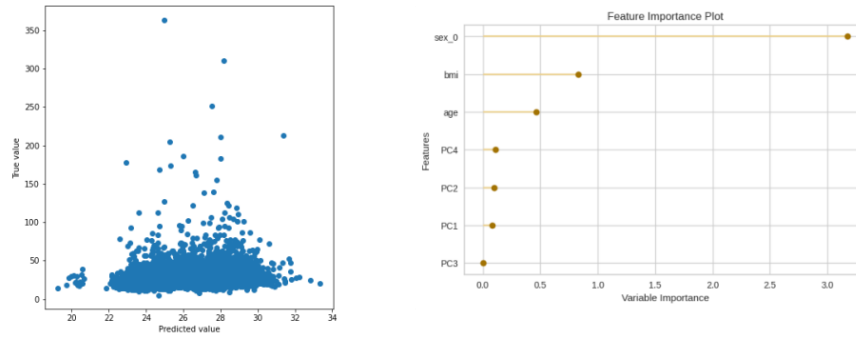


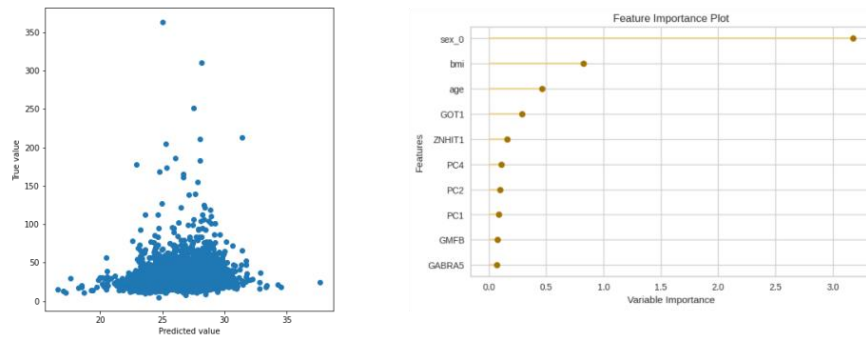
Figure S31 True vs. Predicted value plot (left) and top 10 features (right) for apolipoprotein B* A. covariates model B. genes model C. PRS model and D. combined model.

* statin adjusted values

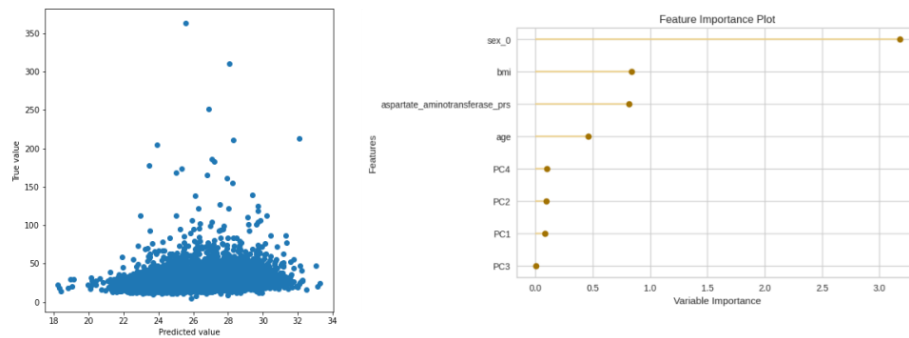
A. Aspartate aminotransferase cov model



B. Aspartate aminotransferase genes model



C. Aspartate aminotransferase prs model



D. Aspartate aminotransferase genes prs model

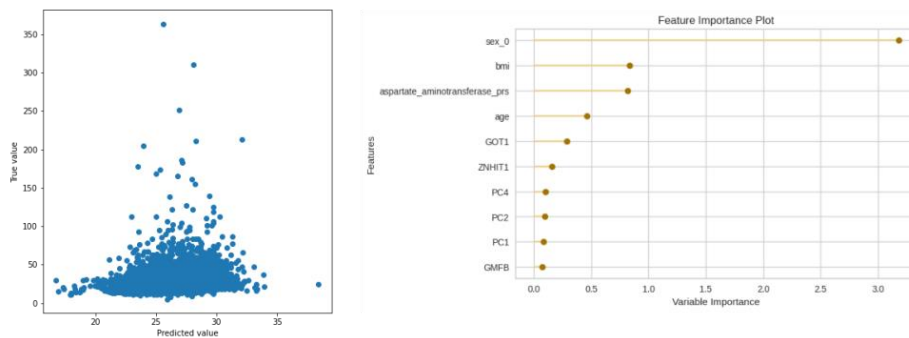
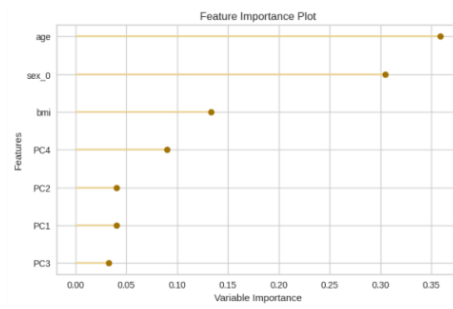
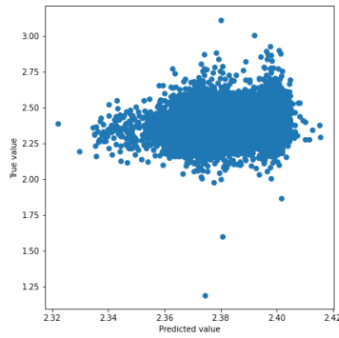
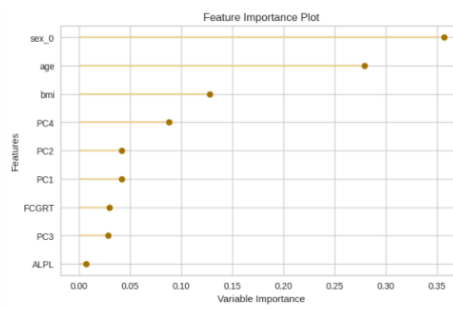
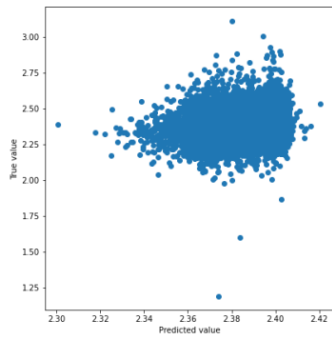


Figure S32 True vs. Predicted value plot (left) and top 10 features (right) for aspartate aminotransferase A. covariates model B. genes model C. PRS model and D. combined model.

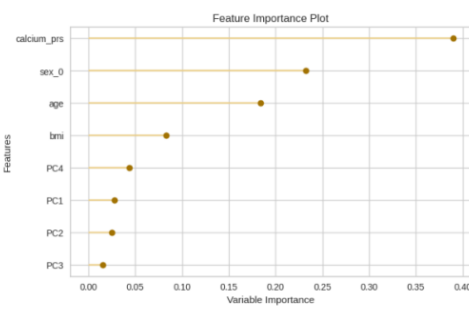
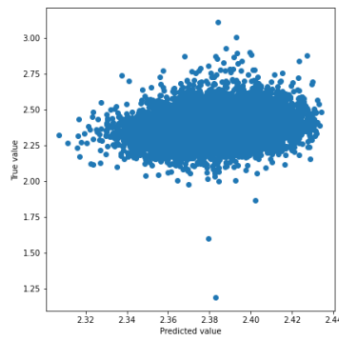
A. Calcium cov model



B. Calcium genes model



C. Calcium prs model



D. Calcium genes prs model

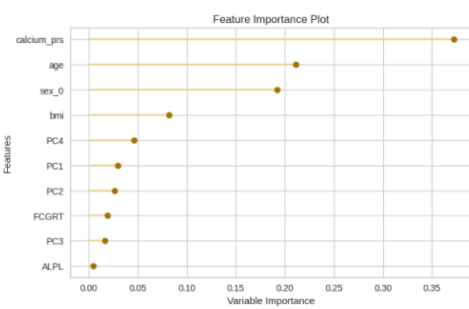
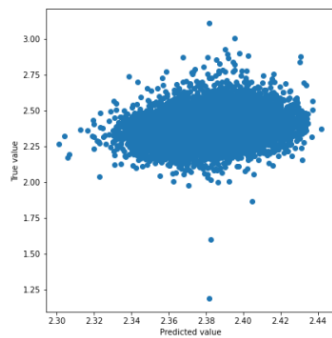
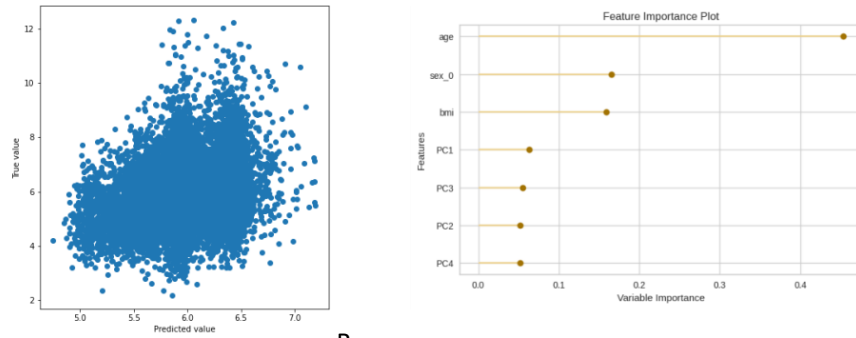
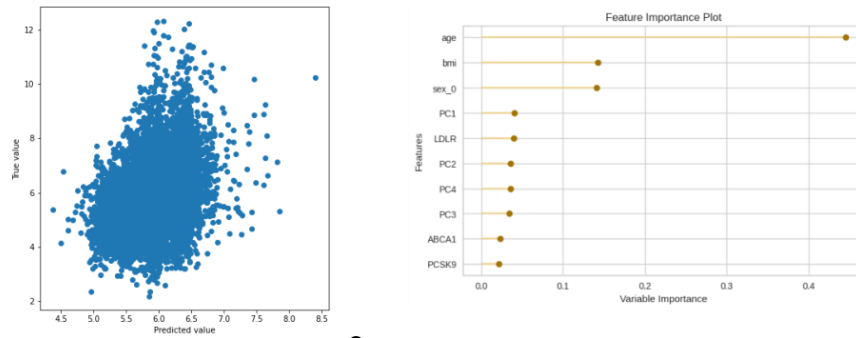


Figure S33 True vs. Predicted value plot (left) and top 10 features (right) for calcium A. covariates model B. genes model C. PRS model and D. combined model.

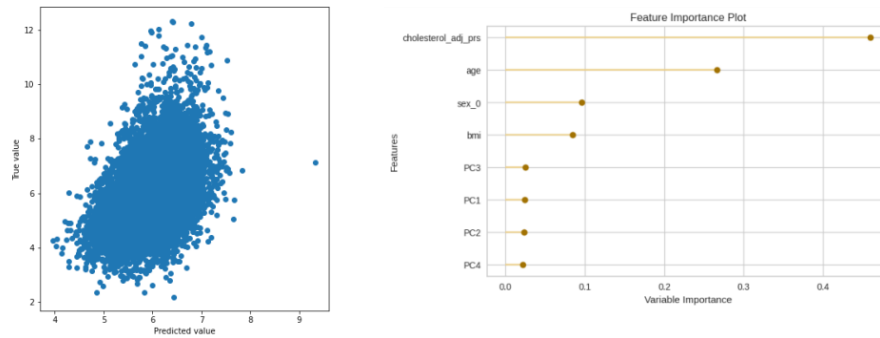
A. Cholesterol * cov model



B. Cholesterol * genes model



C. Cholesterol * prs model



D. Cholesterol * genes prs model

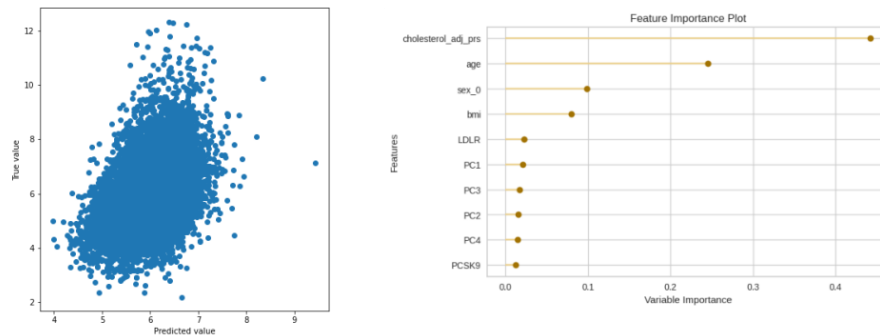
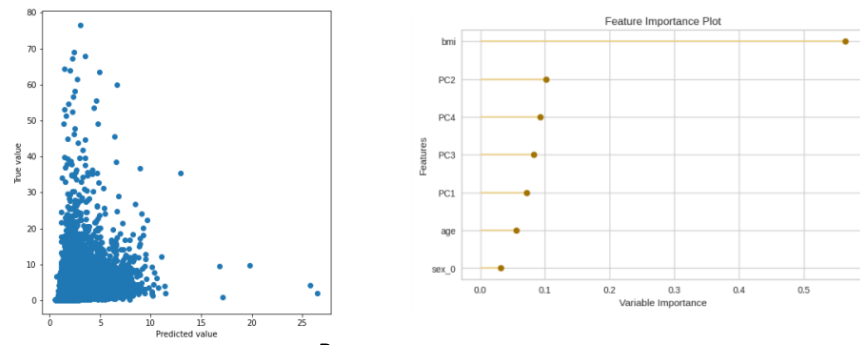
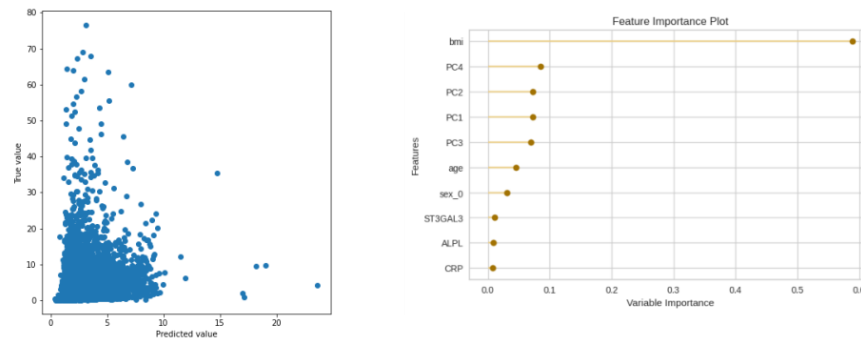


Figure S 34 True vs. Predicted value plot (left) and top 10 features (right) for cholesterol* A. covariates model B. genes model C. PRS model and D. combined model.
* statin adjusted values

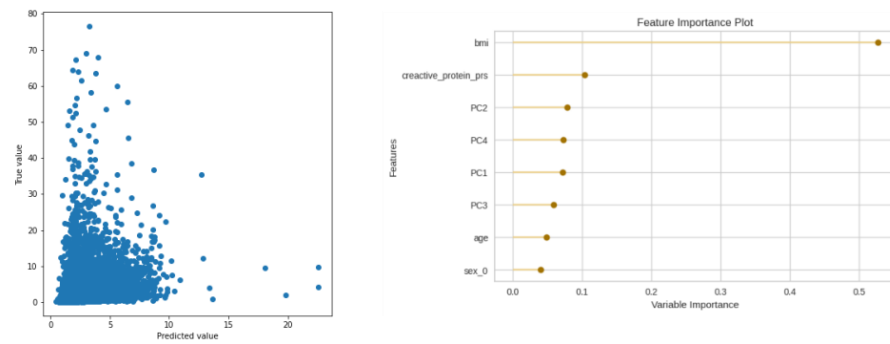
A. Creative protein cov model



B. Creative protein genes model



C. Creative protein prs model



D. Creative protein genes prs model

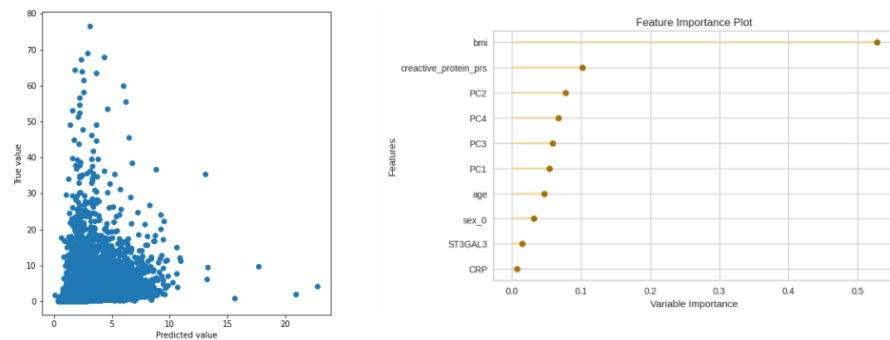
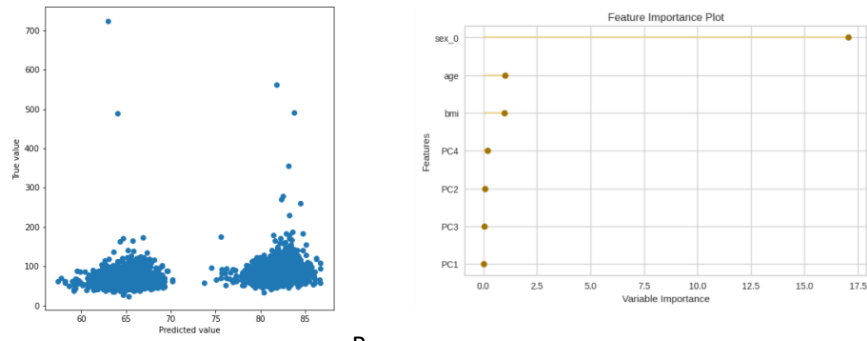
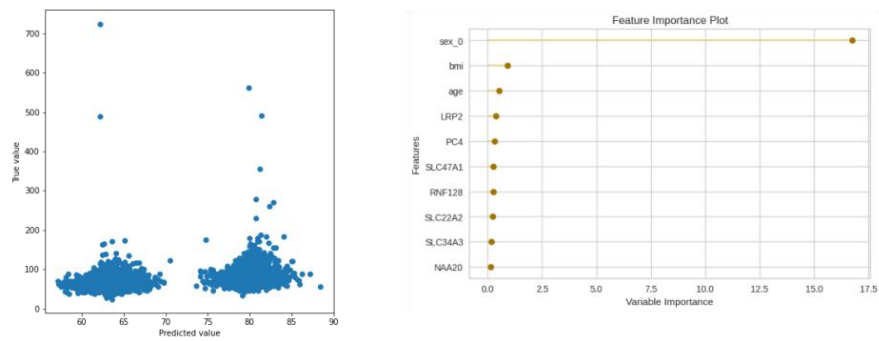


Figure S35 True vs. Predicted value plot (left) and top 10 features (right) for C reactive protein A. covariates model B. genes model C. PRS model and D. combined model.

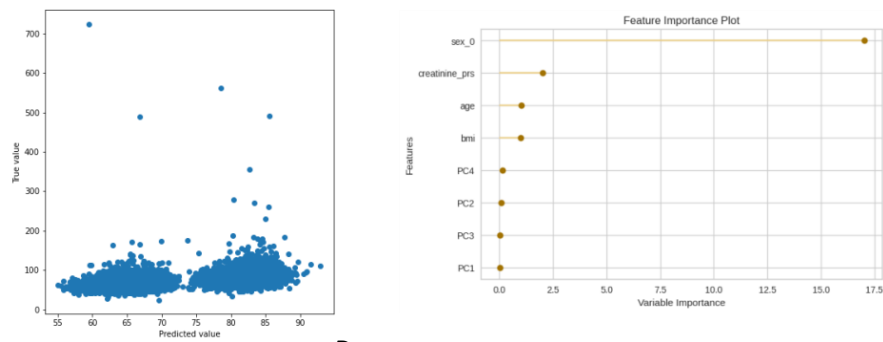
A. Creatinine cov model



B. Creatinine genes model



C. Creatinine prs model



D. Creatinine combined model

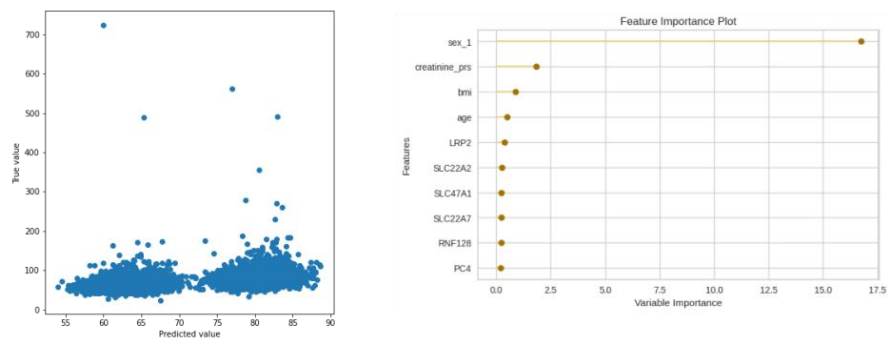
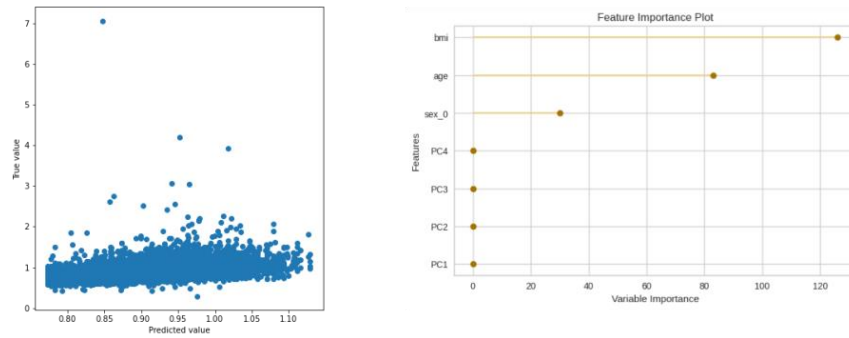
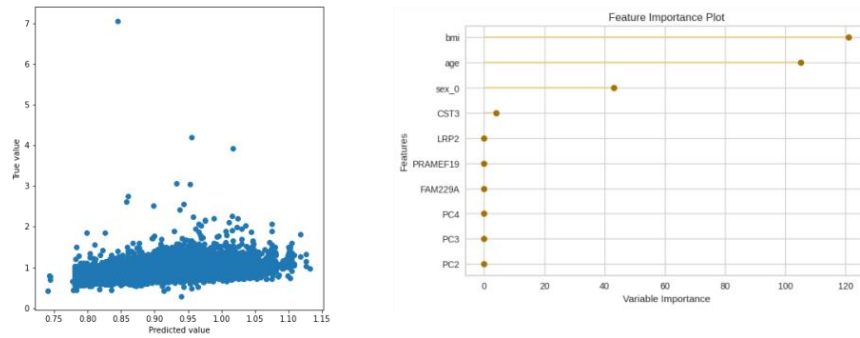


Figure S36 True vs. Predicted value plot (left) and top 10 features (right) for creatinine A. covariates model B. genes model C. PRS model and D. combined model.

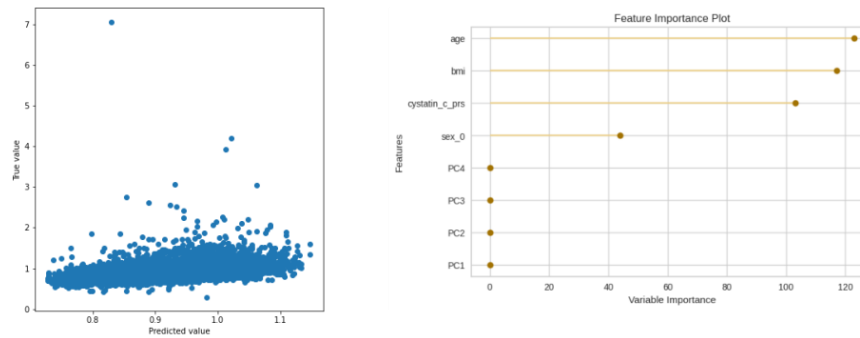
A. Cystatin c cov model



B. Cystatin c genes model



C. Cystatin c prs model



D. Cystatin c genes prs model

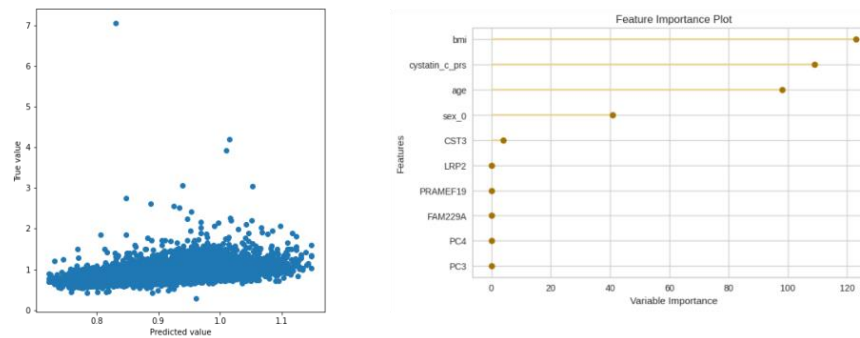
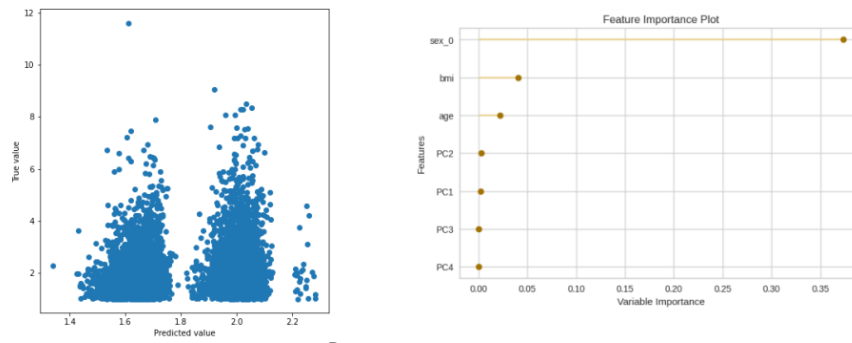
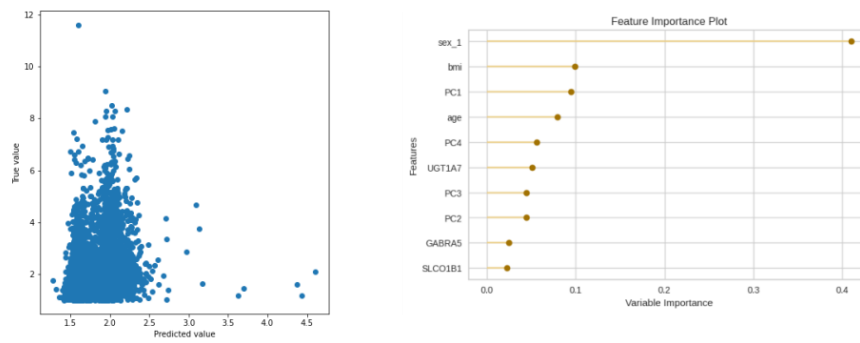


Figure S37 True vs. Predicted value plot (left) and top 10 features (right) for cystatin C A. covariates model B. genes model C. PRS model and D. combined model.

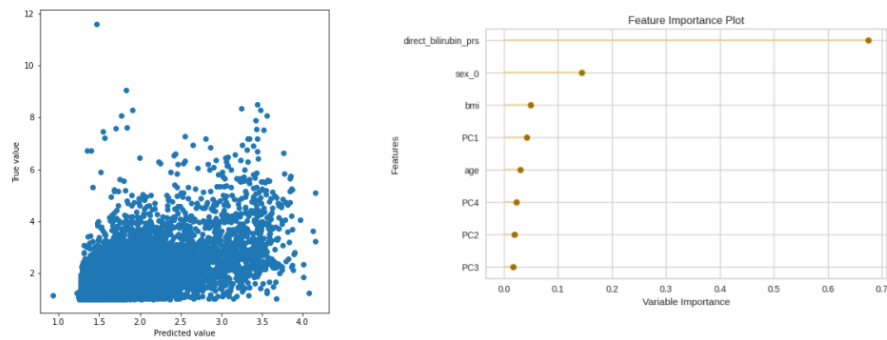
A. Direct bilirubin cov model



B. Direct bilirubin genes model



C. Direct bilirubin prs model



D. Direct bilirubin genes prs model

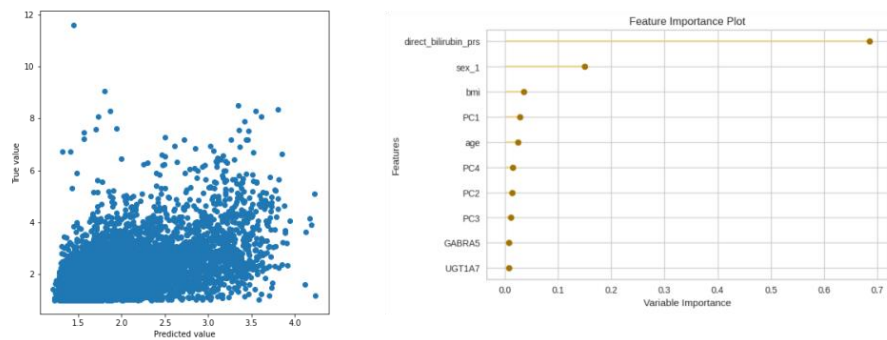
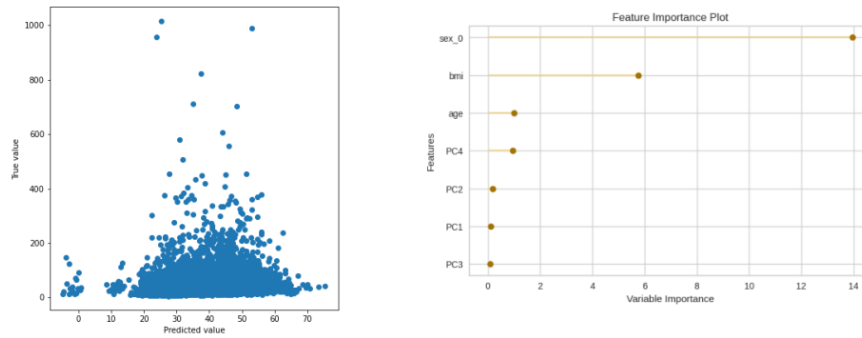
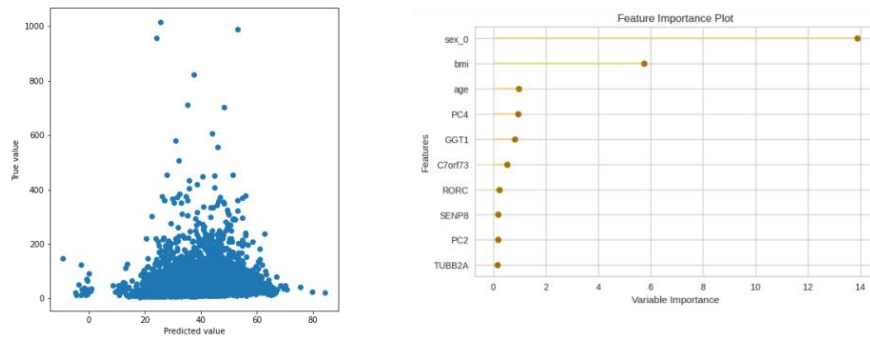


Figure S38 True vs. Predicted value plot (left) and top 10 features (right) for direct bilirubin A. covariates model B. genes model C. PRS model and D. combined model.

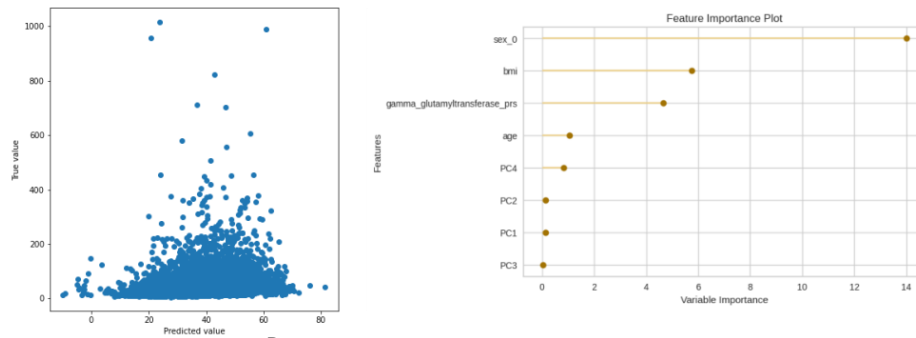
A. Gamma glutamyltransferase cov model



B. Gamma glutamyltransferase genes model



C. Gamma glutamyltransferase prs model



D. Gamma glutamyltransferase genes prs model

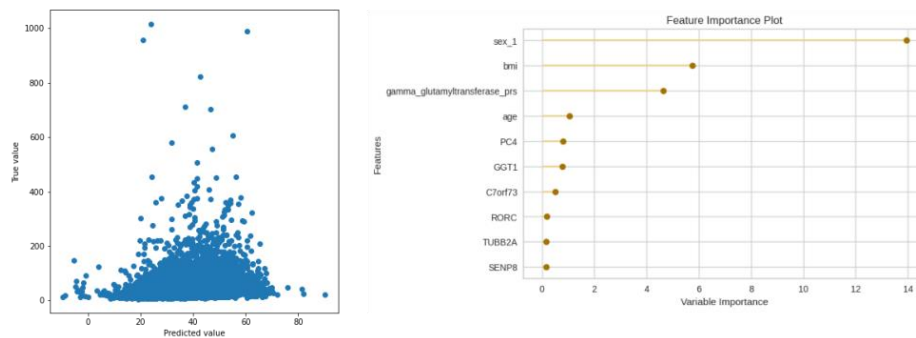
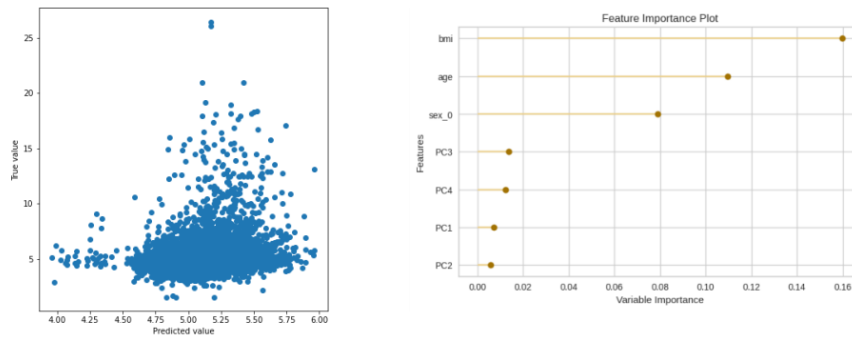
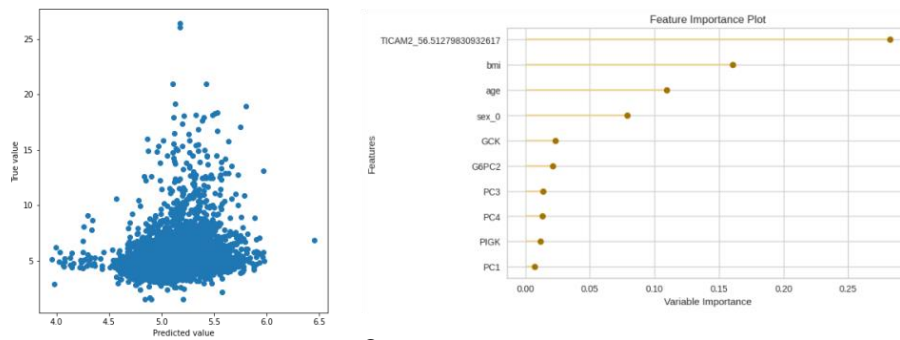


Figure S39 True vs. Predicted value plot (left) and top 10 features (right) for gamma glutamyltransferase A. covariates model B. genes model C. PRS model and D. combined model.

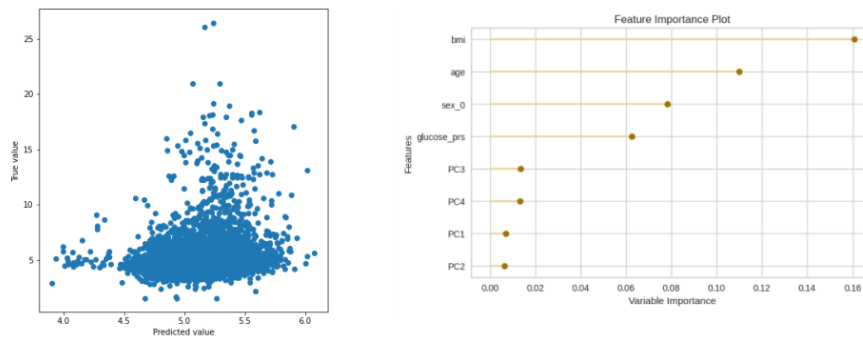
A. Glucose cov model



B. Glucose genes model



C. Glucose prs model



D. Glucose genes prs model

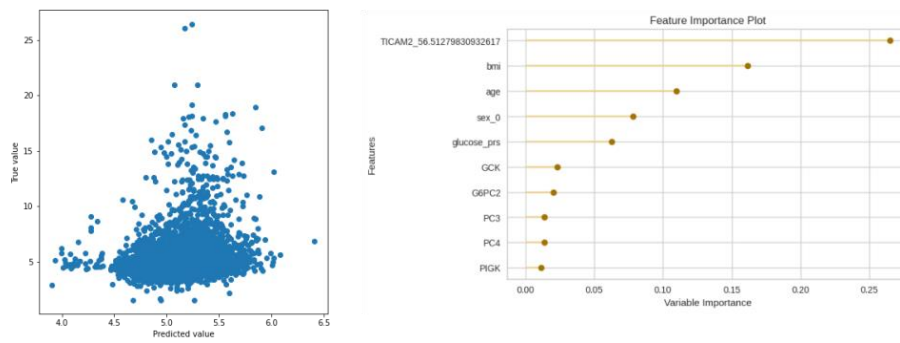
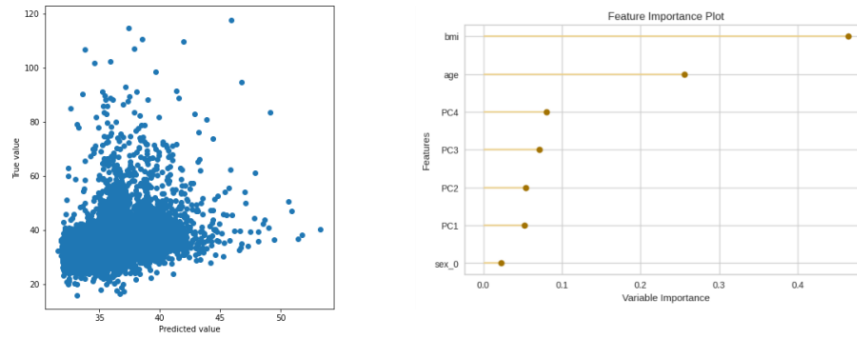
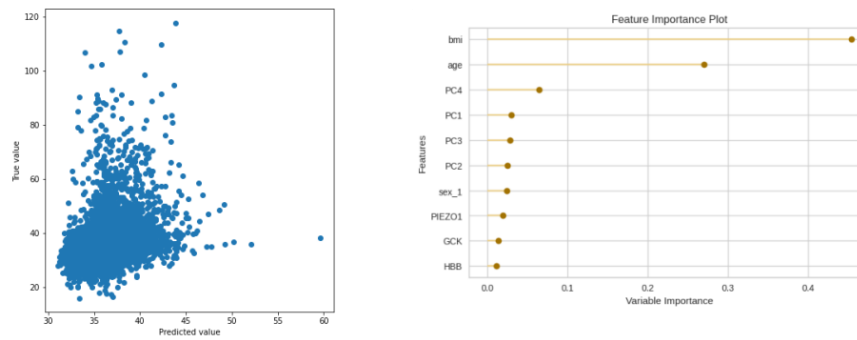


Figure S40 True vs. Predicted value plot (left) and top 10 features (right) for glucose A. covariates model B. genes model C. PRS model and D. combined model.

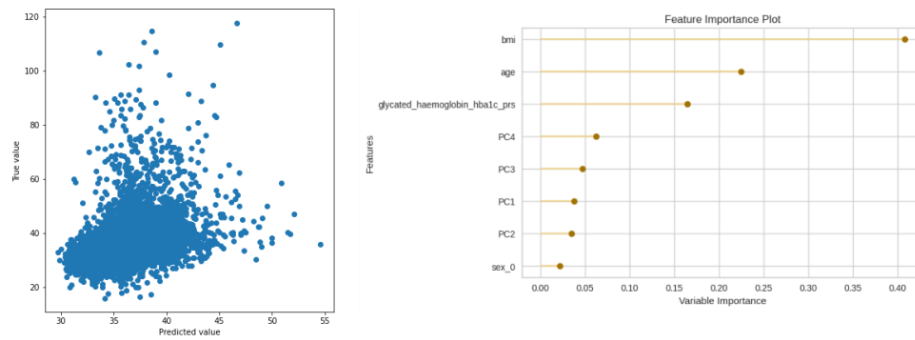
A. Glycated haemoglobin (hba1c) cov model



B. Glycated haemoglobin (hba1c) genes model



C. Glycated haemoglobin (hba1c) prs model



D. Glycated haemoglobin (hba1c) genes prs model

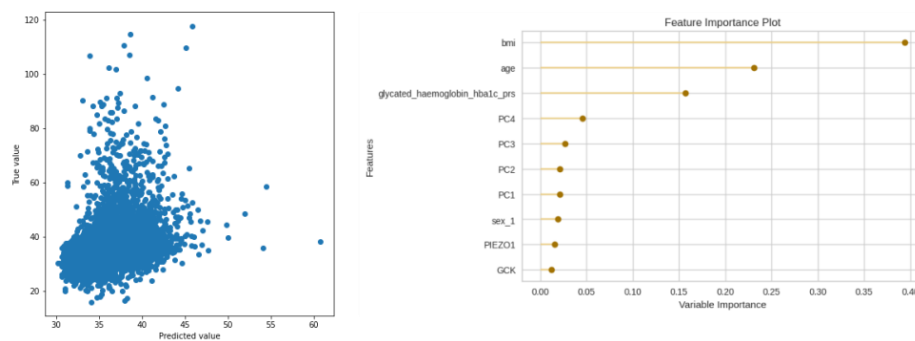
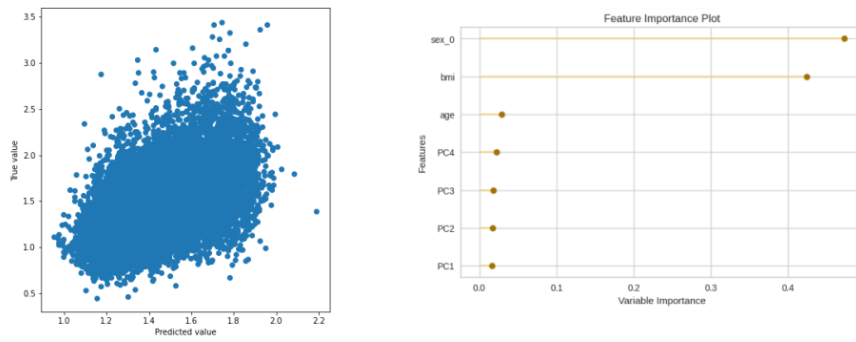
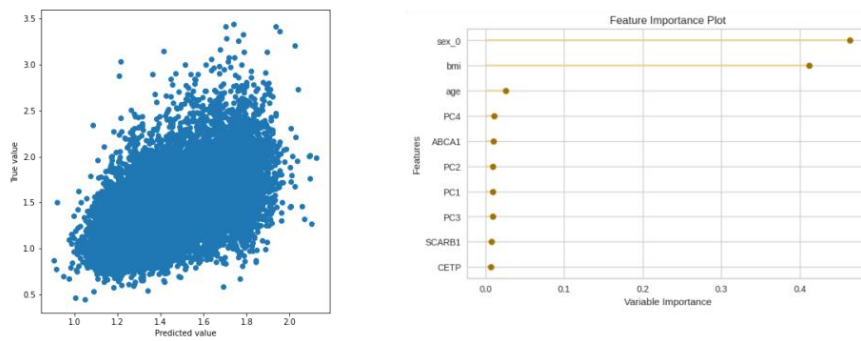


Figure S41 True vs. Predicted value plot (left) and top 10 features (right) for glycated haemoglobin (HbA1c) A. covariates model B. genes model C. PRS model and D. combined model.

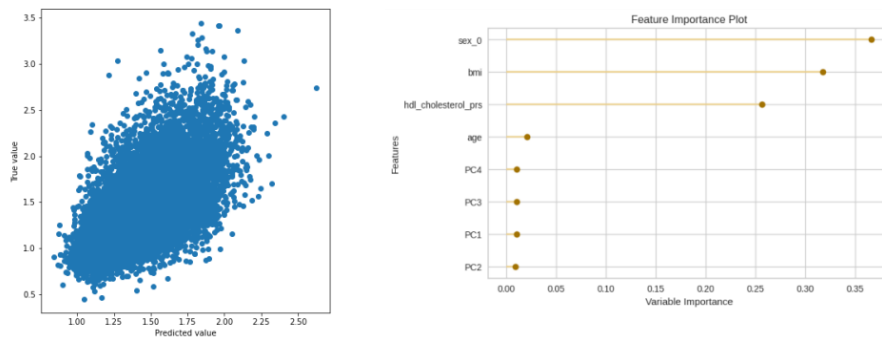
A. Hdl cholesterol cov model



B. Hdl cholesterol genes model



C. Hdl cholesterol prs model



D. Hdl cholesterol genes prs model

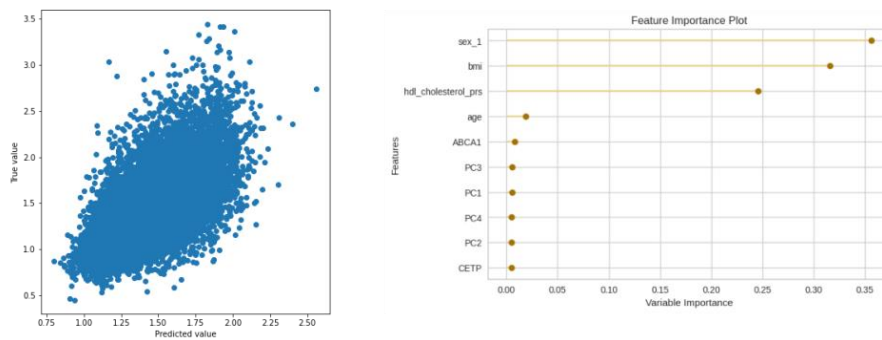
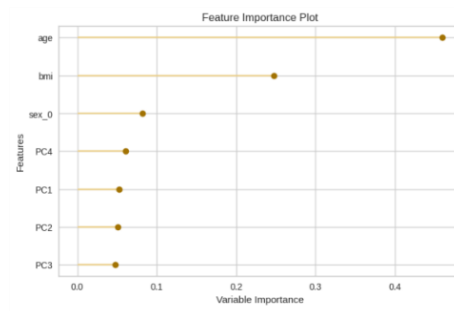
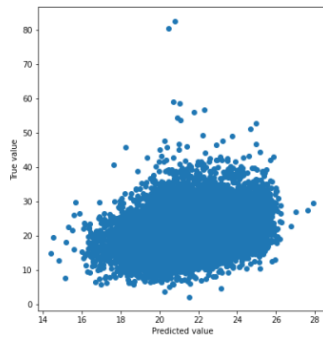
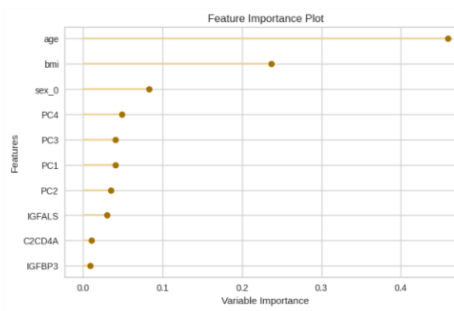
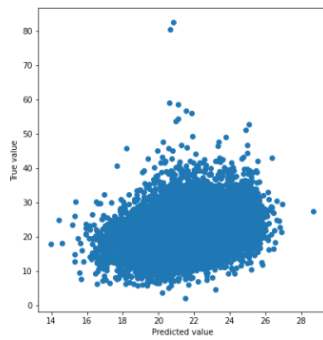


Figure S42 True vs. Predicted value plot (left) and top 10 features (right) for HDL cholesterol A. covariates model B. genes model C. PRS model and D. combined model.

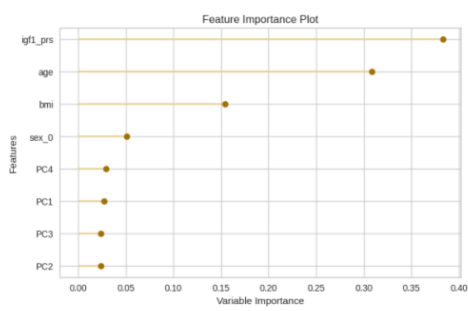
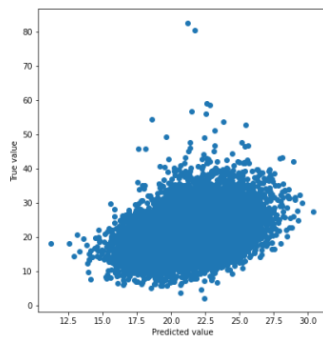
A. Igf1 cov model



B. Igf1 genes model



C. Igf1 prs model



D. Igf1 genes prs model

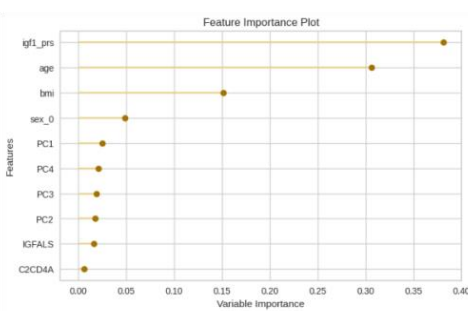
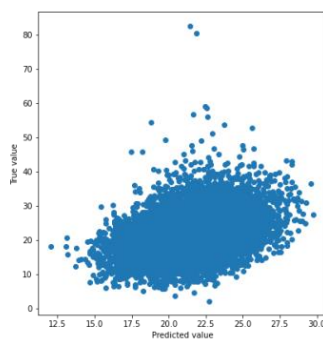
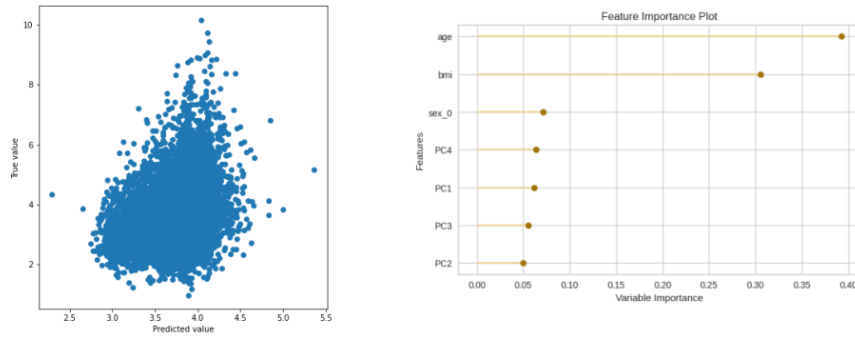
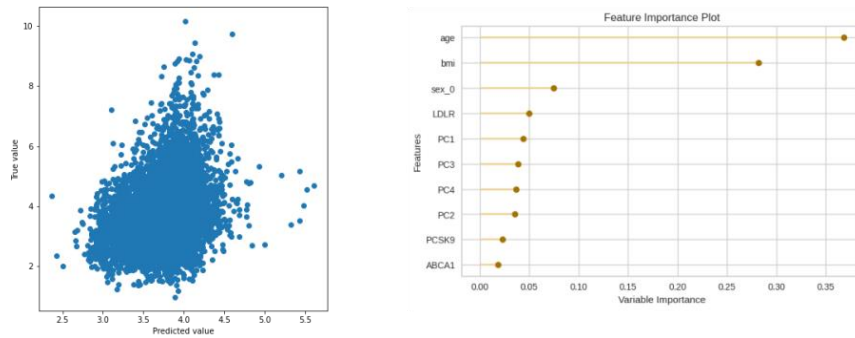


Figure S43 True vs. Predicted value plot (left) and top 10 features (right) for IGF1 A. covariates model B. genes model C. PRS model and D. combined model.

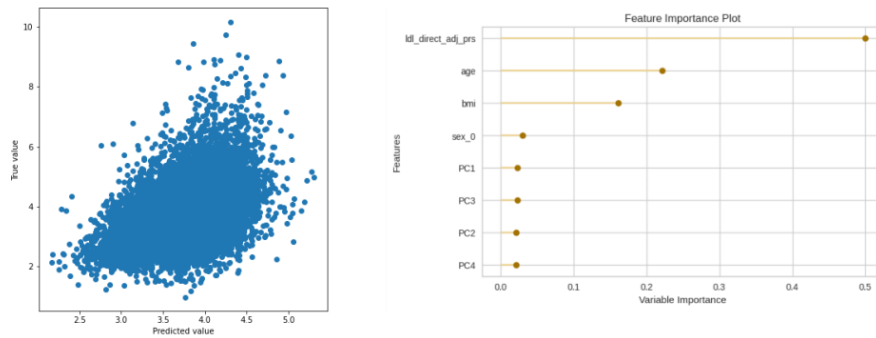
A. Ldl direct * cov model



B. Ldl direct * genes model



C. Ldl direct * prs model



D. Ldl direct * genes prs model

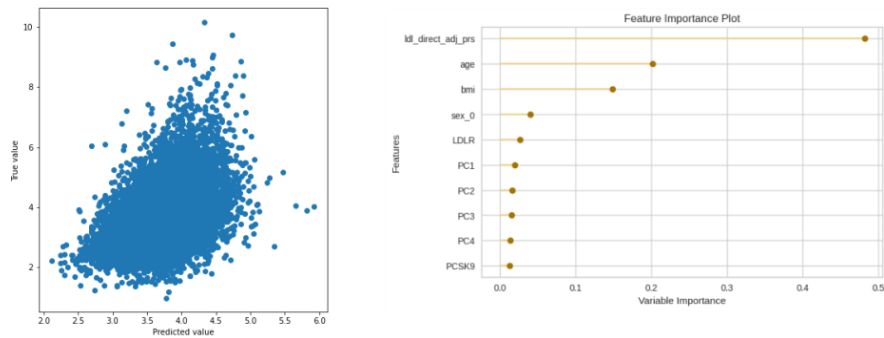
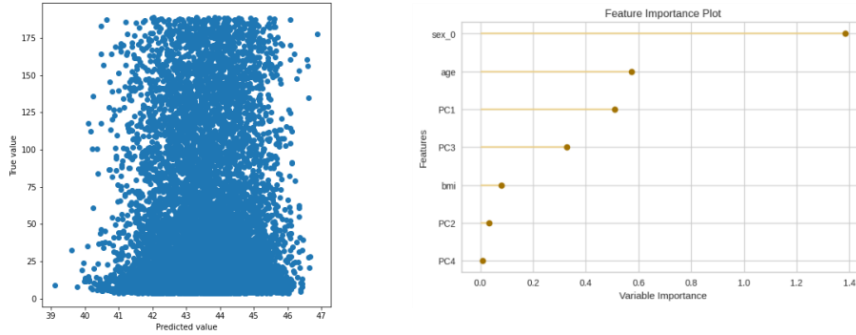
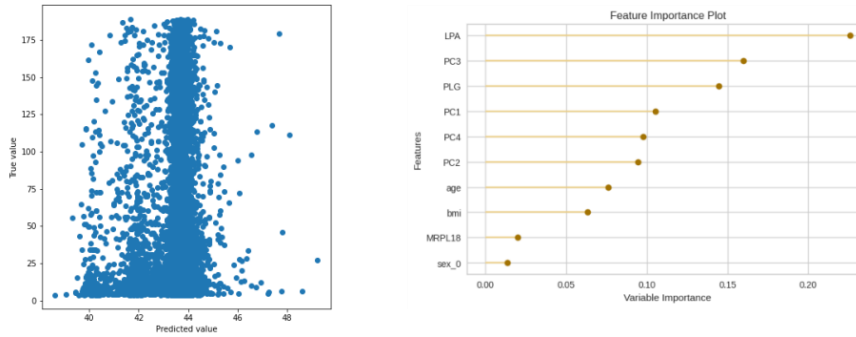


Figure S44 True vs. Predicted value plot (left) and top 10 features (right) for LDL direct* A. covariates model B. genes model C. PRS model and D. combined model.
* statin adjusted values

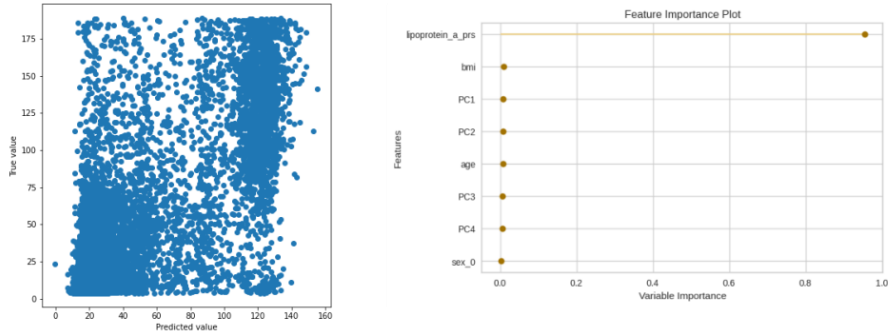
A. Lipoprotein a cov model



B. Lipoprotein a genes model



C. Lipoprotein a prs model



D. Lipoprotein a genes prs model

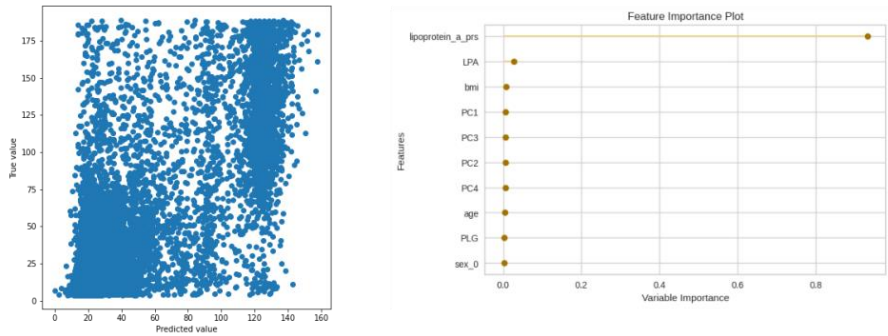
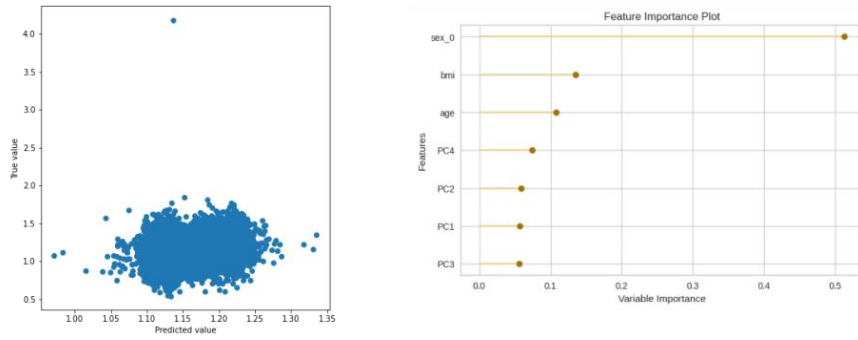
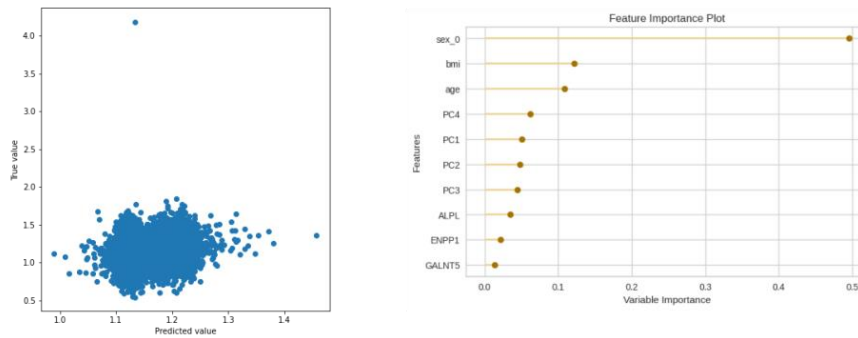


Figure S45 True vs. Predicted value plot (left) and top 10 features (right) for lipoprotein A A. covariates model B. genes model C. PRS model and D. combined model.

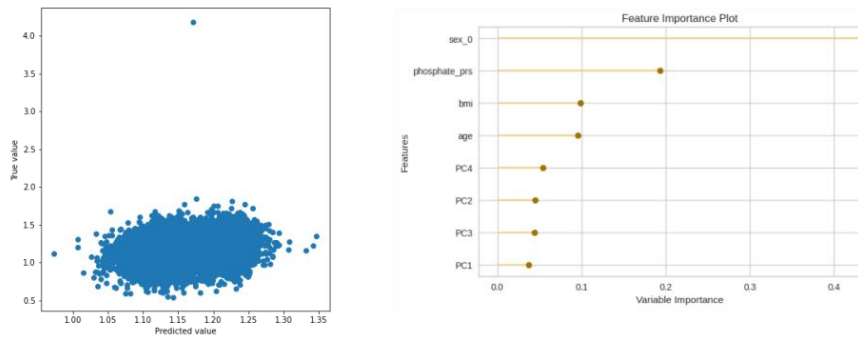
A. Phosphate cov model



B. Phosphate genes model



C. Phosphate prs model



D. Phosphate genes prs model

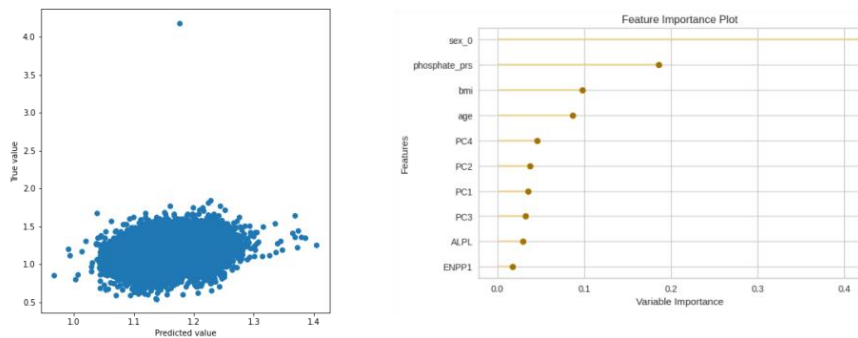
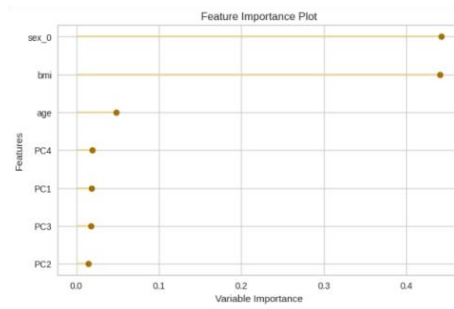
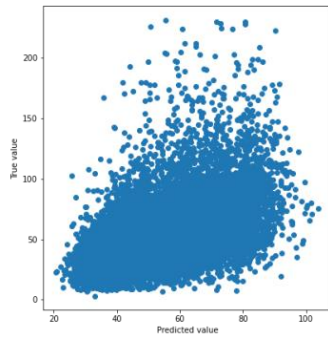
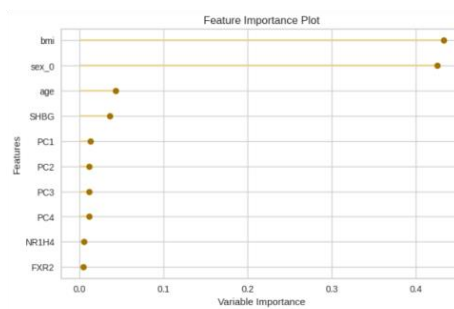
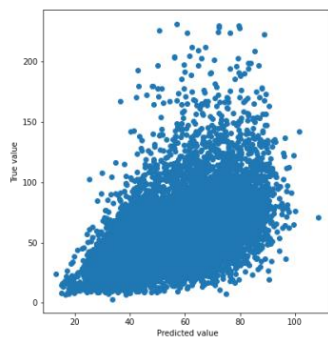


Figure S46 True vs. Predicted value plot (left) and top 10 features (right) for phosphate A. covariates model B. genes model C. PRS model and D. combined model.

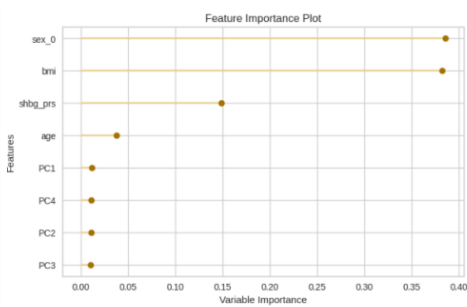
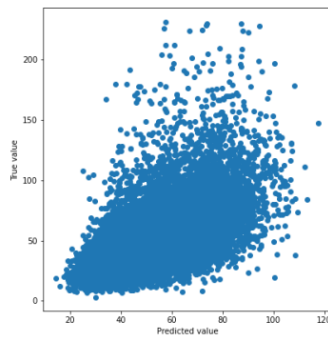
Shbg cov model



Shbg genes model



Shbg prs model



Shbg genes prs model

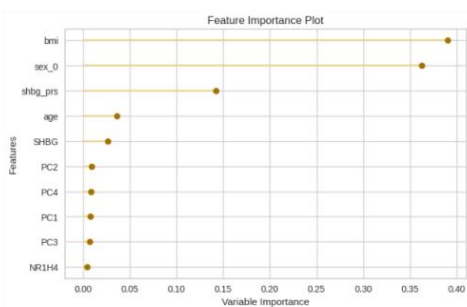
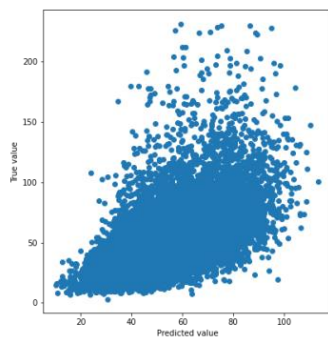
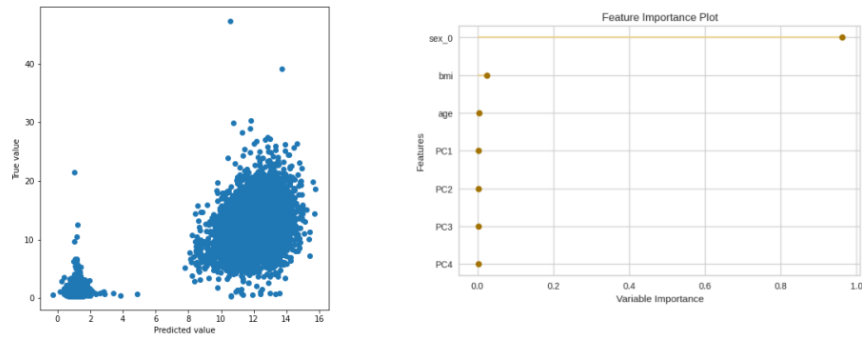
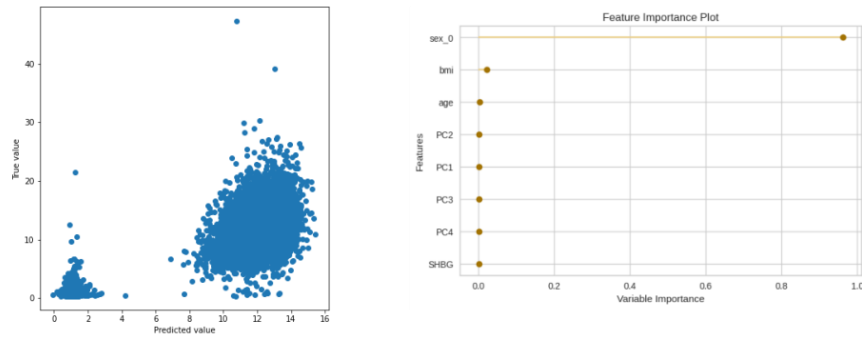


Figure S47 True vs. Predicted value plot (left) and top 10 features (right) for SHBG A. covariates model B. genes model C. PRS model and D. combined model.

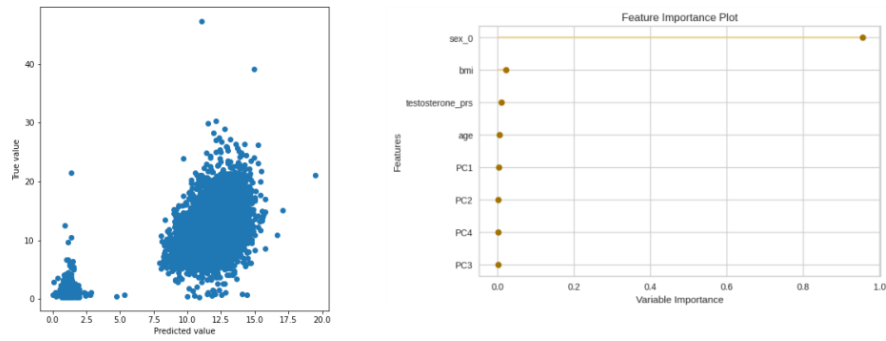
A. Testosterone cov model



B. Testosterone genes model



C. Testosterone prs model



D. Testosterone genes prs model

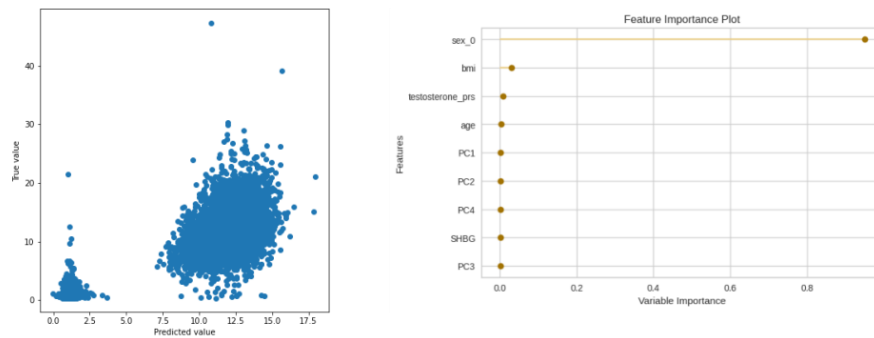
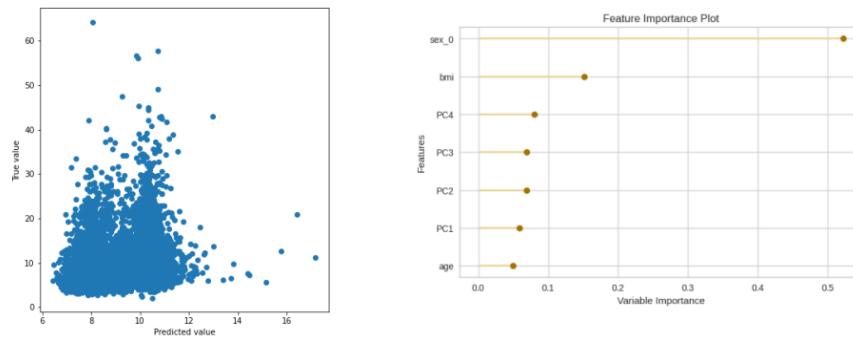
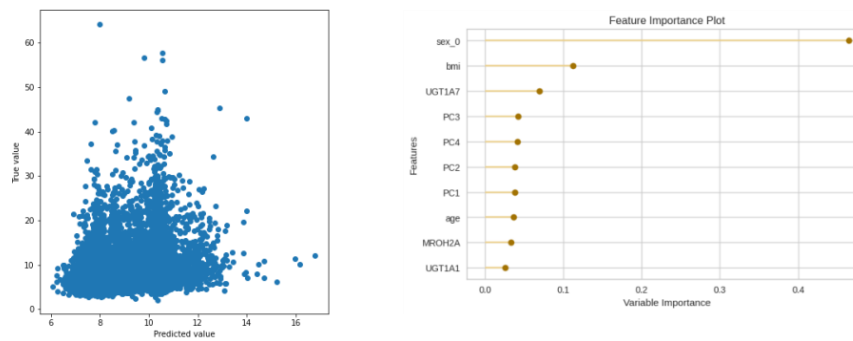


Figure S48 True vs. Predicted value plot (left) and top 10 features (right) for testosterone A. covariates model B. genes model C. PRS model and D. combined model.

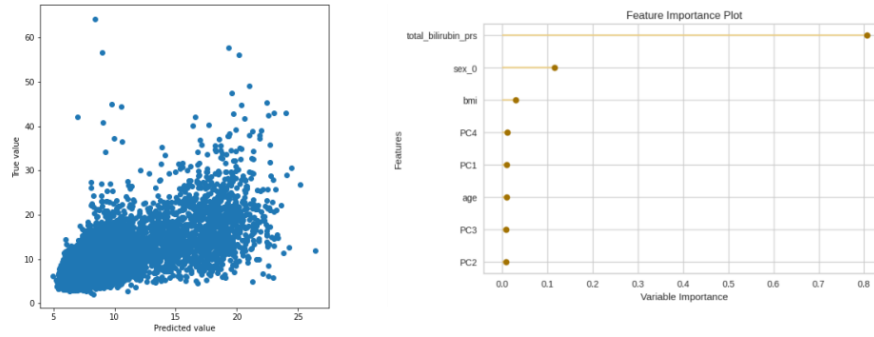
A. Total bilirubin cov model



B. Total bilirubin genes model



C. Total bilirubin prs model



D. Total bilirubin genes prs model

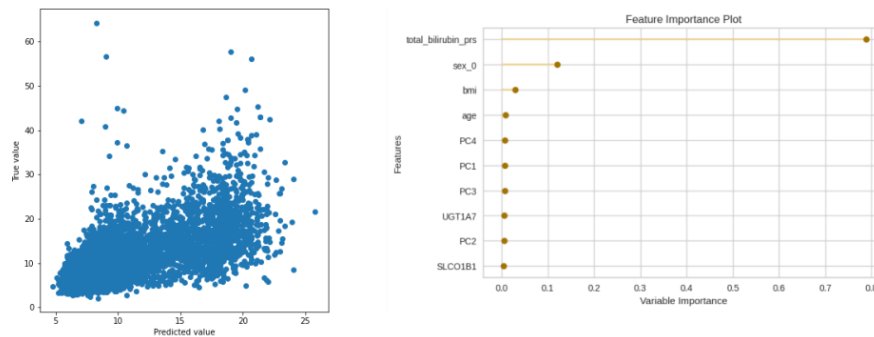
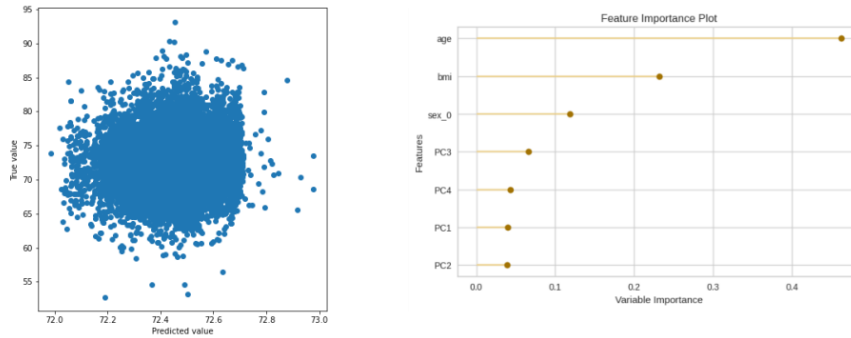
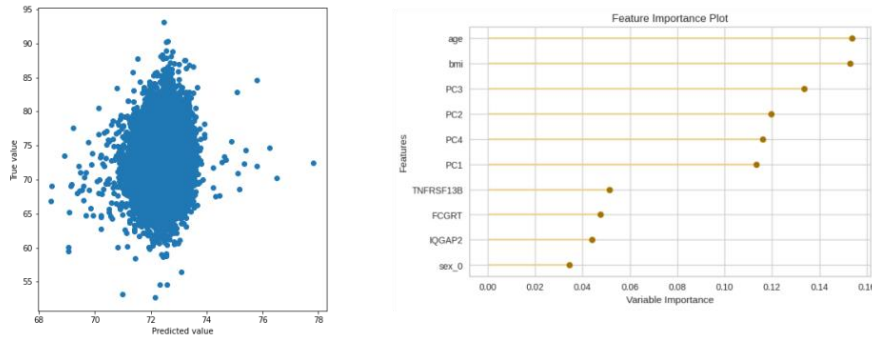


Figure S49 True vs. Predicted value plot (left) and top 10 features (right) for total bilirubin A. covariates model B. genes model C. PRS model and D. combined model.

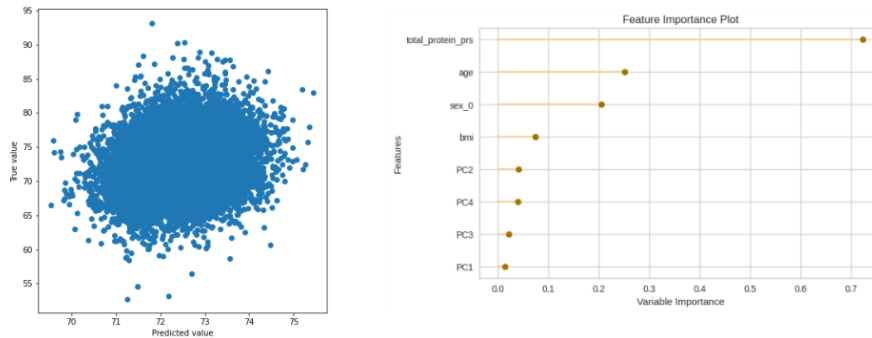
A. Total protein cov model



B. Total protein genes model



C. Total protein prs model



D. Total protein genes prs model

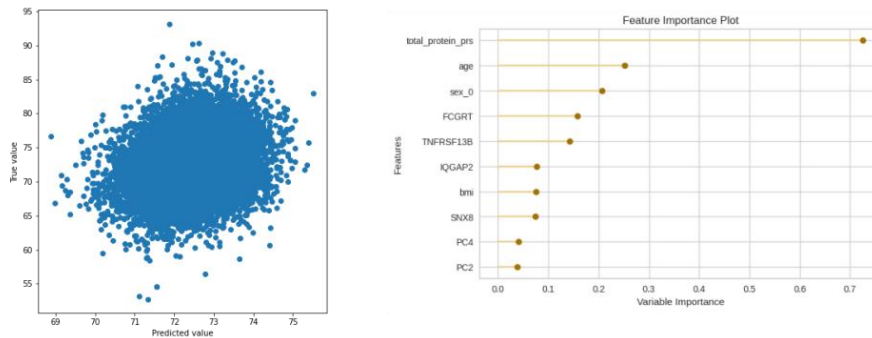
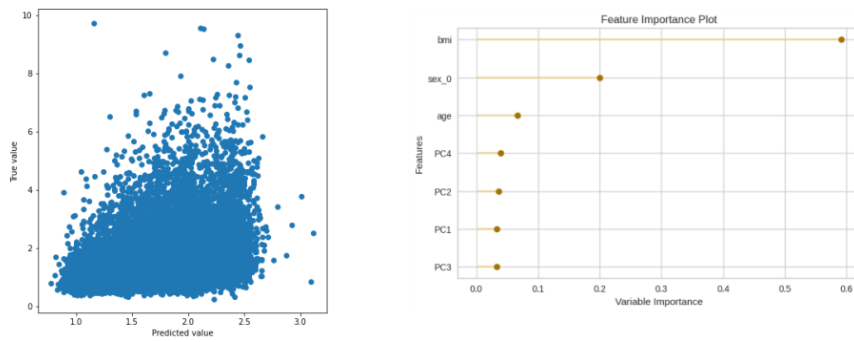
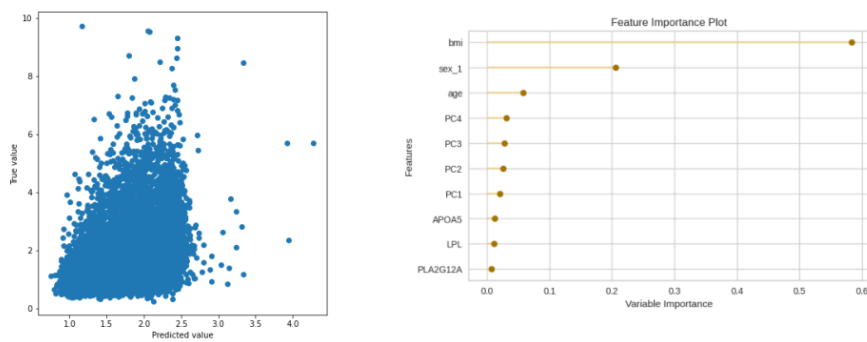


Figure S50 True vs. Predicted value plot (left) and top 10 features (right) for total protein A. covariates model B. genes model C. PRS model and D. combined model.

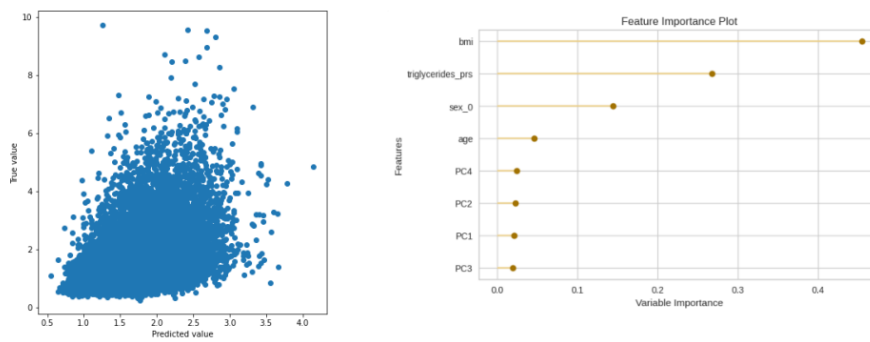
A. Triglycerides cov model



B. Triglycerides genes model



C. Triglycerides prs model



D. Triglycerides genes prs model

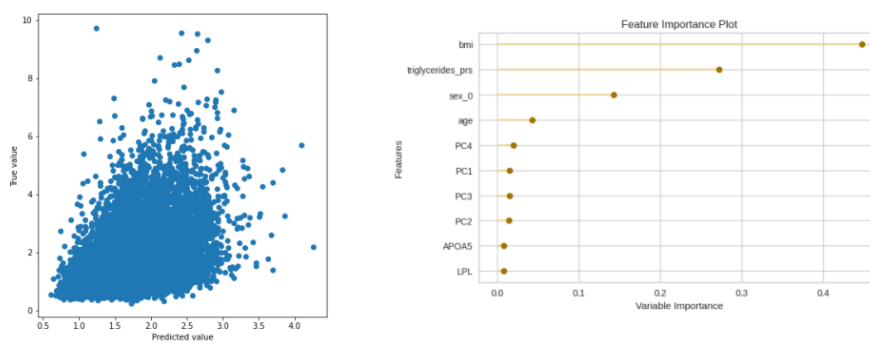
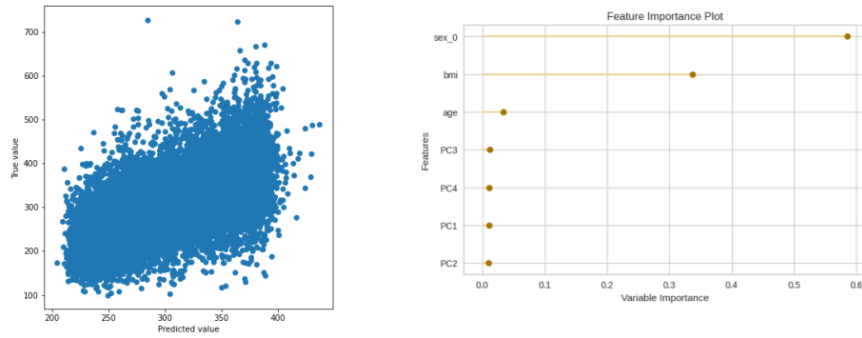
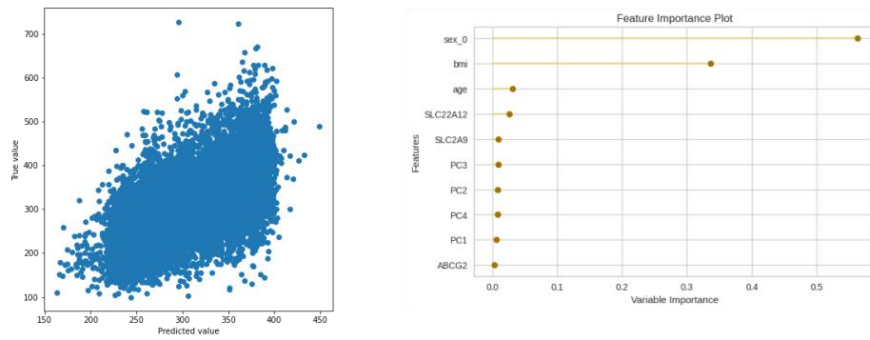


Figure S51 True vs. Predicted value plot (left) and top 10 features (right) for triglycerides A. covariates model B. genes model C. PRS model and D. combined model.

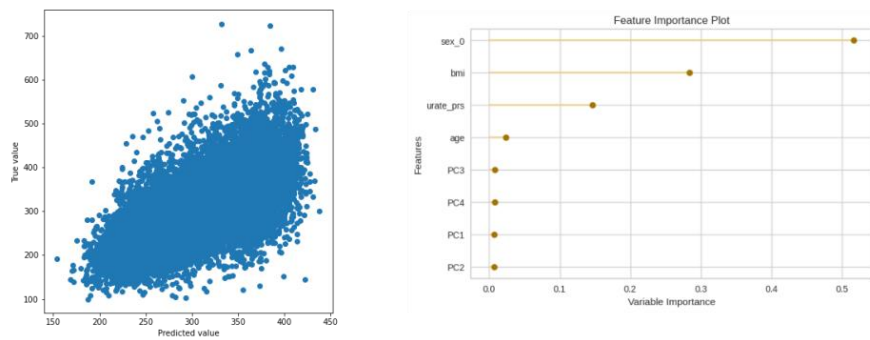
A. Urate cov model



B. Urate genes model



C. Urate prs model



D. Urate genes prs model

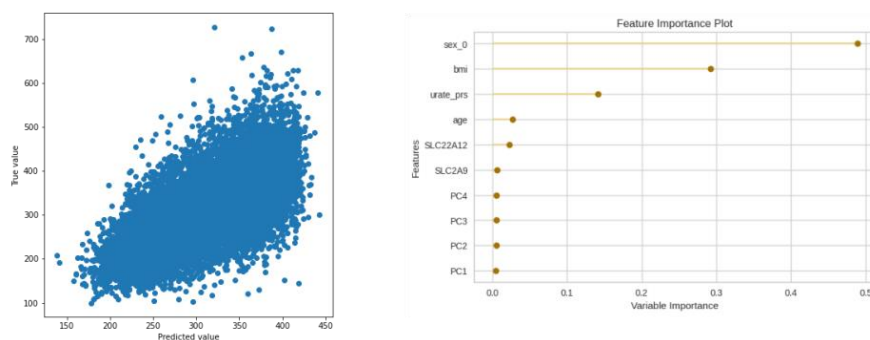
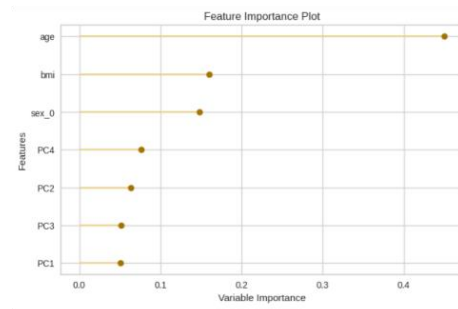
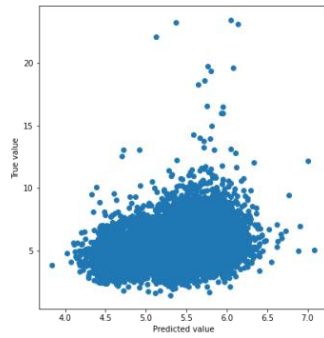
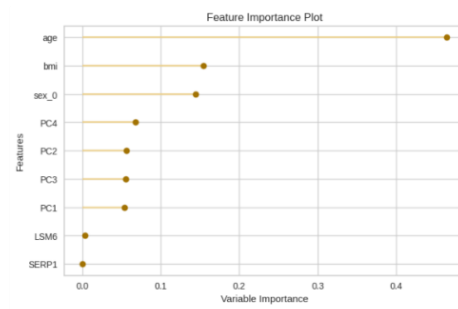
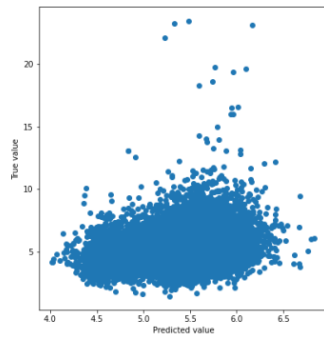


Figure S52 True vs. Predicted value plot (left) and top 10 features (right) for urate A. covariates model B. genes model C. PRS model and D. combined model.

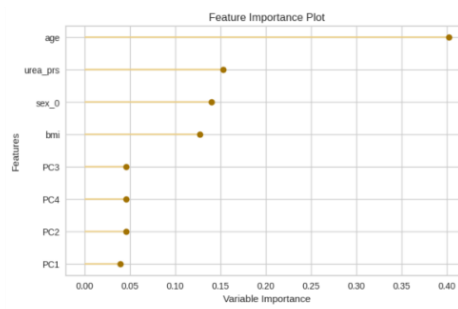
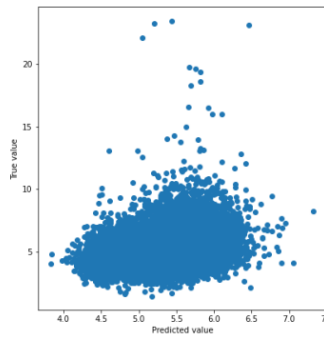
A. Urea cov model



B. Urea genes model



C. Urea prs model



D. Urea genes prs model

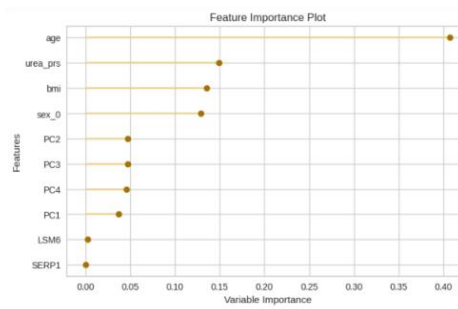
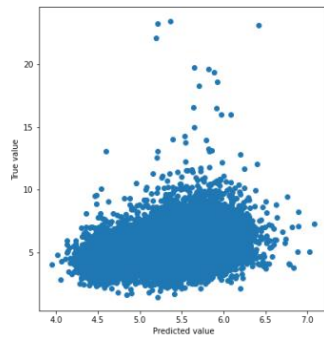
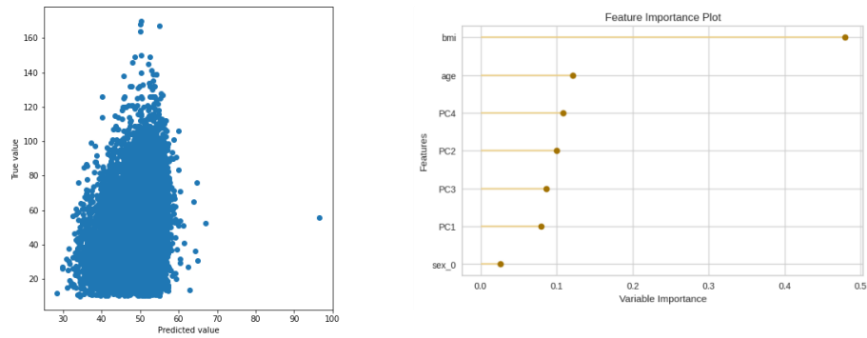
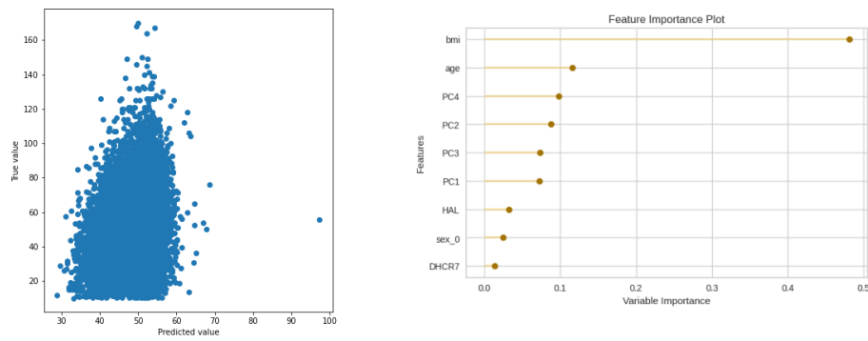


Figure S53 True vs. Predicted value plot (left) and top 10 features (right) for urea A. covariates model B. genes model C. PRS model and D. combined model.

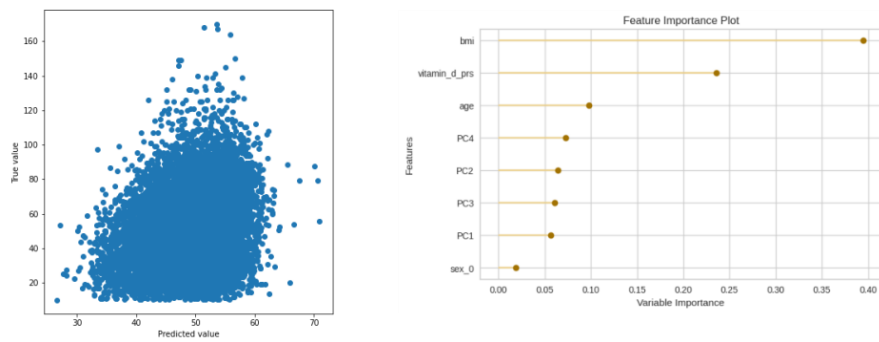
A. Vitamin d cov model



B. Vitamin d genes model



C. Vitamin d prs model



D. vitamin d genes prs model

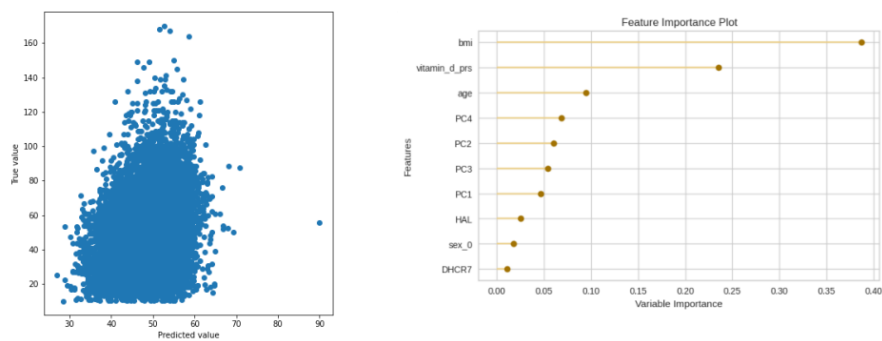


Figure S54 True vs. Predicted value plot (left) and top 10 features (right) for vitamin D A. covariates model B. genes model C. PRS model and D. combined model.

3.3 Publication 3 - Analysis of 72,469 UK Biobank exomes links rare variants to male-pattern hair loss

This publication looks into the genetics of male-pattern hair loss in samples from UK Biobank. Both rare and common variants analyses were done to investigate their influence on the phenotype. Our paper provides further evidence of previously indicated and novel genes, as well as the contribution of rare-variants in MPHL phenotype. Thus, these findings could be the base for future investigations into the contribution of rare variants to MPHL. Information about the supplementary material for this article can be found in subsection 3.3.1 (Publication 3 - Appendix A). For clarity, more information about the machine learning methods used can be found in subsection 3.3.2 (Publication 3 - Appendix B). For this paper, I significantly contributed to the planning of the methodology, the data collection, curation and analysis, especially the analyses relevant to rare variants. I have also contributed to the evaluation and interpretation of the results.



Analysis of 72,469 UK Biobank exomes links rare variants to male-pattern hair loss

Received: 21 September 2022

Accepted: 24 August 2023

Published online: 22 September 2023



Sabrina Katrin Henne¹, Rana Aldisi², Sugirthan Sivalingam^{2,3},
Lara Maleen Hochfeld¹, Oleg Borisov², Peter Michael Krawitz², Carlo Maj^{2,4},
Markus Maria Nöthen¹ & Stefanie Heilmann-Heimbach¹ ✉

Male-pattern hair loss (MPHL) is common and highly heritable. While genome-wide association studies (GWAS) have generated insights into the contribution of common variants to MPHL etiology, the relevance of rare variants remains unclear. To determine the contribution of rare variants to MPHL etiology, we perform gene-based and single-variant analyses in exome-sequencing data from 72,469 male UK Biobank participants. While our population-level risk prediction suggests that rare variants make only a minor contribution to general MPHL risk, our rare variant collapsing tests identified a total of five significant gene associations. These findings provide additional evidence for previously implicated genes (*EDA2R*, *WNT10A*) and highlight novel risk genes at and beyond GWAS loci (*HEPH*, *CEPT1*, *EIF3F*). Furthermore, MPHL-associated genes are enriched for genes considered causal for monogenic trichoses. Together, our findings broaden the MPHL-associated allelic spectrum and provide insights into MPHL pathobiology and a shared basis with monogenic hair loss disorders.

Male-pattern hair loss (MPHL), or androgenetic alopecia, is the most common form of hair loss, with a lifetime prevalence of ~80% in European men. MPHL is characterized by progressive and androgen-dependent hair loss in the frontotemporal region and vertex of the scalp¹. Affected men may experience psychosocial effects², and lack well-tolerated and effective treatment options^{3,4}.

Early twin studies estimated that ~80% of the observed phenotypic variance of MPHL is attributable to genetic factors^{5,6}. Subsequent genome-wide association studies (GWAS) have yielded substantial insights into the genetic basis of MPHL via the identification of more than 600 independent genetic risk variants at more than 350 genomic loci, which together explain ~39% of the phenotypic variance^{7–17}. While these data have highlighted a number of plausible candidate genes and pathways, the majority of GWAS risk variants are common variants (minor allele frequency (MAF) > 1%) located in non-coding areas of the genome, which renders pinpointing of disease mechanisms and causal genes notoriously difficult.

In contrast, fewer data are available concerning the potential contribution to MPHL etiology of rare variants (MAF < 1%). A previous study on MPHL, which analyzed imputed genotyping data from the UK Biobank (UKB), estimated that the contribution of rare variants (MAF 0.0015% – 1%) to MPHL heritability was close to 0%¹³. However, imputed genotyping data do not offer comprehensive insights on rare variants, the systematic study of which has been hampered by the limited availability of whole genome or in the context of (rare) coding variants, whole exome sequencing (WES) data from adequately sized cohorts. Since 2019, the analysis of (rare) variants in coding areas of the genome has been facilitated by the availability of a large WES data set created by UKB^{18,19}. The UKB resource further contains self-report data on MPHL, thereby for the first time enabling the investigation of a potential relevance of rare variants to MPHL pathogenesis.

The aim of the present study therefore was to perform the first exome-based analysis on MPHL in a tranche of 200,629 exomes from

¹Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany. ²Institute for Genomic Statistics and Bioinformatics, University of Bonn, Bonn, Germany. ³Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany.

⁴Center for Human Genetics, University Hospital of Marburg, Marburg, Germany. ✉ e-mail: sheilman@uni-bonn.de

the UKB. Gene-based analyses (SKAT-O and GenRisk) and single-variant tests were used to investigate whether rare variants showed association with MPHL in a final set of 72,469 men. To interpret the association findings, multiple follow-up analyses were performed. A schematic overview of the study workflow is depicted in Fig. 1. Our first systematic analysis of the contribution of rare variants to MPHL etiology broadens the allelic spectrum of previously reported candidate genes (*EDA2R*, *WNT10A*), yields evidence for novel MPHL candidate genes both at and beyond known GWAS loci (*HEPH*, *CEPT1*, *EIF3F*), suggests an association between genotrichoses and the common MPHL phenotype and provides a basis for future investigations of the contribution of rare variants to MPHL pathobiology.

Results

Data set characteristics

After quality control, the final data set comprised the data of 72,469 men aged 39–82 years. Our continuous model, all-model and two-as-control model comprised 72,024 unrelated (kinship < 0.0442) men. Of these, 49,640 with any signs of baldness (pattern 2–4) were classified as cases (case-control ratio 2.2:1) in the all-model, and 33,454 (pattern 3 or 4) were classified as cases in the two-as-control model (case-control ratio 1:1.2). The age distribution per MPHL pattern group is shown in Fig. 2. The extreme model comprised 17,053 unrelated men, of whom 6523 relatively younger men (age < 60 years) with significant balding (pattern 4) were classified as cases and 10,530 elderly men (age ≥ 60 years) with no signs of balding were classified as controls (case-control ratio 1:1.6).

After filtering for per-sample and per-individual missing rates (<5%) and Hardy-Weinberg-Equilibrium ($P_{HWE} > 10^{-6}$), a total of 2,656,761 rare (MAF < 1%), nonsynonymous variants in 18,946 protein-coding genes remained for analysis in the SKAT-O and single-variant association tests (Fig. 1), with 239,082 variants in 18,449 genes meeting the more stringent high impact threshold (frameshift, splice acceptor-, splice donor-, and start- or stop-altering variants, transcript ablations and transcript amplifications). For the GenRisk analyses, a total of 16,211,028 rare (MAF < 1%) variants in 18,848 genes remained after filtering.

Analyses were performed to assess the optimal number of top principal components (PCs) to correct for. In the association tests of imputed genotype data with a variable number of included top PCs, minimum genomic inflation factor values were generated when including 14–20 PCs in the continuous model, 14–15 PCs in the all-model, 14–19 PCs in the two-as-control model, and 5 PCs in the extreme model (see Supplementary Fig. 1). Based on these findings, we opted to correct for 14 PCs in the continuous-, all- and two-as-control models, and for 5 PCs in the extreme model.

Single-variant association analyses

In a first step, we tested for an association of individual rare coding variants to MPHL. The analyses identified two genome-wide significant variants ($P < 8 \times 10^{-9}$) in the continuous- and all-model (Fig. 3, Supplementary Fig. 2, Supplementary Data 1). The two genome-wide significant variants, i.e., 23:66604439:G:A (rs12837393, MAF = 5.5×10^{-3} , $P_{\text{continuous}} = 3.0 \times 10^{-12}$, $\beta_{\text{continuous}} = 0.19$, $P_{\text{all}} = 4.8 \times 10^{-10}$, odds-

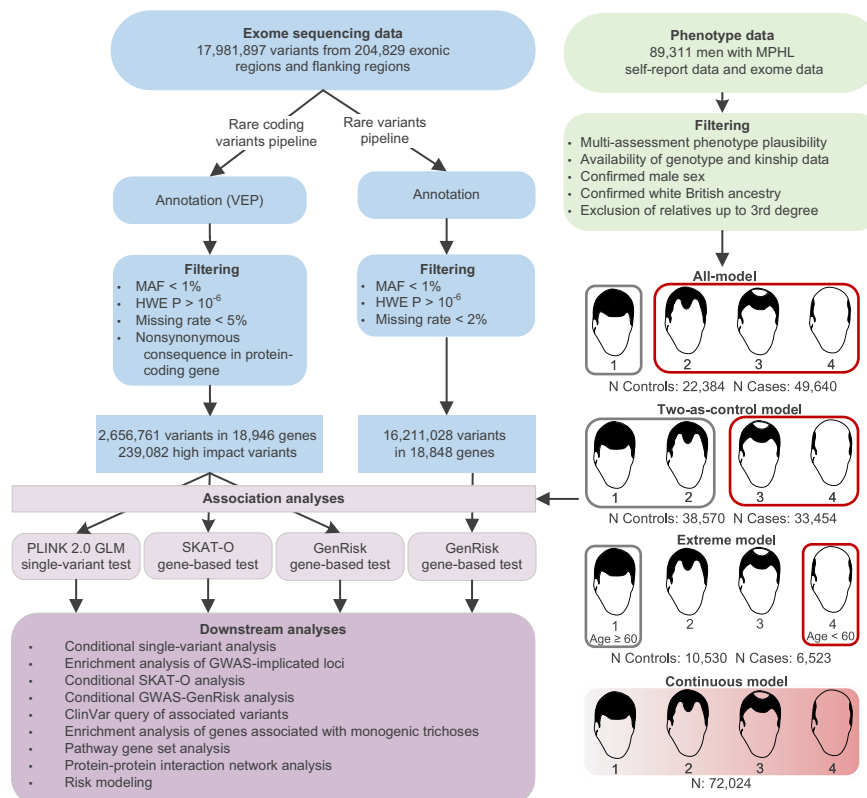


Fig. 1 | Overview of the analysis workflow. Exome and phenotype data obtained from the UKB were processed and used in three types of association analysis: GenRisk, SKAT-O, and single-variant testing. Four different phenotype models were used, of which three distinguishing cases (red) and controls (grey), as well as one continuous phenotype model. To interpret the association findings, several downstream follow-up analyses were performed. VEP ensemble variant effect

predictor, HWE Hardy-Weinberg-equilibrium, MAF minor allele frequency, GWAS genome-wide association study, MPHL male-pattern hair loss. MPHL pattern diagrams adapted from the UK Biobank survey accessible at <https://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=100423> and reproduced by kind permission of UK Biobank ©.

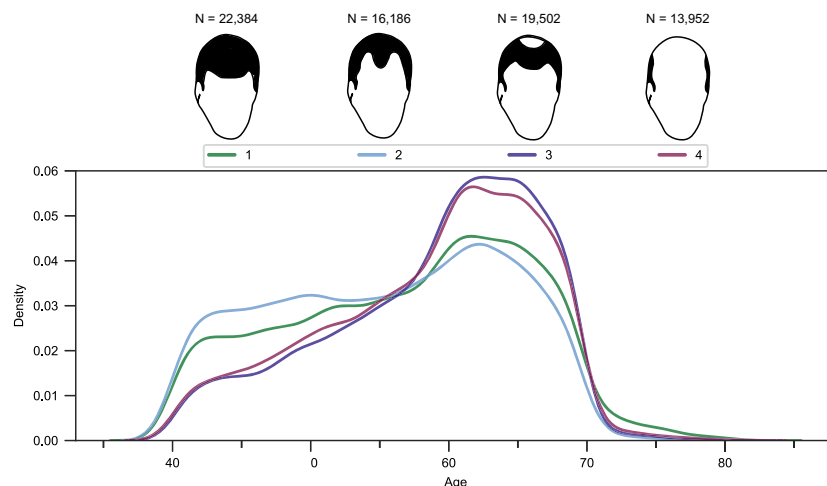


Fig. 2 | Phenotypic distribution within the final set of 72,024 men in the continuous-, all- and two-as-control model. Density plot showing the age distribution per male-pattern hair loss (MPHL) pattern group. The number of individuals in each

MPHL pattern group is shown above the plot. MPHL pattern diagrams adapted from the UK Biobank survey accessible at <https://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=100423> and reproduced by kind permission of UK Biobank ©.

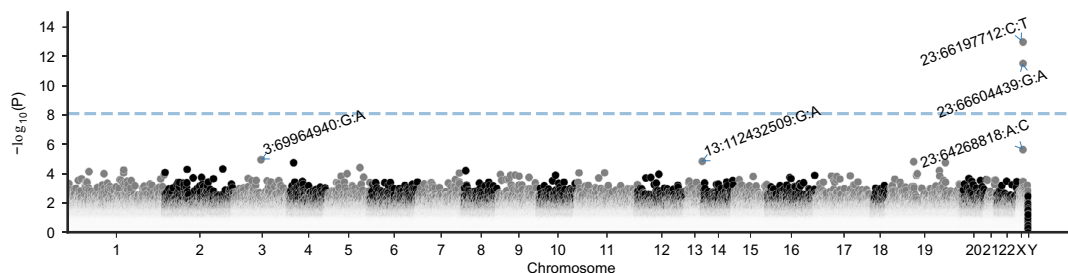


Fig. 3 | Results of the single-variant analysis for the continuous model.

Manhattan plot of single-variant association results for the continuous model. The dashed line denotes the selected genome-wide threshold for multiple testing in

single-variant tests (8×10^{-9}). The y-axis depicts $-\log_{10}(P)$ of the unadjusted P -value obtained from linear regression (two-sided). The top 5 variants were annotated.

ratio $[OR]_{all} = 1.53$, $r^2_{sentinel\ SNP} = 1.6 \times 10^{-4}$, $D'_{sentinel\ SNP} = 0.35$) and 23:66197712:C:T (rs151003259, $MAF = 2.0 \times 10^{-3}$, $P_{continuous} = 1.0 \times 10^{-13}$, $\beta_{continuous} = -0.35$, $P_{all} = 2.9 \times 10^{-10}$, $OR_{all} = 0.59$, $r^2_{sentinel\ SNP} = 4.5 \times 10^{-7}$, $D'_{sentinel\ SNP} = 1.0$) (GRCh38), are missense variants located within *EDA2R* and *HEPH* respectively. Notably, the T allele of 23:66197712:C:T was exclusively observed in combination with the MPHL risk allele ($MAF > 0.99$) of the respective GWAS sentinel SNP.

To assess whether the observed single-variant associations were independent of common variant associations previously identified through GWAS, all single-variant analyses were repeated with conditioning for 622 lead SNPs previously implicated in a UKB-based GWAS on MPHL¹³ (Supplementary Fig. 3, Supplementary Data 1). Neither of the previously significant single variants retained genome-wide significance after conditioning. While an association signal was retained for the variant 23:66604439:G:A in *EDA2R* ($P_{all} = 4.0 \times 10^{-4}$), the 23:66197712:C:T variant in *HEPH* was not independent of the GWAS lead SNPs ($P_{all} = 0.35$). Several variants retained a relatively low P -value even after conditioning, indicating a strong association that was independent from common GWAS variants. For instance, among the top ten variants post-conditioning were 3:69964940:G:A (rs149617956, located in *MITF*, $P_{continuous} = 5.4 \times 10^{-6}$), 2:218882368:C:A (rs121908119, located in *WNT10A*, $P_{two-as-control} = 6.1 \times 10^{-6}$), 21:44499878:C:T (rs138480801, located in *TSPEAR*, $P_{two-as-control} = 6.9 \times 10^{-6}$), 11:46366461:G:T (rs901998, located in *DGKZ*, $P_{two-as-control} = 1.1 \times 10^{-5}$), and 23:67711453:C:A (rs1800053, located in *AR*, $P_{continuous} = 1.8 \times 10^{-5}$).

Gene-based association analyses

To assess the cumulative contribution of rare variants to MPHL, we performed gene-based association analyses using SKAT-O²⁰ and GenRisk²¹, a new burden association test which upweights rarer and more deleterious variants (based on CADD). We applied the GenRisk test to a data set of both coding and non-coding rare variants, as well as to coding rare variants identical to the variant set used in the SKAT-O analysis. The SKAT-O analysis based on 2,656,761 variants from all ten variant consequence categories identified two genes with a genome-wide significant association ($P < 2.6 \times 10^{-6}$) to MPHL: *EDA2R* ($P_{continuous} = 1.4 \times 10^{-8}$); and *HEPH* ($P_{continuous} = 7.3 \times 10^{-9}$) (Fig. 4, Supplementary Data 2). No significantly associated genes were identified based on SKAT-O analyses of high-impact variants, with the top association, *WNT10A*, yielding a P -value of 7.8×10^{-6} in the two-as-control model (Supplementary Fig. 4, Supplementary Data 2).

The GenRisk analyses identified a total of three significantly associated genes ($P < 2.6 \times 10^{-6}$) across the four phenotype models: *EDA2R* ($P_{continuous} = 1.8 \times 10^{-6}$), *CEPT1* ($P_{all-model} = 2.1 \times 10^{-6}$), and *WNT10A* ($P_{two-as-control} = 2.2 \times 10^{-6}$) (Fig. 5, Supplementary Data 3). The *CEPT1* association finding is likely attributable to a combination of coding and non-coding variants with high CADD scores, and mainly driven by the MPHL pattern groups 2 and 3. Whether this reflects a biological aspect has to be determined by further analyses. The GenRisk analyses based on only coding variants further identified a significant association with *EIF3F* ($P_{two-as-control} = 2.5 \times 10^{-6}$) (Fig. 6, Supplementary Data 3).

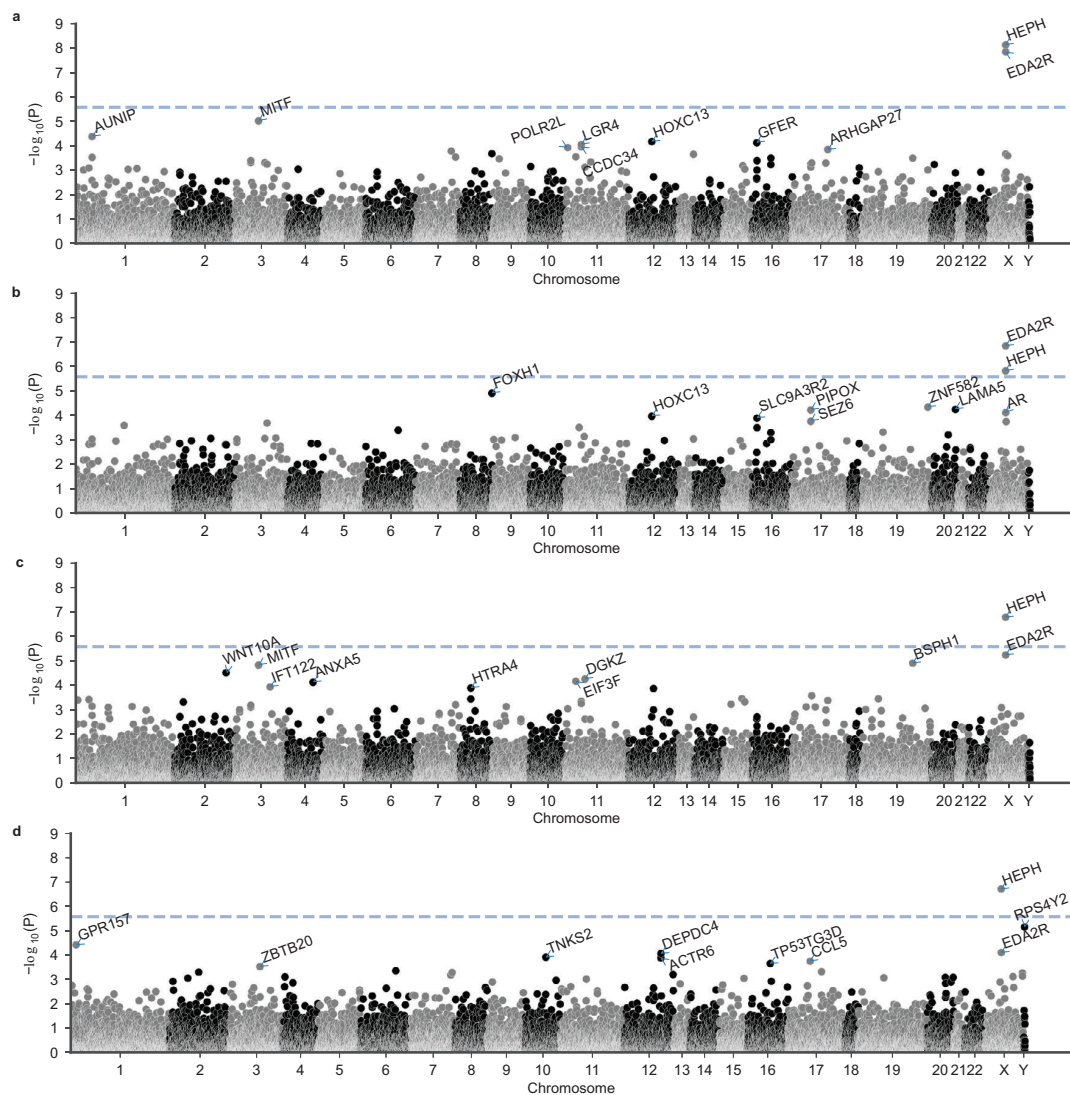


Fig. 4 | Results of the SKAT-O gene-based analysis. Results are shown for (a) the continuous model; (b) the all-model; (c) the two-as-control model; and (d) the extreme model. The y-axes depict $-\log_{10}(P)$ of the unadjusted P -value. The dashed

line denotes the Bonferroni threshold for multiple testing in the gene-based analyses (2.6×10^{-6}). The top 10 genes per analysis were annotated.

Comparison with an in-house data set on human hair follicle expression²² revealed that all five MPHL-associated genes (*EDA2R*, *HEPH*, *CEPT1*, *WNT10A*, *EIF3F*) are expressed in human hair follicles. Of these, *EDA2R*, *HEPH* and *WNT10A* are located at previously implicated MPHL-GWAS risk loci. An enrichment of a less stringent set of gene associations ($P < 3 \times 10^{-3}$ in the SKAT-O or GenRisk analyses) was observed in regions ± 1 Mb of published MPHL-GWAS lead SNPs ($P = 5.6 \times 10^{-15}$, overlap 192/595 genes). This was supported by the FUMA GENE2FUNC analysis, which identified an enrichment of these gene associations and MPHL GWAS findings reported in the GWAS catalog.

Conditional SKAT-O analyses were performed in order to determine whether the significant associations findings for *EDA2R* and *HEPH* were driven by the genome-wide significantly associated variants 23:66604439:G:A and 23:66197712:C:T, respectively. The P -values of *HEPH* and *EDA2R* both before and after the exclusion of these two variants are shown in Table 1. Notably, the association with *EDA2R* appears to have been driven very strongly by 23:66604439:G:A. In contrast, the effect of 23:66197712:C:T seems to have been less

pronounced, since the conditional analyses for *HEPH* generated low P -values (albeit non genome-wide significant), particularly in the two-as-control and the extreme model.

Conditional GWAS-GenRisk analysis

A conditional GWAS-GenRisk analysis was performed to test whether common variants implicated by GWAS are independent from GenRisk gene scores (Supplementary Data 4). The distribution of the differences in $-\log_{10}(P)$ with and without GenRisk gene score correction is shown in Supplementary Fig. 5. These data indicate no systematic dependence between common variants implicated by GWAS and GenRisk gene scores, as a large majority of tested common variants (99.89%) are not or only minimally impacted ($|\Delta - \log_{10}(P)| < 1$) by correction for any gene score. However, the GenRisk scores of the genes *EDA2R* and *WNT10A* show some attenuation of the common variant GWAS signal at their respective loci. In contrast, the associated gene *HEPH* and e.g., the *AR* gene do not show any such attenuation (Supplementary Fig. 6).

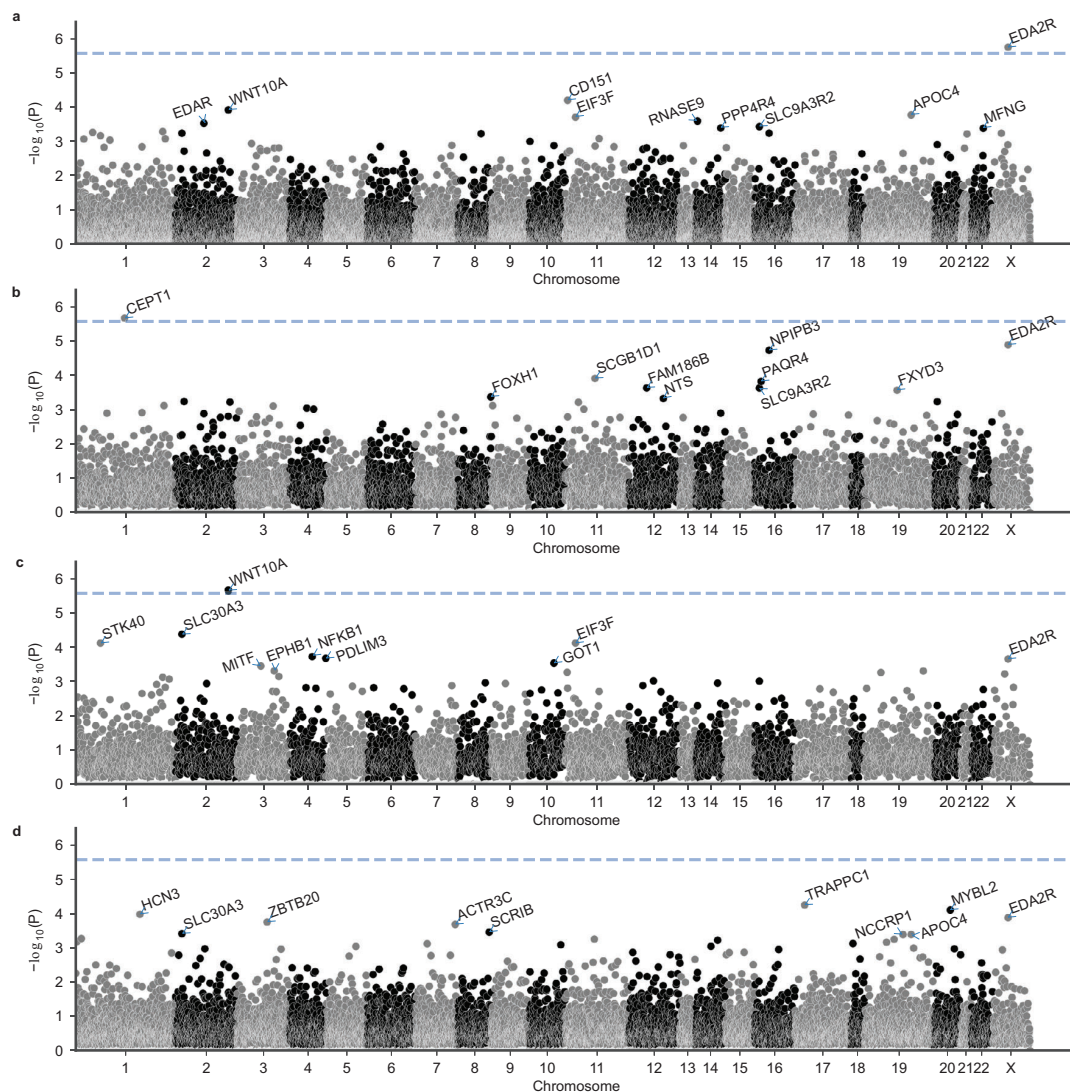


Fig. 5 | Results of the GenRisk gene-based analysis. Results are shown for (a) the continuous model; (b) the all-model; (c) the two-as-control model; and (d) the extreme model. The y-axes depict $-\log_{10}(P)$ of the unadjusted P -value. The dashed

line denotes the Bonferroni threshold for multiple testing in the gene-based analyses (2.6×10^{-6}). The top 10 genes per analysis were annotated.

Overlap with genotrichoses

The inspection of ClinVar (Supplementary Data 5) revealed that MPHL-associated variants comprise several variants that have been reported as pathogenic for monogenic trichoses. A systematic enrichment analysis of genotrichosis-associated genes^{23–27} amongst a less stringent set of gene associations ($P < 3 \times 10^{-3}$ in the SKAT-O or GenRisk analyses) revealed a significant enrichment ($P = 1.1 \times 10^{-4}$). The total overlap across all association analyses comprised the genes *WNT10A*, *HOXC13*, *DSP*, *LPAR6*, *ALX4*, *EDAR*, *CDH3*, *HR*, and *SPINK5*. Notably, two of the top associated single variants (albeit not genome-wide significant), i.e., 2:218882368:C:A ($P_{\text{two-as-control}} = 4.1 \times 10^{-5}$) and 21:44499878:C:T ($P_{\text{two-as-control}} = 9.0 \times 10^{-6}$), which are located in *WNT10A* and *TSPEAR* respectively, were reported to be pathogenic for ectodermal dysplasia in previous studies^{28,29}.

Pathway gene set and network analyses

Pathway-based gene set enrichment analysis of a less stringent set of 559 MPHL-associated genes ($P < 3 \times 10^{-3}$ in either the SKAT-O or the GenRisk analyses) revealed an enrichment of MPHL-associated genes in TGF-beta signaling (false discovery rate [FDR] = 0.040) and SMAD2/

3:SMAD4 transcriptional regulation (FDR = 0.021) (Supplementary Data 6). A protein-protein interaction network analysis of a less stringent set of 86 MPHL-associated genes ($P < 3 \times 10^{-4}$ in the SKAT-O or the GenRisk analyses) detected enrichments with ectodermal dysplasia genes (FDR = 2.6×10^{-3} , overlapping genes *EDA2R*, *WNT10A*, *EDAR*, *HOXC13* and *IFT122*) and genes assigned to the gene ontology term hair follicle development (FDR = 0.014, overlapping genes *WNT10A*, *EDAR*, *LAMA5*, *HOXC13*, *LGR4* and *ALX4*) (Supplementary Fig. 7).

Risk modeling

To evaluate the contribution of rare variants to MPHL, a risk prediction model integrating MPHL polygenic risk scores (PRS) and GenRisk gene-based scores was created (Fig. 7), as based on rare variants (MAF < 1%), age, sequencing batch and top PCs. The PRS-only risk model achieved medium discriminative power similar to the MPHL PRS model previously published by Hageaars et al.¹⁴ in distinguishing no hair loss (pattern 1) from severe hair loss (pattern 4), at least moderate hair loss (pattern 3–4) and at least slight hair loss (pattern 2–4), as measured by the area under the curve (AUC) ($\text{AUC}_{\text{severe}} = 0.791$, $\text{AUC}_{\text{moderate}} = 0.732$, $\text{AUC}_{\text{slight}} = 0.693$) when

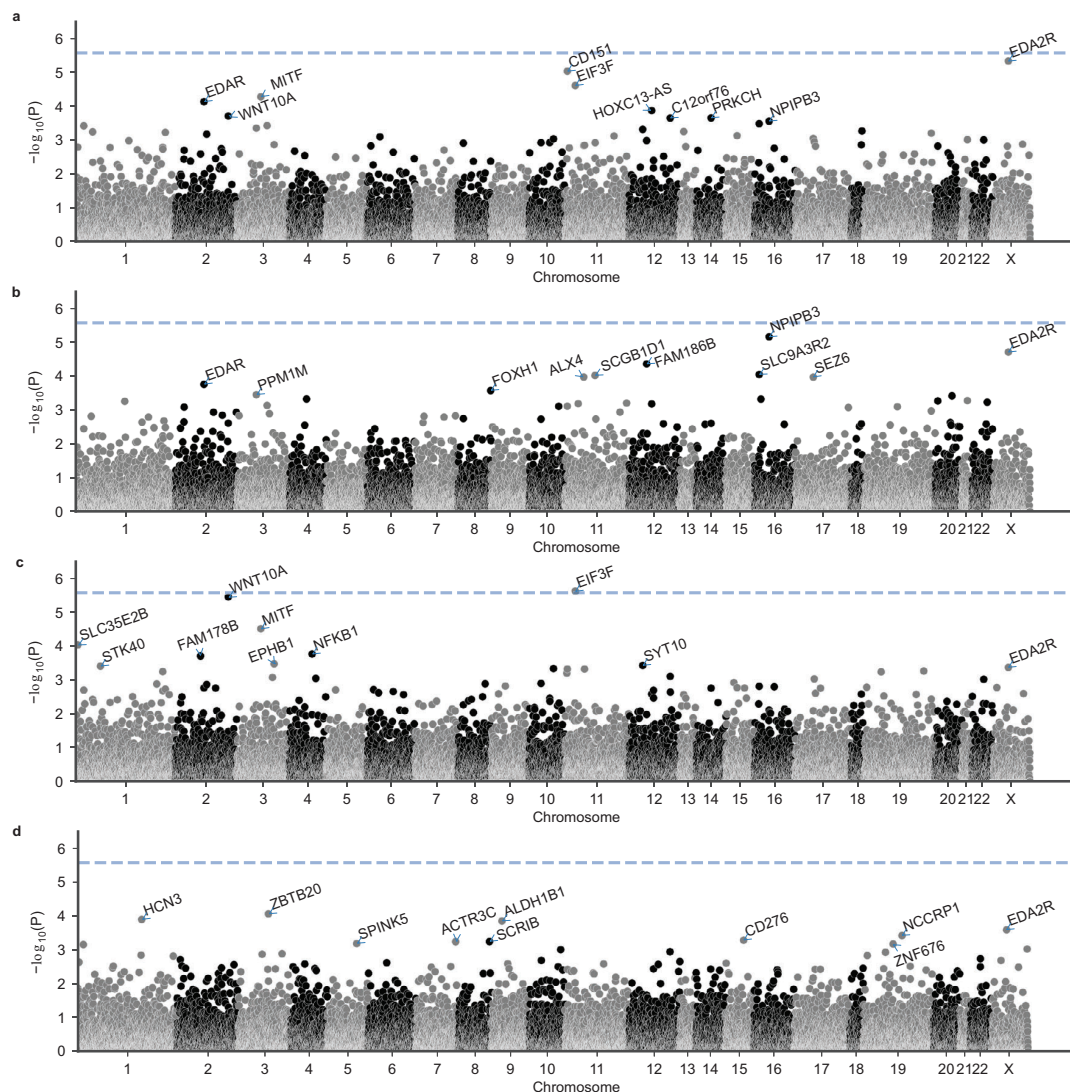


Fig. 6 | Results of the GenRisk gene-based analysis based on only coding variants. Results are shown for (a) the continuous model; (b) the all-model; (c) the two-as-control model; and (d) the extreme model. The y-axes depict $-\log_{10}(P)$ of the

unadjusted P -value. The dashed line denotes the Bonferroni threshold for multiple testing in the gene-based analyses (2.6×10^{-6}). The top 10 genes per analysis were annotated.

considering the full cohort of 72,024 males. In the test data set, the PRS-only model yielded slightly lower predictive power ($AUC_{\text{severe}} = 0.725$, $AUC_{\text{moderate}} = 0.687$, $AUC_{\text{slight}} = 0.647$). A risk model based exclusively on the gene-based risk score, which integrated all gene-based scores into one, showed low discriminative power ($AUC_{\text{severe}} = 0.560$, $AUC_{\text{moderate}} = 0.557$, $AUC_{\text{slight}} = 0.508$). Integration of PRS and gene-based risk scores generated only minimal to no increase in discriminative power compared to the PRS-only model ($AUC_{\text{severe}} = 0.726$, $AUC_{\text{moderate}} = 0.686$, $AUC_{\text{slight}} = 0.646$). Despite the high number of associated genes, this largely confirms earlier observations that rare variants explain only a minor fraction of the genetic risk for MPHIL at population-level¹³.

Discussion

MPHL is a complex, common trait for which a large number of risk loci and variants have already been characterized via analyses of common variation^{7–17}. The main aim of the present study was to analyze the extent to which rare variants contribute to MPHIL. A previous study of MPHIL, which was based on imputed genotyping data from the UKB, showed that the contribution of rare variants (MAF between 0.0015%

and 1%) to MPHIL heritability was close to 0%¹³. To reassess this finding, we accessed a large exome sequencing data set from the UKB in order to perform a systematic analysis of rare variants in coding areas of the genome.

In line with previous reports that suggest a minor contribution of rare variants to MPHIL heritability, our risk prediction models showed that the inclusion of gene-based scores that are based on rare variants into existing risk prediction models based on common variants made little to no contribution to discriminative power between cases and controls. This is also reflected in the low number of significant association findings in our single-variant analysis. Both rare variant associations identified ($P < 8 \times 10^{-9}$) have already been reported at genome-wide significance in GWAS¹³.

The SKAT-O and GenRisk gene-based analyses detected significant associations with rare variants in five genes ($P < 2.6 \times 10^{-6}$), which, while limited, offers important insights into MPHIL biology, and may be etiologically relevant for individual risk. The identified gene associations comprise both previously implicated and novel MPHIL candidate genes. Genes previously implicated by GWAS include *EDA2R* (ectodysplasin A2 receptor), one of the flanking genes at the most

strongly associated MPHL GWAS locus on chromosome (chr) X (*AR/EDA2R* locus)³⁰ and *WNT10A* (Wnt Family Member 10A), the likely causal gene at the chr.2q35 risk locus for which a functional interaction with another MPHL risk locus has been shown³¹. These findings suggest that both common and rare variation in these genes contributes to MPHL etiology. The analyses further identified an association with *HEPH* (Hephaestin), which, while being located less than 500 kb upstream of *EDA2R*, has not been previously considered a candidate gene. However, recent reports have indicated that *HEPH* plays a crucial role in hair development through its ferroxidase activity³². In addition to the insights that our rare coding variant analyses yielded at GWAS loci, they also implicate novel MPHL candidate genes beyond GWAS loci, namely *CEPT1* (Choline/ethanolamine phosphotransferase 1) and *EIF3F* (Eukaryotic translation initiation factor 3 subunit F). *CEPT1* encodes the terminal enzyme in the Kennedy pathway of phospholipid

biosynthesis³³. While no reports specifically linking *CEPT1* and hair (loss) biology exist, there is evidence for a link between phospholipid metabolism and hair biology. For example, the topical administration of phospholipids was shown to promote hair growth in mice³⁴, and overexpression of group X-secreted phospholipase A₂ in mice led to alopecia and changes in hair cycling³⁵. *EIF3F* encodes a subunit of the eukaryotic initiation factor 3 (eIF-3) complex. Recent reports suggest a potential involvement of *EIF3F* in hair pigmentation, as a patient with two heterozygous variants in *EIF3F* presented with skin and hair hypopigmentation³⁶, and a heterozygous *EIF3F* knock-out resulted in abnormal coat pigmentation in mice³⁷. This is of interest as the transformation of pigmented terminal hair follicles to unpigmented vellus hair follicles is a pathophysiological feature of MPHL³⁸. Additionally, *EIF3F* has been shown to act as a negative regulator of cell proliferation in cancer cells³⁹, and was shown to regulate Notch signaling⁴⁰, which in turn is involved in hair follicle stem cell fate determination⁴¹.

Our conditional single-variant analysis further identified a number of strong associations independent from common GWAS variants. Among the top ten variant associations from this analysis are variants located within the genes *AR* (androgen receptor), *WNT10A*, *TSPEAR* (Thrombospondin Type Laminin G Domain and EAR Repeats), *MITF* (Melanocyte Inducing Transcription Factor) and *DGKZ* (Diacylglycerol Kinase Zeta). Given that these rare, nonsynonymous coding variants achieved low *P*-values - albeit above the threshold for genome-wide significance - despite the generally low power of the single-variant analyses, these may constitute independent candidate genes. The two genome-wide significant single variant associations 23:66604439:G:A (in *EDA2R*) and 23:66197712:C:T (in *HEPH*) did not retain genome-wide significance after conditioning, pointing to a (partial) interdependence between these variants and common GWAS variants, which was more pronounced for 23:66197712:C:T, while 23:66604439:G:A retained a partial signal. We further observed that (i) the rare MPHL risk allele of the 23:66604439:G:A variant occurs

Table 1 | Results of conditional SKAT-O analysis, involving the removal of the two variants that showed genome-wide significance in the single-variant analyses (23:66197712:C:T and 23:66604439:G:A)

	Gene	<i>P</i>	<i>P</i> _{conditioned}
Continuous model	<i>HEPH</i>	7.3×10^{-9}	1.3×10^{-4}
	<i>EDA2R</i>	1.4×10^{-8}	0.84
All-model	<i>HEPH</i>	1.5×10^{-7}	1.3×10^{-2}
	<i>EDA2R</i>	1.4×10^{-7}	0.79
Two-as-control model	<i>HEPH</i>	1.7×10^{-7}	2.8×10^{-5}
	<i>EDA2R</i>	5.9×10^{-6}	0.81
Extreme model	<i>HEPH</i>	2.0×10^{-7}	5.7×10^{-6}
	<i>EDA2R</i>	8.0×10^{-5}	0.12

The SKAT-O *P*-value (unadjusted) before and after conditioning is shown according to gene and phenotype model.

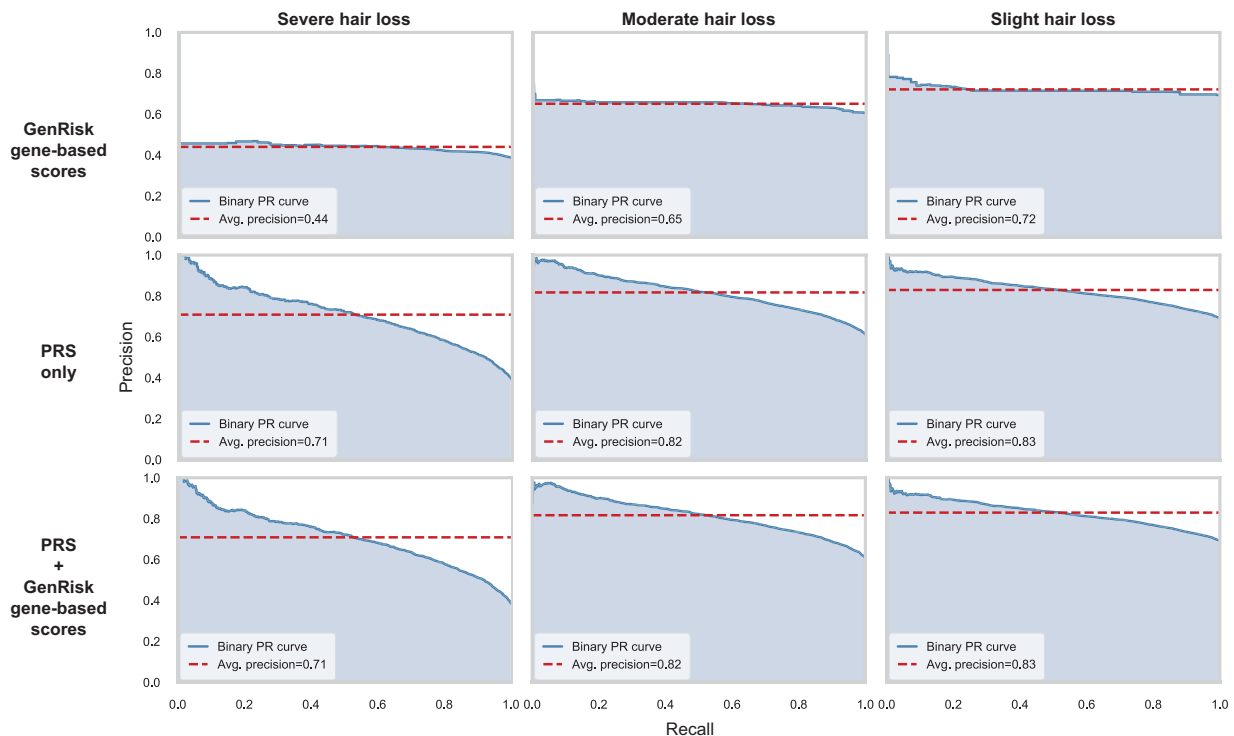


Fig. 7 | Precision-recall-curves of the created MPHL risk models based on PRS only, GenRisk gene-based scores, and PRS combined with GenRisk gene-based scores. The models were tested in terms of prediction of no hair loss (pattern 1) vs

severe hair loss (pattern 4), at least moderate hair loss (pattern 3-4), and at least slight hair loss (pattern 2-4). PRS polygenic risk score, PR precision-recall, Avg. average.

exclusively on the common MPHL risk haplotype previously reported by Hillmer et al.⁴² (rs2497935-A, rs962458-A, rs12007229-C, rs12396249-G) and (ii) the rare protective allele of the 23:66197712:C:T variant occurs almost exclusively on a lower-risk haplotype with only the rs962458-A risk allele. While the 23:66604439:G:A variant exclusively occurs on the previously reported MPHL risk haplotype, a partial signal remains in the conditional analyses, which may point to an independent effect of the rare variant and the risk haplotype. However, at this point, a causal role of either variant can neither be confirmed nor excluded.

As genes identified through our rare variant gene-based association tests were enriched for genes at known MPHL GWAS loci (lead SNP ± 1 Mb), these data underline the importance of studies that assess the entire allelic spectrum of disease associations, and their potential to highlight causal genes at GWAS risk loci. A conditional GWAS-GenRisk analysis was performed and found no systematic dependence between common GWAS-implicated variants and GenRisk gene scores. The analysis however identified risk loci where the GWAS association signal appears to be (partially) driven by both common and rare variants, namely chr.Xq12 (*EDA2R*) and chr.2q35 (*WNT10A*). Associated loci which are not impacted by any GenRisk gene score may be due to a low contribution of rare deleterious variants to the association. However, further investigation into the extent of dependence between common variants and GenRisk gene scores is required.

The X-chromosome has long been at the center of genetic analyses on MPHL. Early studies focused on the X-linked androgen receptor gene (*AR*), due to the strict androgen dependency of the phenotype. Although the results have been conflicting in regards to the likely causal variants and genes, the *AR/EDA2R* locus has consistently been the most strongly associated genomic region for MPHL, although neither the precise causal variants nor the causal genes have been confirmed⁴³. In the present study, we identified significant associations with two X-chromosomal genes, namely *EDA2R* and *HEPH*, thereby yielding new or additional evidence for these candidate genes. Our analyses did not identify significant associations of rare variants in the *AR* gene ($P_{\text{SKAT-O binary}} = 7.6 \times 10^{-5}$). This is in line with previous Sanger-sequencing-based studies of the *AR* coding sequence, which did not identify any significant associations between the *AR* and MPHL^{44,45}. Although we cannot exclude the possibility that our analysis lacked statistical power to detect such an association, one might also speculate that a potential involvement of the *AR* gene in MPHL pathobiology is impacted primarily by regulatory common variants, rather than rare variants in or around its coding sequence.

Moreover, a less stringent set of MPHL-associated genes overlapped with and were enriched for genes that have been reported as the cause of monogenic trichoses, namely *WNT10A* (odonto-onychodermal dysplasia and Schöpf-Schulz-Passarge syndrome), *HOXC13* (pure hair and nail ectodermal dysplasia), *DSP* (Carvajal syndrome), *LPAR6* (hypotrichosis 6), *ALX4* (total alopecia in frontonasal dysplasia), *EDAR* (ectodermal dysplasia), *CDH3* (ectodermal dysplasia), *HR* (hypotrichosis 4 and alopecia universalis), and *SPINK5* (Netherton syndrome). Notably, most of these genes either cause ectodermal dysplasias, hypotrichoses or alopecia. However, as we detected variants with a previously reported likely or known pathogenic association with genotrichoses in both cases and controls, no definitive statement can be made as to whether the presence of or variable expressivity of a genotrichosis may have led to a misclassification in the MPHL self-report. Generally, an overlap between genotrichoses and MPHL-associated genes would be biologically plausible, as different levels of impairment of key hair follicle signaling pathways would be expected to result in differing phenotypes. For example, GWAS have previously yielded evidence for an association between hair curl and MPHL⁹. Together, these findings may indicate an overlap in causal genes between genotrichoses and MPHL.

Rare coding variants in the associated genes identified in this study have been previously associated with phenotypes such as mean corpuscular haemoglobin (*EIF3F*) and urea (*HEPH*)^{46,47}. Suggestive associations have further been identified between testosterone levels and *EDA2R*, and alcohol use and *EIF3F*. Some of these associations may present interesting links – for instance, epidemiological studies have (albeit with conflicting evidence) found associations between MPHL and alcohol consumption⁴⁸.

The present analyses utilized four different phenotype models. Our continuous model represented a 1:1 representation of the progressive phenotype, which may however be most sensitive to misclassifications in the self-report. Our all-model provided a simple description of the phenotype by considering unaffected men as controls and men with any type of balding (frontal or vertex) as cases. The purpose of our extreme model was to achieve complete separation between cases and controls, despite the age-dependent and progressive nature of MPHL. This involved considering men with complete baldness of the scalp below 60 years of age as cases, and unaffected men aged 60 years or older as controls. The aim of this approach is to facilitate detection of variants and genes contributing to balding in relatively younger men and may provide higher statistical power, as these supercontrols are among the 10% of men least affected by MPHL¹ and are unlikely to develop a significant degree of balding during their lifetime. However, this phenotype model comes at the expense of sample size, which was reduced by nearly 80% compared to the other phenotype models. The purpose of the two-as-control model was to address the possibility of misclassifications in the self-reporting of balding. Misclassifications may be possible for UK Biobank MPHL patterns 1 and 2 (unaffected vs frontal balding), since we are of the opinion that the presence of balding in the frontotemporal regions of the scalp may be subjectively over- or underestimated in the absence of a dermatological assessment. In the present study, the different phenotype models yielded partially distinct gene associations, for example *WNT10A* and *EIF3F*, which consistently showed stronger signals in the two-as-control model. This may be an indication that distinct mechanisms contribute to more severe stages of balding, which are easier to detect using this case-control separation. All in all, the phenotype models employed in this study provide different perspectives on the MPHL phenotype and can account for certain possible errors in the self-report.

In this study, we performed two types of gene-based analyses: SKAT-O and GenRisk. SKAT-O is a well-established tool for gene-based association analyses and has the ability to detect associations in the presence of mixed effect directions at the variant level. GenRisk employs a scoring system that uses a beta distribution weighting schema for allele frequency, which is similar to SKAT-O, and pathogenicity scores (CADD score), to upweight rare and deleterious variants. As a result, GenRisk does not require variant consequence filtering. Moreover, GenRisk generates individual-level gene-based scores, which can be used in downstream analyses such as association analyses and risk prediction modeling. GenRisk was recently used to identify associations between rare genetic variants and blood biomarkers, identifying both known and novel associations (preprint)⁴⁹. In the present study, both methods yielded partially distinct gene associations. While the inclusion of non-coding variants and non-protein-coding genes in the GenRisk analysis may yield overall more comprehensive results, the association signal may encompass a greater overlap with GWAS. The GenRisk analysis of coding variants only, on the other hand, offers an increased focus on high-impact coding variants, without severely reducing the number of variants through e.g., high-impact variant consequence filters. The analyses employed in this study therefore address different hypotheses. While each method offers different biological insights, some identified gene associations are consistent between SKAT-O

and GenRisk, and Fisher's exact tests show a significant overlap of a less stringent set of associations ($P < 3 \times 10^{-3}$) between the two analyses across all phenotype models ($OR_{\text{continuous}} = 54.3$, $P_{\text{continuous}} = 1.5 \times 10^{-17}$; $OR_{\text{all-model}} = 92.8$, $P_{\text{all-model}} = 2.1 \times 10^{-24}$; $OR_{\text{two-as-control}} = 71.3$, $P_{\text{two-as-control}} = 3.5 \times 10^{-20}$; $OR_{\text{extreme model}} = 99.1$, $P_{\text{extreme model}} = 3.9 \times 10^{-19}$). However, given the novelty of the approach, corroboration of the GenRisk results in further studies is desirable.

To our knowledge the present study represents the first systematic analysis of the contribution of rare variants to MPHL etiology. While rare variants in coding regions of the genome seem to make only a small contribution to MPHL genetic risk at population-level and may have little value for risk prediction, they may nonetheless contribute significantly to individual risk. In line with this hypothesis, we observed only a marginal contribution to the overall MPHL risk prediction of gene-based burden scores with respect to the PRS. Since prediction model performances are typically assessed on overall data set metrics (such as AUC) it can be expected that the impact of variables informative only for a small proportion of samples can be marginal (e.g., there might be few individuals whose MPHL genetic risk can be attributed to damaging rare variants in a specific MPHL susceptibility gene). Instead, PRS by providing a gradient-risk in the overall data set can model the genetic risk throughout the population, therefore representing a global genetic risk variable. While gene-based burden scores may not be particularly suited for risk prediction models in the general population, they are a powerful instrument to detect gene associations and can therefore be helpful to dissect the genetic architecture of complex traits such as MPHL. As demonstrated with our study, the analysis of rare variants additionally offers important insights into associated alleles, genes and pathways, as well as pleiotropy, thereby improving our understanding of MPHL pathobiology. While the present study provides first insights into the contribution of rare variants to MPHL pathobiology based on a tranche of 200,629 exomes from the UK Biobank, the final data set of ~450,000 exomes has been released while completing the present analyses. This data set represents a considerable increase in sample size. Continued investigation on the role of rare variants for MPHL using this larger data set is therefore warranted.

In summary, the findings of our analysis broaden the allelic spectrum of previously reported candidate genes (*EDA2R*, *WNT10A*), yield evidence for novel MPHL candidate genes both at (*HEPH*) and beyond (*CEPT1*, *EIF3F*) known GWAS loci and suggest an association between genotrichoses and the common MPHL phenotype. Together, they provide a basis for future investigations into MPHL pathobiology and the contribution of rare variants to MPHL. Investigations of the functional relevance of rare variants and their interactions with common variants at and beyond risk loci will eventually improve our understanding of MPHL pathobiology and may lead to improved risk prediction and identification of affected pathways and can pave the way for the development of personalized therapies.

Methods

Phenotype data

The UK Biobank study has been approved by the North West Multi-centre Research Ethics Committee as a Research Tissue Bank and all UKB participants provided written informed consent. The UKB 200k release contains exome- and MPHL self-report data from 89,311 men^{18,19}. These MPHL self-report data were recorded at up to four UKB assessment center visits. Using a touch-screen questionnaire, participants scored their hair loss on a scale of 1 to 4, as based on four pictograms (Supplementary Fig. 8): 1 – Unaffected; 2 – frontotemporal balding; 3 – balding of the frontotemporal region and vertex; and 4 – complete baldness of the top of the scalp.

In the present study, four phenotype models were used: (i) a continuous model, which considers hair loss patterns 1–4 on a

continuous scale, (ii) an all-model, in which controls (pattern 1) were compared to cases (pattern 2–4); (iii) an extreme model, in which supercontrols (pattern 1, age ≥ 60) were compared to severe cases (pattern 4, age < 60); and (iv) a two-as-control model, in which controls (pattern 1–2) were compared to cases (pattern 3–4) in order to address the possibility of misclassifications between pattern 1 and 2 in the self-assessment.

For individuals who provided MPHL data at more than one assessment center visit, additional steps were performed in order to check the self-report data for sanity, and to select an entry for use in the analyses. The most recently recorded MPHL pattern was selected for analysis, unless an improvement in MPHL status was recorded. Due to the progressive nature of MPHL, an improvement is implausible. To avoid the need to exclude individuals who reported an improved MPHL status and to instead identify a plausible MPHL pattern, the following steps were performed: (i) if two balding patterns were available, and the difference between the patterns was no larger than 1, the higher pattern was used; (ii) if 3 balding patterns were available, a pattern that was recorded 2 times was used; (iii) if 4 balding patterns were available, a pattern that was recorded 3 times was used. If no plausible MPHL pattern could be identified in this manner, the individual was excluded. To account for the age-dependency of MPHL, in case of multiple assessments, for cases, we selected the lowest age at which the highest MPHL pattern was recorded. For controls, we selected the highest age at which no (pattern 1) or mild (pattern 2) hair loss was recorded.

To select participants for the present analysis, the following four criteria were used: (i) no grounds for exclusion found in the MPHL multi-entry sanity check; (ii) availability of genotype and kinship data; (iii) genetically and self-reported male sex with no sex chromosome aneuploidy; and (iv) self-reported white British ethnicity, as well as very similar genetic ancestry, as based on a principal components analysis of the genotypes. In addition, related individuals up to the third degree were excluded on the basis of UKB kinship coefficients (kinship coefficient ≥ 0.0442). Iterative exclusion was performed for one individual in a related pair, with individuals with a larger number of related individuals being excluded preferentially. An unexpected improvement of MPHL was observed in 2235 individuals. Of these, 293 were excluded since no plausible MPHL pattern could be nominated. A total of 72,469 of the 89,311 male UKB participants fulfilled these criteria. Exclusion of related individuals was performed separately for each phenotype model, resulting in the following final sample counts: 72,024 (continuous, all- and two-as-control models) and 17,053 (extreme model).

Variant data

Exome sequencing variant data for the 200,643 participants in the UKB 200k release were downloaded from the UKB in PLINK format. The data comprised 17,981,897 variants, which were captured from 204,829 autosomal and gonosomal exonic regions ± 100 bp flanking regions. For the SKAT-O and single-variant analyses, the data were quality controlled in PLINK 2.0⁵⁰ with respect to per-individual missing rate ($< 5\%$), per-variant missing rate ($< 5\%$), and Hardy-Weinberg equilibrium ($P > 10^{-6}$). Variants were filtered for a MAF $< 1\%$ based on their frequency in the different phenotype model data subsets. Variants were converted to variant call format (VCF) and annotated using the Ensembl Variant Effect Predictor (VEP)(v104)⁵¹. Variants with a predicted nonsynonymous consequence of moderate or high impact in a protein-coding gene (as based on Ensembl gene annotation release 104) were selected. These variant criteria comprised missense, insertion, deletion, splice acceptor-, splice donor-, and start- or stop-altering variants, as well as transcript and regulatory region ablations.

For the GenRisk analyses, variant data were quality controlled in PLINK 2.0 with respect to per-variant missing rate ($< 2\%$) and

Hardy-Weinberg equilibrium ($P > 10^{-6}$). Variants were filtered for an MAF $< 1\%$, as based on their frequency in the data set. The variant data were annotated with CADD (Combined Annotation Dependent Depletion) scores (CADD v1.6)⁵² and gene features based on the GRCh38 NCBI RefSeq refflat table from the UCSC Genome Browser^{53,54}.

Imputed genotype data were downloaded from the UKB in BGEN format. These data comprised information concerning 97,059,328 variants and were converted to PLINK format using PLINK 2.0 and the *ref-first* parameter.

Correction for population stratification

To account for population stratification, analyses were performed to estimate the optimal number of top PCs to include in our statistical models. For this purpose, GWAS were performed on the UKB imputed genotype data using a varying number of included PCs, and the genomic inflation factor λ was determined. Imputed genotype data were processed in PLINK 2.0, with preservation of the imputed genotype dosages. The data were quality controlled for info score (≥ 0.3) and minor allele count (≥ 20) and filtered for each of the four phenotype model subsets. GWAS were performed in PLINK 2.0, with correction for age and the 1–20 top PCs, as pre-calculated by the UKB based on imputed SNP genotype data. In the extreme model, age correction was omitted, since this phenotype model differentiates based on age.

Association analyses

The association analyses of the continuous model, all-model and the two-as-control model were corrected for age, sequencing batch and the top 14 PCs. In the association analyses of the extreme model, correction was made for sequencing batch and the top five PCs only. GWAS-style single variant analyses of the filtered exome data were performed in PLINK 2.0 using the *glm* function with covariate normalization to mean 0, variance 1. LD of the single-variant associations with the sentinel GWAS SNP was calculated in PLINK 2.0 using the *ld* function. For the 23:66197712:C:T variant in *HEPH*, the nearest GWAS lead SNP was used: 23:66001818:T:A (rs771150309, MPHL risk allele = major allele = T). For the 23:66604439:G:A variant in *EDA2R*, the closest GWAS lead SNP (23:66418642 (rs5965195)) was not contained in the employed imputed genotyping data release, and the nearest significant SNP was used instead: 23:66418267:G:A (rs4827473, MPHL risk allele = major allele = A).

Two types of gene-based analyses were performed: SKAT-O²⁰ and GenRisk²¹. SKAT-O was applied to the filtered exome data using the *SKATBinary.SSD.All* (for binary phenotype definitions) and the *SKAT.SSD.All* (for the continuous phenotype definition) functions with default settings in the SKAT R package (v2.0.1)²⁰. Data were converted to PLINK 1 binary format using PLINK 2.0 for use as input files. Variants were assigned to genes based on the VEP annotation approach described above. In addition to the nonsynonymous variant consequence threshold imposed through the present filtering steps, more stringent thresholds were applied in this analysis by restricting inclusion to variants of high impact, as based on VEP annotation (splice acceptor, splice donor, stop- or start-altering and frameshift variants, as well as transcript ablations). The GenRisk analysis was performed on the filtered exome data in VCF format using the GenRisk Python package (v0.2.5)²¹. GenRisk was applied separately to i) rare variants (MAF $< 1\%$) annotated to any gene and ii) only coding variants, using the identical variant set as used in the SKAT-O analyses. Gene-based scores were generated using weighted MAF (beta density function with parameters $a=1$ and $b=25$) and raw CADD scores as functional annotation, whereby variants with a lower MAF or higher CADD score were upweighted. The association analysis of the gene-based scores and previously described covariates was performed using linear regression (continuous model) or LI-logistic regression (all-, two-as-control and extreme models).

The P -value threshold for genome-wide significance in single-variant association analyses was selected as 8×10^{-9} , as empirically determined by Karczewski et al. based on analyses of 394,841 UK Biobank exomes⁴⁶. P -value thresholds for the SKAT-O and GenRisk gene-based analyses were determined using Bonferroni correction based on the maximum number of genes tested, resulting in a threshold of 2.6×10^{-6} (corresponding to 18,946 genes tested in the SKAT-O analysis).

Enrichment analyses

To improve the feasibility of enrichment analyses and obtain a more comprehensive gene list of approximately 500 genes, a less stringent P -value threshold of $P < 3 \times 10^{-3}$ was selected, resulting in a less stringent set of 595 MPHL-associated genes. Testing was performed for an enrichment of this less stringent set of MPHL-associated genes in genes located ± 1 Mb of previously published GWAS lead SNPs^{7–17}. Enrichment testing using a one-tailed Fisher's exact test from *scipy* (v1.8.1) was performed with a background list comprising the final tested genes per phenotype model. Using the same method, analyses were also performed to test for an enrichment of this less stringent set of MPHL-associated genes in genes causative for monogenic trichoses. A list of 65 known trichosis genes was created, as based on previous publications^{23–27}. The genes and their corresponding condition are listed in Supplementary Table 1.

ClinVar query

An inspection was made to determine whether MPHL-associated rare variants have been described as pathogenic or likely pathogenic on ClinVar. ClinVar data were downloaded as VCF (accessed 02.05.2022) and filtered for nominally significant single variants ($P < 0.05$ in any phenotype model). Information on associated conditions was extracted for variants listing a clinical significance of *pathogenic*, *likely pathogenic* or *conflicting interpretations of pathogenicity*.

Conditional analyses

To evaluate the dependence of association signals on specific variants, conditional analyses were performed. Gene-level conditional analyses of the genes *EDA2R* and *HEPH* were conducted using SKAT-O, as previously described, after the exclusion of two variants that showed genome-wide significance in the single-variant analyses (23:66197712:C:T and 23:66604439:G:A).

Using data from the continuous model, we tested whether SNPs previously implicated in GWAS were independent from GenRisk gene scores. Imputed genotype data from the UKB were filtered for MAF ($> 1\%$), info score (> 0.3), per-variant missing rate ($< 5\%$), Hardy-Weinberg equilibrium ($P > 10^{-6}$). Association analyses of the filtered imputed genotype data were performed in PLINK 2.0 using the *glm* function with covariate normalization to mean 0, variance 1, and corrected for age and 14 top PCs. The analyses were performed per locus, defined based on 622 SNPs that were identified as independent MPHL lead SNPs in a UKB-based GWAS¹³ ± 500 kb flanking regions. For each gene per locus, the analysis was additionally corrected for the respective GenRisk gene scores. Resulting P -values and effect sizes were then compared between the uncorrected and gene-corrected analyses.

To test whether the rare single-variant associations were independent from common SNPs previously implicated in GWAS, imputed genotype data from the UKB were filtered for info score (> 0.3) and 622 SNPs that were identified as independent MPHL lead SNPs in a UKB-based GWAS¹³. GWAS-style exome single variant analyses were then performed as described above, with the inclusion of genotypes for lead SNPs on the same chromosome as covariates.

Pathway gene set and network analyses

Gene set analysis was performed using FUMA GENE2FUNC⁵⁵ (v1.4.0) with default settings. MPHL-associated genes ($P < 3 \times 10^{-3}$ in any gene-

based test) were used as input gene list. All tested genes were supplied as a background list. The results were filtered for pathway gene set categories, namely canonical pathways, curated gene sets, computational gene sets, chemical and genetic perturbation, hallmark gene sets, Reactome, KEGG and WikiPathways. To further obtain an overview of protein interactions and co-expression, a STRING (v11.5)⁵⁶ protein network analysis was performed using a less stringent set of MPHL-associated genes. Here, a threshold of $P < 3 \times 10^{-4}$ was selected in order to obtain a more manageable network of <100 genes.

Risk modeling

To test whether the inclusion of rare variants improves common variant-based risk modeling of MPHL, GenRisk was used to create a risk prediction model integrating MPHL PRS with GenRisk gene-based scores, which were generated as described above. In order to establish a PRS model, a GWAS of imputed genotyping data from the UKB was performed based on data from our continuous model. Individuals with no exome sequencing data were selected to ensure no sample overlap, and filtered using the criteria described previously, resulting in 105,565 unrelated (both within this sample and with the 72,024 individuals of the continuous, all- and two-as-control models) male individuals. The imputed genotype data were quality-controlled (info score >0.3, per-variant missing rate <5%, HWE $P > 10^{-6}$) and filtered for common variants (MAF >1%). The GWAS was performed in PLINK 2.0 using the *glm* function, and corrected for age and 18 top PCs (estimated as the optimal number of top PCs for this sample based on λ calculation).

PRS were calculated for the cohort of 72,024 males using PRSice-2 (v2.3.5)⁵⁷ using autosomal and X-chromosomal SNPs $P < 7.85 \times 10^{-3}$ (best-fit PRS P -value threshold) and otherwise default settings. AUCs for the full cohort were computed using the pROC R package (v1.18)⁵⁸. The cohort of 72,024 males was split 25–50–25%, with 25% being used for weighting genes and summing all gene-based scores into one gene-based risk score per individual. Training of the integrated risk prediction models was performed using 50% of the samples with 10-fold cross-validation, with the remaining 25% of samples being used as an independent testing set. The risk prediction model was generated based on data from our continuous model with age, sequencing batch and the top 14 PCs being included as features. To evaluate the contribution of rare variants, the performances of risk models that included gene-based scores and PRS were compared with risk models that included PRS only.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

This research has been conducted using data from UK Biobank under Application Numbers 24661 and 102444. The individual-level genetic and phenotypic data are available under restricted access; access can be obtained by application through the UK Biobank platform. The data generated that support the findings of this study are provided in the Supplementary Data. The CADD score data used in this study are available in the University of Washington CADD score database https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/whole_genome_SNVs.tsv.gz. The gene feature annotation data used in this study are available in the Ensembl database under release number 104 https://ftp.ensembl.org/pub/release-104/gtf/homo_sapiens/Homo_sapiens.GRCh38.104.chr.gtf.gz and in the UCSC Genome Browser <https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/refFlat.txt.gz>. The ClinVar data used in this study are available from the ClinVar database https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/archive_2.0/2022/clinvar_20220430.vcf.gz.

References

- Hamilton, J. B. Patterned loss of hair in man; types and incidence. *Ann. N. Y. Acad. Sci.* **53**, 708–728 (1951).
- Stough, D. et al. Psychological effect, pathophysiology, and management of androgenetic alopecia in men. *Mayo Clin. Proc.* **80**, 1316–1322 (2005).
- Varothai, S. & Bergfeld, W. F. Androgenetic alopecia: an evidence-based treatment update. *Am. J. Clin. Dermatol.* **15**, 217–230 (2014).
- Traish, A. M., Hassani, J., Guay, A. T., Zitzmann, M. & Hansen, M. L. Adverse side effects of 5 α -reductase inhibitors therapy: persistent diminished libido and erectile dysfunction and depression in a subset of patients. *J. Sex. Med.* **8**, 872–884 (2011).
- Heath, A. C., Nyholt, D. R., Gillespie, N. A. & Martin, N. G. Genetic basis of male pattern baldness. *J. Invest. Dermatol.* **121**, 1561–1564 (2003).
- Rexbye, H. et al. Hair loss among elderly men: etiology and impact on perceived age | The Journals of Gerontology: Series A | Oxford Academic. *J. Gerontol.* **60**, 1077–1082 (2005).
- Li, R. et al. Six novel susceptibility Loci for early-onset androgenetic alopecia and their unexpected association with common diseases. *PLoS Genet.* **8**, e1002746 (2012).
- Heilmann-Heimbach, S. et al. Meta-analysis identifies novel risk loci and yields systematic insights into the biology of male-pattern baldness. *Nat. Commun.* **8**, 14694 (2017).
- Heilmann, S. et al. Androgenetic alopecia: identification of four genetic risk loci and evidence for the contribution of WNT signaling to its etiology. *J. Invest. Dermatol.* **133**, 1489–1496 (2013).
- Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
- Hillmer, A. M. et al. Susceptibility variants for male-pattern baldness on chromosome 20p11. *Nat. Genet.* **40**, 1279–1281 (2008).
- Richards, J. B. et al. Male-pattern baldness susceptibility locus at 20p11. *Nat. Genet.* **40**, 1282–1284 (2008).
- Yap, C. X. et al. Dissection of genetic variation and evidence for pleiotropy in male pattern baldness. *Nat. Commun.* **9**, 5407 (2018).
- Hagenaars, S. P. et al. Genetic prediction of male pattern baldness. *PLoS Genet.* **13**, e1006594 (2017).
- Pirastu, N. et al. GWAS for male-pattern baldness identifies 71 susceptibility loci explaining 38% of the risk. *Nat. Commun.* **8**, 1584 (2017).
- Brockschmidt, F. F. et al. Susceptibility variants on chromosome 7p21.1 suggest HDAC9 as a new candidate gene for male-pattern baldness. *Br. J. Dermatol.* **165**, 1293–1302 (2011).
- Adhikari, K. et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat. Commun.* **7**, 10815 (2016).
- Szustakowski, J. D. et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
- 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**, 1001779 (2015).
- Lee, S. et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
- Aldisi, R. et al. GenRisk: a tool for comprehensive genetic risk modeling. *Bioinformatics* **38**, 2651–2653 (2022).
- Herrera-Rivero, M., Hochfeld, L. M., Sivalingam, S., Nöthen, M. M. & Heilmann-Heimbach, S. Mapping of cis-acting expression quantitative trait loci in human scalp hair follicles. *BMC Dermatol.* **20**, 16 (2020).
- Betz, R. C., Cabral, R. M., Christiano, A. M. & Sprecher, E. Unveiling the roots of monogenic genodermatoses: genotrichoses as a paradigm. *J. Invest. Dermatol.* **132**, 906–914 (2012).

24. Wright, J. T. et al. Ectodermal dysplasias: classification and organization by phenotype, genotype and molecular pathway. *Am. J. Med. Genet. A* **179**, 442–447 (2019).
25. Hayashi, R. & Shimomura, Y. Update of recent findings in genetic hair disorders. *J. Dermatol.* **49**, 55–67 (2022).
26. Ü Basmanav, F. B. et al. Mutations in three genes encoding proteins involved in hair shaft formation cause uncombable hair syndrome. *Am. J. Hum. Genet.* **99**, 1292–1304 (2016).
27. Duverger, O. & Morasso, M. I. To grow or not to grow: hair morphogenesis and human genetic hair disorders. *Semin Cell Dev. Biol.* **25–26**, 22–33 (2014).
28. Peled, A. et al. Mutations in TSPEAR, encoding a regulator of notch signaling, affect tooth and hair follicle morphogenesis. *PLoS Genet.* **12**, 1006369 (2016).
29. Krøigård, A. B., Clemmensen, O., Gjørup, H., Hertz, J. M. & Bygum, A. Odonto-onycho-dermal dysplasia in a patient homozygous for a WNT10A nonsense mutation and mild manifestations of ectodermal dysplasia in carriers of the mutation. *BMC Dermatol.* **16**, 3 (2016).
30. Heilmann-Heimbach, S., Hochfeld, L. M., Paus, R. & Nöthen, M. M. Hunting the genes in male-pattern alopecia: how important are they, how close are we and what will they tell us? *Exp. Dermatol.* **25**, 251–257 (2016).
31. Hochfeld, L. M. et al. Evidence for a functional interaction of WNT10A and EBF1 in male-pattern baldness. *PLoS One* **16**, e0256846 (2021).
32. Helman, S. L. et al. The biology of mammalian multi-copper ferroxidases. *BioMetals* **36**, 263–281 (2022).
33. Funai, K. et al. Skeletal muscle phospholipid metabolism regulates insulin sensitivity and contractile function. *Diabetes* **65**, 358–370 (2016).
34. Choi, S. H. et al. Hair growth promoting potential of phospholipids purified from porcine lung tissues. *Biomol. Ther. (Seoul.)* **23**, 174–179 (2015).
35. Yamamoto, K. et al. Hair follicular expression and function of group X secreted phospholipase A2 in mouse skin. *J. Biol. Chem.* **286**, 11616 (2011).
36. Nicoli, E.-R. et al. eP198: EIF3F compound heterozygous genotype-phenotype association. *Genet. Med.* **24**, S123 (2022).
37. Groza, T. et al. The international mouse phenotyping consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.* **51**, D1038–D1045 (2023).
38. Heilmann-Heimbach, S., Hochfeld, L. M., Henne, S. K. & Nöthen, M. M. Hormonal regulation in male androgenetic alopecia-Sex hormones and beyond: evidence from recent genetic studies. *Exp. Dermatol.* **29**, 814–827 (2020).
39. Gomes-Duarte, A., Lacerda, R., Menezes, J. & Romão, L. eIF3: a factor for human health and disease. *RNA Biol.* **15**, 26–34 (2018).
40. Moretti, J. et al. The translation initiation factor 3f (eIF3f) exhibits a deubiquitinase activity regulating notch activation. *PLoS Biol.* **8**, 1000545 (2010).
41. Hu, X. M. et al. A systematic summary of survival and death signalling during the life of hair follicle stem cells. *Stem Cell Res. Ther.* **12**, 453 (2021).
42. Hillmer, A. M. et al. Recent positive selection of a human androgen receptor/ectodysplasin A2 receptor haplotype and its relationship to male pattern baldness. *Hum. Genet.* **126**, 255–264 (2009).
43. Henne, S. K., Nöthen, M. M. & Heilmann-Heimbach, S. Male-pattern hair loss: comprehensive identification of the associated genes as a basis for understanding pathophysiology. *Medizinische Genetik* **35**, 3–14 (2023).
44. Cobb, J. E., White, S. J., Harrap, S. B. & Ellis, J. A. Androgen receptor copy number variation and androgenetic alopecia: a case-control study. *PLoS ONE* **4**, e5081 (2009).
45. Brockschmidt, F. F. et al. Fine mapping of the humanAR/EDA2R locus in androgenetic alopecia. *Br. J. Dermatol.* **162**, 899–903 (2010).
46. Karczewski, K. J. et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom.* **2**, 100168 (2022).
47. Wang, Q. et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nat.* **2021** 597:7877 **597**, 527–532 (2021).
48. Severi, G. et al. Androgenetic alopecia in men aged 40–69 years: Prevalence and risk factors. *Br. J. Dermatol.* **149**, 1207–1213 (2003).
49. Aldisi R. et al. Gene-based burden scores identify rare variant associations for 28 blood biomarkers. Preprint at <https://doi.org/10.21203/RS.3.RS-2271894/V1> (2023).
50. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
51. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
52. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 1–12 (2021).
53. Karolchik, D. et al. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
54. O’Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733 (2016).
55. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
56. 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 (2019).
57. Choi, S. W. & O’Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).
58. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* **12**, 77 (2011).

Acknowledgements

This research was conducted using the UK Biobank resource under Application Numbers 24661 and 102444. We would like to thank the Core Unit for Bioinformatics Data Analysis for providing computing resources and support in setting up the initial pipeline. We further thank Christine Schmal for her proofreading and feedback on the manuscript.

Author contributions

L.M.H., O.B., P.M.K., C.M., M.M.N., S.H.-H. conceived and designed the research goals and analyses. S.H.-H. supervised the project. P.M.K. provided computing resources and analysis tools. S.K.H., S.S., R.A. curated the data. S.K.H., R.A., S.S., O.B., C.M. performed the analyses. S.K.H. wrote the initial manuscript draft and S.K.H., R.A., L.M.H., C.M., M.M.N., S.H.-H. revised the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare the following competing interests: M.M.N., S.H.-H. and L.M.H. receive salary payments from Life & Brain GmbH and M.M.N. holds shares in Life & Brain GmbH. M.M.N. has received fees for membership in an Advisory Board of HMG Systems Engineering GmbH, and for membership in the Medical-Scientific Editorial Office of the Deutsches Ärzteblatt. M.M.N. is a member of the excellence cluster ImmunoSensation². All this concerned activities outside the submitted work. S.K.H., R.A., S.S., O.B., P.M.K. and C.M. declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-41186-w>.

Correspondence and requests for materials should be addressed to Stefanie Heilmann-Heimbach.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

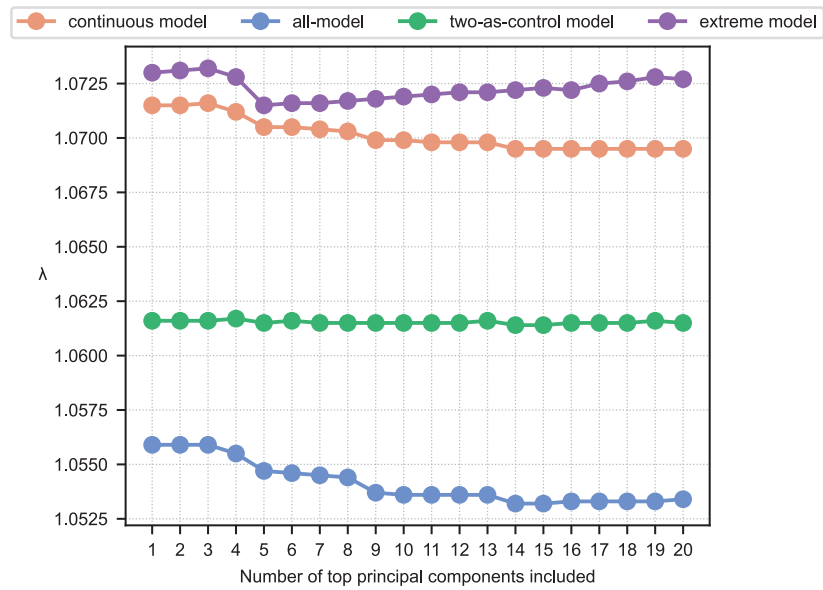
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

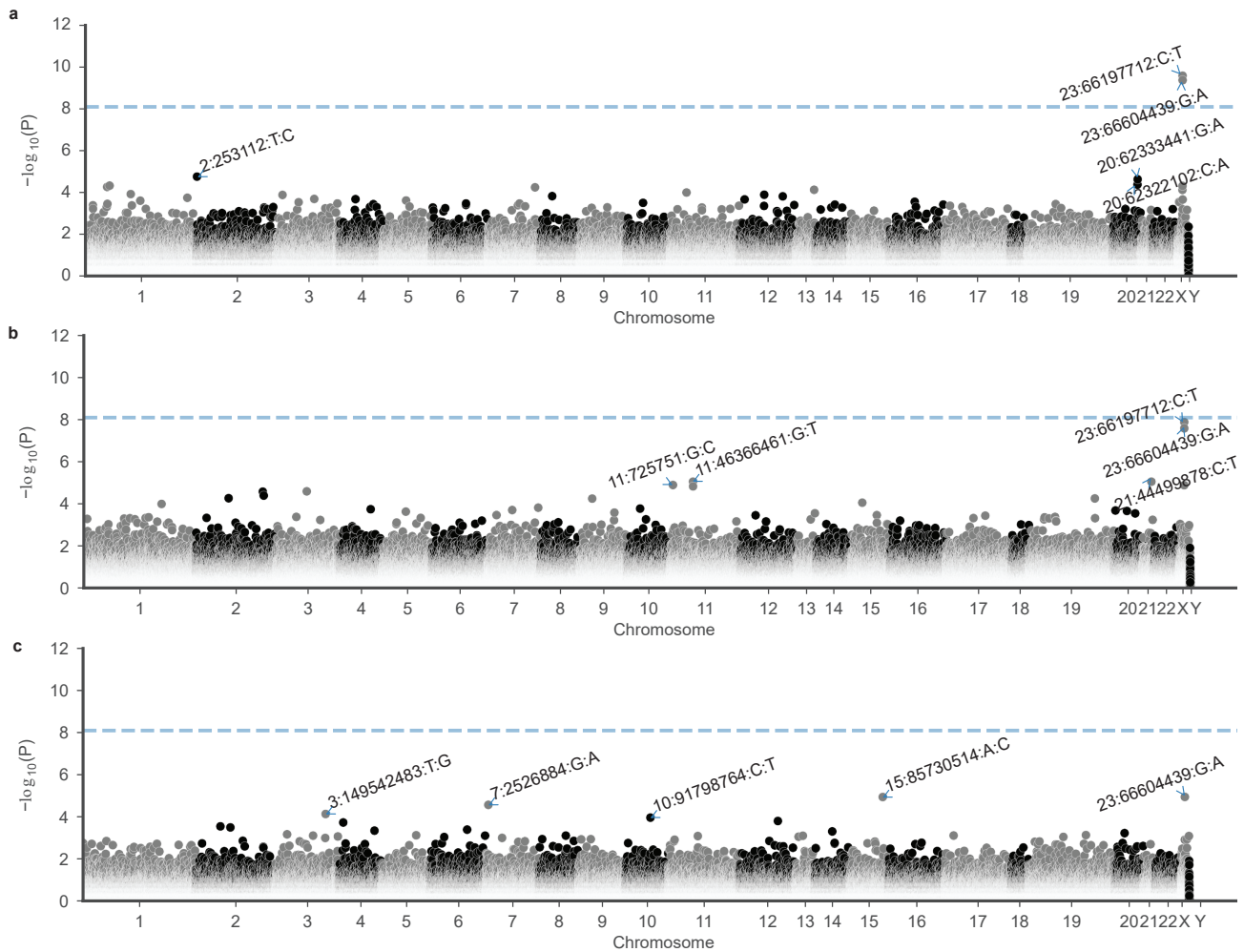
3.3.1 Publication 3 - Appendix A

This appendix contains the supplementary material for publication 3. The supplementary figures are shown below. Because of size issues, the tables and other supplementary information cannot be included in this thesis, but all files can be directly downloaded from the paper using this link: <https://doi.org/10.1038/s41467-023-41186-w>

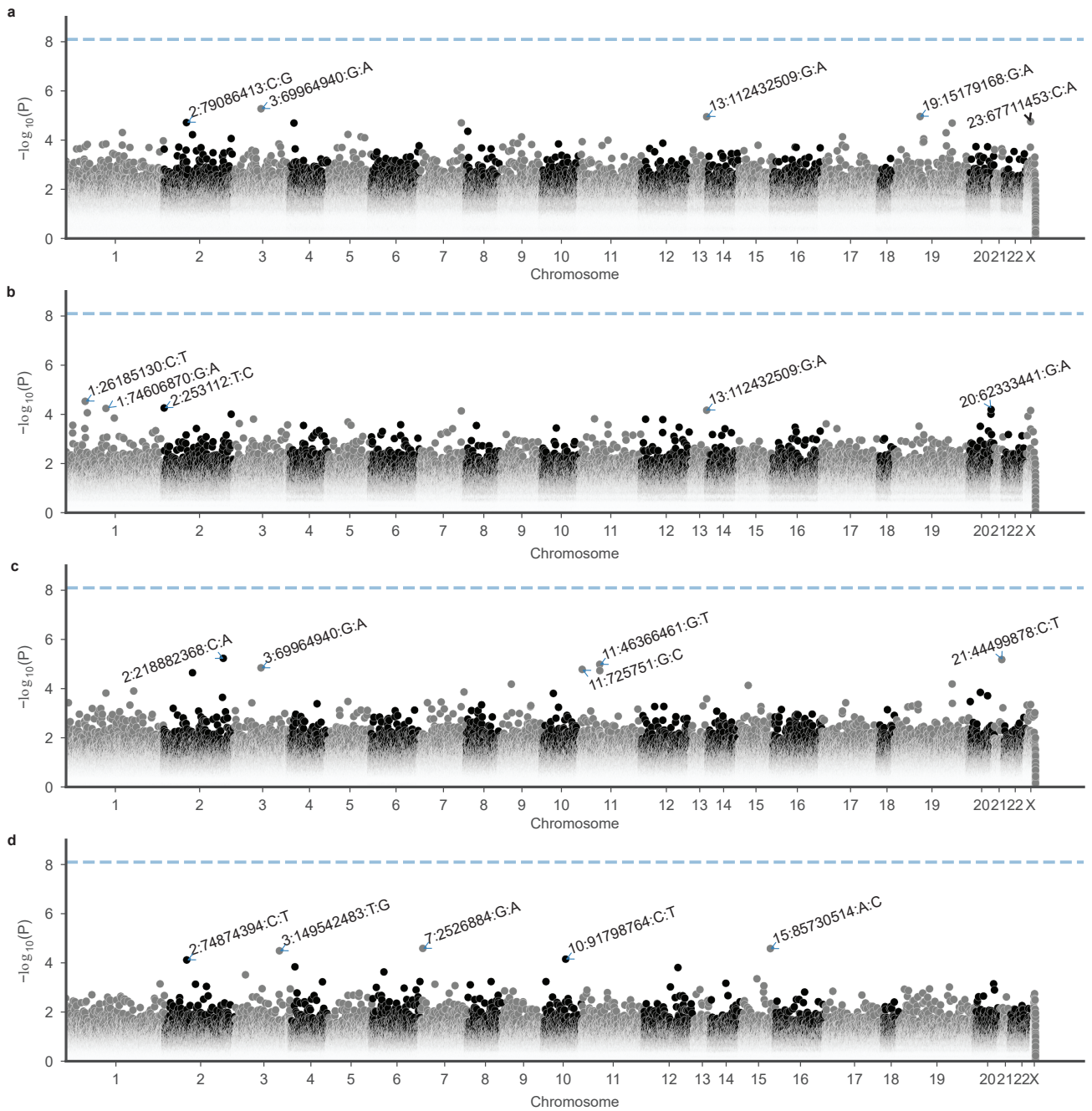
Supplementary Figures



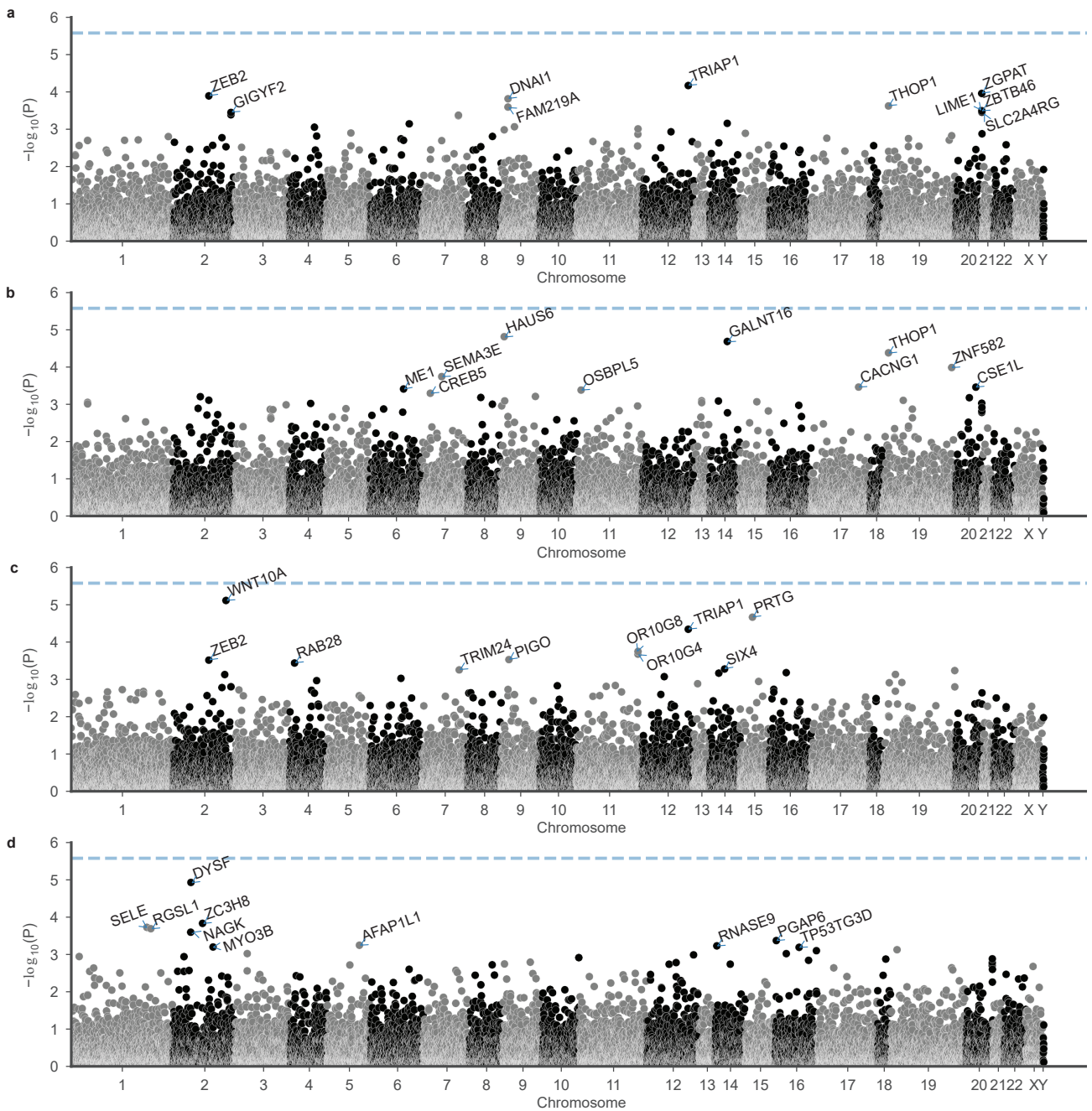
Supplementary Figure 1: Genomic inflation factor λ in GWAS with a varying number of included principal components. Genomic inflation factor λ according to the number of top principal components corrected for in a GWAS of imputed genotype data in the continuous model (orange), all-model (blue), the two-as-control model (green) and the extreme model (purple). The λ values generated were lowest when using 14-15 PCs in the continuous, all- and two-as-control models and 5 PCs in the extreme model, respectively.



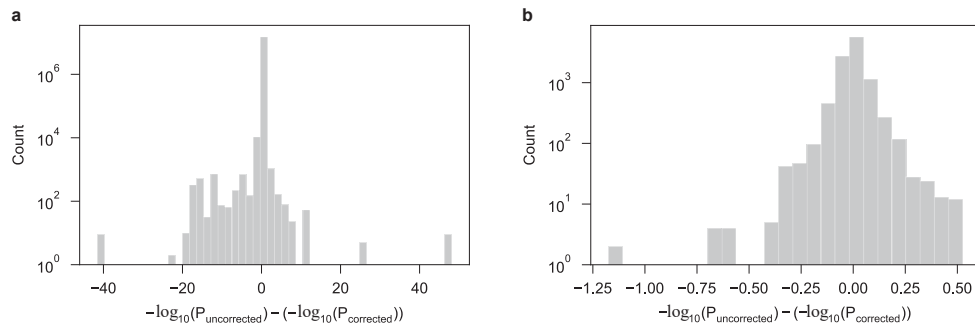
Supplementary Figure 2: Results of the single-variant analysis of additional phenotype models. Results are shown for **a** the all-model; **b** the two-as-control model; and **c** the extreme model. Only variants that were tested in the respective SKAT-O analysis are included. The dashed line denotes the selected genome-wide threshold for multiple testing in single-variant tests (8×10^{-9}). The y-axes depict $-\log_{10}(P)$ obtained from logistic regression (two-sided, unadjusted). The top 5 variants per analysis were annotated.



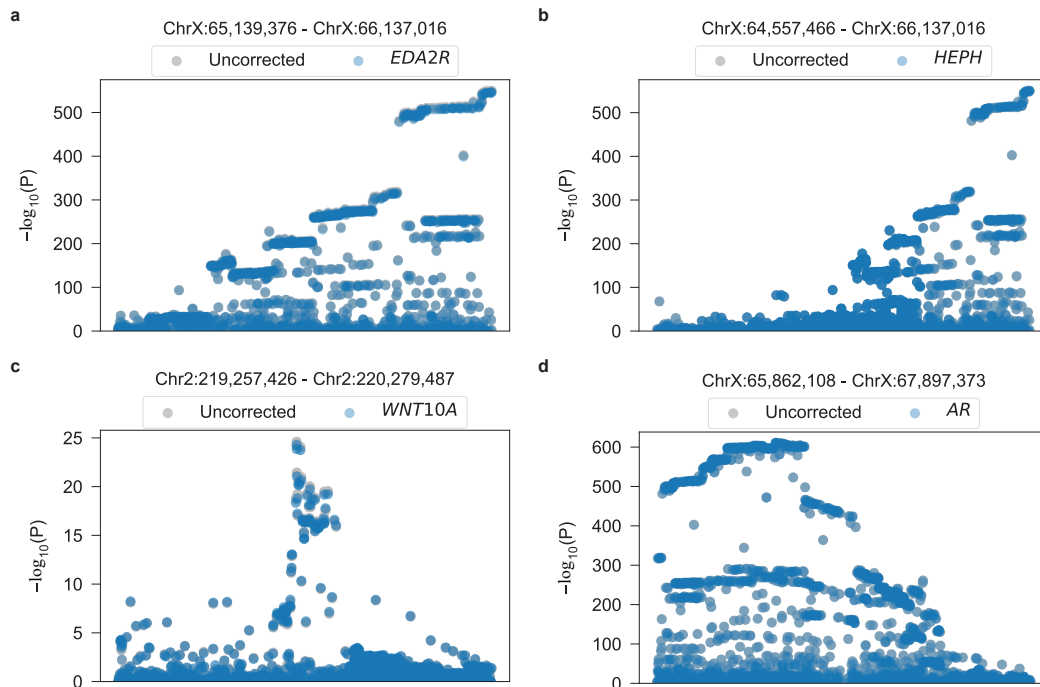
Supplementary Figure 3: Results of the single-variant analysis conditioned for 622 GWAS lead SNPs. Results are shown for **a** the continuous model; **b** the all-model; **c** the two-as-control model; and **d** the extreme model. Only variants that were tested in the respective SKAT-O analysis are included. The dashed line denotes the selected genome-wide threshold for multiple testing in single-variant tests (8×10^{-9}). The y-axes depict $-\log_{10}(P)$ obtained from linear regression (a) or logistic regression (b – d) (two-sided, unadjusted). The top 5 variants per analysis were annotated.



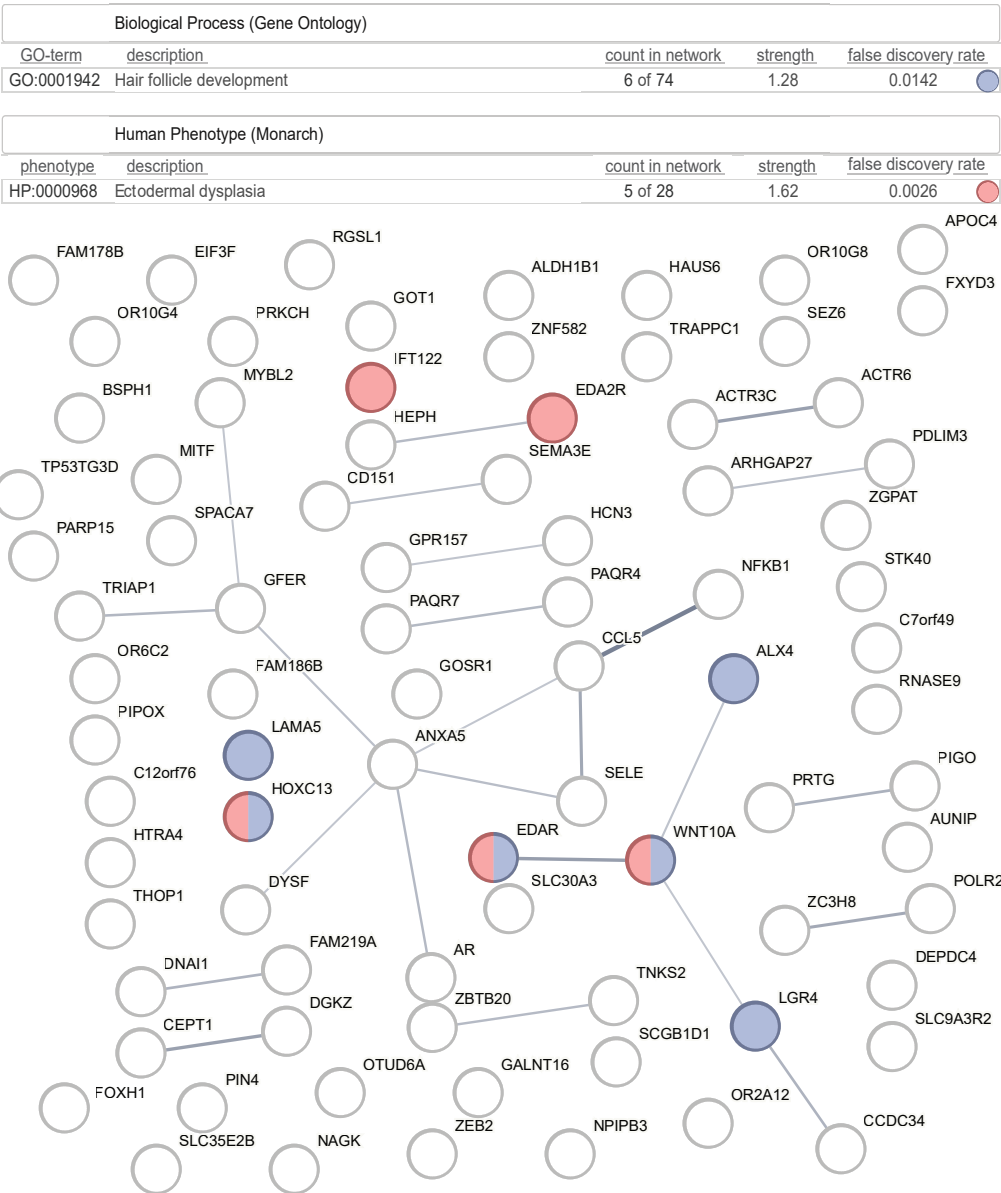
Supplementary Figure 4: Results of the SKAT-O gene-based analysis with a high impact variant threshold. Results are shown for **a** the continuous model; **b** the all-model; **c** the two-as-control model; and **d** the extreme model. The dashed line denotes the Bonferroni threshold for multiple testing in SKAT-O analyses (2.6×10^{-6}). The y-axes depict $-\log_{10}(P)$ obtained from linear regression (a) or logistic regression (b – d) (two-sided, unadjusted). The top 10 genes per analysis were annotated.



Supplementary Figure 5: Distribution of the decrease in $-\log_{10}(P)$ -values in the conditional GWAS-GenRisk analysis. Distributions are shown across **a** all tested common variants and **b** GWAS lead SNPs. $P_{\text{corrected}}$ refers to the p-value generated when correcting for any single GenRisk gene score of a gene at the respective locus. Uncorrected and corrected p-values were generated by linear regression (two-sided, unadjusted).





Supplementary Figure 6: Conditional GWAS-GenRisk results. Results are shown for the loci of **a** *EDA2R*, **b** *HEPH*, **c** *WNT10A*, and **d** *AR*. Association results without correction for GenRisk gene scores are shown in gray, association results after correction for GenRisk gene scores of a single gene are shown in blue, with the gene denoted in the respective legend. The y-axes depict $-\log_{10}(P)$ obtained from linear regression (two-sided, unadjusted).





Supplementary Figure 7: Results of the STRING protein interaction network analysis of a less stringent set of MPHL-associated genes ($P < 3 \times 10^{-4}$ in the SKAT-O or GenRisk analyses). Functional enrichments performed by STRING based on one-sided hypergeometric tests are shown in the top panels, along with the corresponding false discovery rate (p-value corrected using the Benjamini-Hochberg method). Genes belonging to the enriched functional annotation terms are shown in blue (hair follicle development), orange (ectodermal dysplasia), or both. Line thickness indicates the strength of the data supporting the interaction.

Which of the following best describes your hair/balding pattern?

 Pattern 1


 Pattern 2


 Pattern 3


 Pattern 4


Do not know

Prefer not to answer

 Back

 Info

 Help

 Next

Supplementary Figure 8: Screenshot of the touchscreen questionnaire used to capture state of hair/balding pattern in the UK Biobank. Reproduced by kind permission of UK Biobank ©.

Supplementary Tables

Supplementary Table 1: List of known causative genes for monogenic trichoses, as used for enrichment testing.

Gene symbol	Ensembl ID	Gene name	Condition	References
<i>ABCA5</i>	ENSG00000154265	ATP Binding Cassette Subfamily A Member 5	Generalized hypertrichosis	¹
<i>ADAM17</i>	ENSG00000151694	ADAM Metallopeptidase Domain 17	Structural hair defects	²
<i>ALX4</i>	ENSG00000052850	ALX Homeobox 4	Total alopecia in frontonasal dysplasia	²
<i>ANTXR1</i>	ENSG00000169604	ANTXR Cell Adhesion Molecule 1	Growth retardation, alopecia, pseudoanodontia (GAPO) syndrome	²
<i>APCDD1</i>	ENSG00000154856	APC Down-Regulated 1	Hypotrichosis 1	¹⁻³
<i>ATP7A</i>	ENSG00000165240	ATPase Copper Transporting Alpha	Menkes disease	²
<i>BCS1L</i>	ENSG00000074582	BCS1 Homolog, Ubiquinol-Cytochrome C Reductase Complex Chaperone	Björnstad syndrome	^{1,3}
<i>C3orf52</i>	ENSG00000114529	Chromosome 3 Open Reading Frame 52	Woolly hair/hypotrichosis	¹
<i>CDH3</i>	ENSG00000062038	Cadherin 3	Ectodermal dysplasia, ectrodactyly, and macular dystrophy syndrome	²⁻⁴
<i>CDSN</i>	ENSG00000204539	Corneodesmosin	Hypotrichosis 2	¹⁻³
<i>CLDN1</i>	ENSG00000163347	Claudin 1	Ichthyosis, leukocyte vacuoles, alopecia, and sclerosing cholangitis	²
<i>DCAF17</i>	ENSG00000115827	DDB1 And CUL4 Associated Factor 17	Woodhouse-Sakati syndrome	²
<i>DLX3</i>	ENSG00000064195	Distal-Less Homeobox 3	Trichodontoosseous syndrome	²

<i>DSC3</i>	ENSG00000134762	Desmocollin 3	Hypotrichosis and recurrent skin vesicles	2,3
<i>DSG4</i>	ENSG00000175065	Desmoglein 4	Hypotrichosis 6	1-3
<i>DSP</i>	ENSG00000096696	Desmoplakin	Carvajal syndrome	1-3
<i>EDA</i>	ENSG00000158813	Ectodysplasin A	Ectodermal dysplasia	1-4
<i>EDAR</i>	ENSG00000135960	Ectodysplasin A Receptor	Ectodermal dysplasia	1-4
<i>EDARADD</i>	ENSG00000186197	EDAR Associated Death Domain	Ectodermal dysplasia	1-4
<i>EPS8L3</i>	ENSG00000198758	EPS8 Like 3	Hypotrichosis 5	1
<i>ERCC2</i>	ENSG00000104884	ERCC Excision Repair 2, TFIIH Core Complex Helicase Subunit	Trichothiodystrophy	2
<i>ERCC3</i>	ENSG00000163161	ERCC Excision Repair 3, TFIIH Core Complex Helicase Subunit	Trichothiodystrophy	2
<i>FGF13</i>	ENSG00000129682	Fibroblast Growth Factor 13	Generalized hypertrichosis	1,2
<i>FOXE1</i>	ENSG00000178919	Forkhead Box E1	Hypothyroidism, with spiky hair and cleft palate	2
<i>FOXN1</i>	ENSG00000109101	Forkhead Box N1	T-cell immunodeficiency, alopecia, and nail dystrophy	2
<i>GJB6</i>	ENSG00000121742	Gap Junction Protein Beta 6	Clouston syndrome	1,2
<i>GTF2H5</i>	ENSG00000272047	General Transcription Factor IIH Subunit 5	Trichothiodystrophy	2
<i>HOXC13</i>	ENSG00000123364	Homeobox C13	Pure hair and nail ectodermal dysplasia	1,2

<i>HR</i>	ENSG00000168453	HR Lysine Demethylase And Nuclear Receptor Corepressor	Hypotrichosis 4, alopecia universalis	²
<i>IKBKG</i>	ENSG00000269335	Inhibitor Of Nuclear Factor Kappa B Kinase Regulatory Subunit Gamma	Incontinentia pigmenti, alopecia	^{2,4}
<i>JUP</i>	ENSG00000173801	Junction Plakoglobin	Naxos disease	¹⁻³
<i>KRT25</i>	ENSG00000204897	Keratin 25	Woolly hair/hypotrichosis	¹
<i>KRT71</i>	ENSG00000139648	Keratin 71	Woolly hair	^{1,2}
<i>KRT74</i>	ENSG00000170484	Keratin 74	Hypotrichosis 3, woolly hair	¹⁻³
<i>KRT75</i>	ENSG00000170454	Keratin 75	Pseudofolliculitis barbae	^{2,3}
<i>KRT81</i>	ENSG00000205426	Keratin 81	Monilethrix	¹⁻⁴
<i>KRT83</i>	ENSG00000170523	Keratin 83	Monilethrix	¹⁻⁴
<i>KRT85</i>	ENSG00000135443	Keratin 85	Pure hair and nail ectodermal dysplasia	¹⁻⁴
<i>KRT86</i>	ENSG00000170442	Keratin 86	Monilethrix	¹⁻⁴
<i>LIPH</i>	ENSG00000163898	Lipase H	Hypotrichosis 7	¹⁻³
<i>LPAR6</i>	ENSG00000139679	Lysophosphatidic Acid Receptor 6	Hypotrichosis 8	¹⁻³
<i>LSS</i>	ENSG00000160285	Lanosterol Synthase	Hypotrichosis 14	¹
<i>MBTPS2</i>	ENSG00000012174	Membrane Bound Transcription Factor Peptidase, Site 2	Ichthyosis follicularis, atrichia, and photophobia syndrome	^{2,3}
<i>MPLKIP</i>	ENSG00000168303	M-Phase Specific PLK1 Interacting Protein	Trichothiodystrophy	²

<i>NECTIN1</i>	ENSG00000110400	Nectin Cell Adhesion Molecule 1	Cleft lip/palate-ectodermal dysplasia syndrome	1,4
<i>PADI3</i>	ENSG00000142619	Peptidyl Arginine Deiminase 3	Uncombable hair syndrome	5
<i>PKP1</i>	ENSG00000081277	Plakophilin 1	Ectodermal dysplasia/skin fragility syndrome	2,4
<i>PORCN</i>	ENSG00000102312	Porcupine O-Acyltransferase	Goltz syndrome	4
<i>RBM28</i>	ENSG00000106344	RNA Binding Motif Protein 28	Alopecia, neurological defects, and endocrinopathy syndrome	2,3
<i>RIN2</i>	ENSG00000132669	Ras And Rab Interactor 2	Macrocephaly, alopecia, cutis laxa, and scoliosis	2
<i>RPL21</i>	ENSG00000122026	Ribosomal Protein L21	Hypotrichosis 12	1,3
<i>SLC29A3</i>	ENSG00000198246	Solute Carrier Family 29 Member 3	Histiocytosis-lymphadenopathy plus syndrome	2,3
<i>SNRPE</i>	ENSG00000182004	Small Nuclear Ribonucleoprotein Polypeptide E	Hypotrichosis 11	1,2
<i>SOX18</i>	ENSG00000203883	SRY-Box Transcription Factor 18	Hypotrichosis-lymphedema-telangiectasia syndrome	2,3
<i>SOX9</i>	ENSG00000125398	SRY-Box Transcription Factor 9	Hypertrichosis terminalis	1,2
<i>SPINK5</i>	ENSG00000133710	Serine Peptidase Inhibitor Kazal Type 5	Netherton syndrome	2,3
<i>ST14</i>	ENSG00000149418	ST14 Transmembrane Serine Protease Matriptase	Ichthyosis with hypotrichosis	2,3

<i>TCHH</i>	ENSG00000159450	Trichohyalin	Uncombable hair syndrome	5
<i>TGM3</i>	ENSG00000125780	Transglutaminase 3	Uncombable hair syndrome	5
<i>TP63</i>	ENSG00000073282	Tumor Protein P63	Ectodermal dysplasia	1,2,4
<i>TRAF6</i>	ENSG00000175104	TNF Receptor Associated Factor 6	Ectodermal dysplasia	4
<i>TRPS1</i>	ENSG00000104447	Transcriptional Repressor GATA Binding 1	Trichorhinophalangeal syndrome	1,2
<i>VDR</i>	ENSG00000111424	Vitamin D Receptor	Vitamin D-dependent rickets with alopecia	2,3
<i>WNT10A</i>	ENSG00000135925	Wnt Family Member 10A	Odonto-onycho-dermal dysplasia, Schopf-Schulz-Passarge syndrome	1,2,4

Supplementary References

1. Hayashi, R. & Shimomura, Y. Update of recent findings in genetic hair disorders. *J Dermatol* **49**, 55–67 (2022).
2. Duverger, O. & Morasso, M. I. To grow or not to grow: Hair morphogenesis and human genetic hair disorders. *Semin Cell Dev Biol* **25–26**, 22–33 (2014).
3. Betz, R. C., Cabral, R. M., Christiano, A. M. & Sprecher, E. Unveiling the roots of monogenic genodermatoses: Genotrichoses as a paradigm. *Journal of Investigative Dermatology* vol. 132 906–914 Preprint at <https://doi.org/10.1038/jid.2011.408> (2012).
4. Wright, J. T. *et al.* Ectodermal dysplasias: Classification and organization by phenotype, genotype and molecular pathway. *Am J Med Genet A* **179**, 442–447 (2019).
5. Ü. Basmanav, F. B. *et al.* Mutations in Three Genes Encoding Proteins Involved in Hair Shaft Formation Cause Uncombable Hair Syndrome. *Am J Hum Genet* **99**, 1292–1304 (2016).

3.3.2 Publication 3 - Appendix B

This appendix clarifies the machine learning methods used in publication 3.

Association analyses

The gene-based association analyses were done using linear regression for the continuous model and logistic regression for the binary models. The covariates used are mentioned in the paper.

Risk modeling

For risk modeling, different machine learning models were trained with cross-validation using a training set to select the best fitting model, based on the AUC, for the each MPHL model, while a part of the dataset was left out for external evaluation, as explained in the paper. The list of models tested can be seen in the GenRisk documentations (see subsection 3.1.1). Table B.1 includes which ML model was used for each MPHL model.

Table B.1 List of the best fit ML model for each MPHL model.

MPHL model	ML type
Severe (pattern 1 vs 4) PRS only	Logistic Regression
Severe (pattern 1 vs 4) gene-based scores only	Gradient Boosting Classifier
Severe (pattern 1 vs 4) PRS + gene-based scores	Logistic Regression
Moderate (pattern 1 vs 3-4) PRS only	Gradient Boosting Classifier
Moderate (pattern 1 vs 3-4) gene-based scores only	Gradient Boosting Classifier
Moderate (pattern 1 vs 3-4) PRS + gene-based scores	Gradient Boosting Classifier
Slight (pattern 1 vs 2-4) PRS only	Logistic Regression
Slight (pattern 1 vs 2-4) gene-based scores only	Gradient Boosting Classifier
Slight (pattern 1 vs 2-4) PRS + gene-based scores	Logistic Regression

4. Discussion with references

A relevant part of the genetic architecture of complex traits is yet unknown, despite the many studies that have been done on it. To investigate this, more and more methods and hypotheses are being developed. In this thesis, I focus on the effects of rare pathogenic variants on complex phenotypes. In general, we hypothesize that rare high effect variants contribute to the genetics of complex traits and can account for part of the missing heritability. Furthermore, I also study the integration of rare and common variants into genetic risk assessment pipeline and how that might improve genetic risk prediction for complex phenotypes.

In the first publication, we present our own implemented python tool, GenRisk, which we developed to serve as a framework for the following publications and studies. GenRisk implements a gene-based scoring system that upweights rare and pathogenic variants, a method that has been previously used in other applications (Mossotto et al., 2019; Curtis, 2022). The pipeline takes in a VCF as an input file, along with an annotation file that contains the allele frequency and a pathogenicity score of the users' choice and outputs a matrix of gene-based scores for each individual in the cohort. This matrix can then be further used for association analysis and predication modeling, both of which are also implemented in the pipeline. We have also implemented a module for calculating the PRS for the cohort. After the publication of this paper, we have further maintained and improved on the GenRisk tool by allowing binary PLINK (Chang et al., 2015) files (i.e. bed/bim/fam) as input for the scoring module. We also added in a new module for pathway-based scoring, in which the gene-based scores in each pathway are summed then normalized by number of genes in said pathway.

While there are many rare variants testing pipelines, like RVtests (Zhan et al., 2016), which provides a framework for multiple single-variant level and gene-level burden test, these pipelines only provide summary statistics output for a specific phenotype. On the other hand, GenRisk provides a framework which outputs an individual-level burden score matrix that can be reused for multiple downstream analyses and phenotypes. The pipeline has many modules, which can be used on the gene-based scores matrix provided by GenRisk, but also on any other matrix of the same structure. Furthermore, the weighting of the allele frequency, the pathogenicity scores and the threshold of the allele frequency can be determined by the user, allowing a more flexible calculation of the scores. For example, pathogenicity scores that have been generated recently like MetaRNN (Li et al., 2022), which uses deep learning approaches, can be used to generate the GenRisk scores. The parameters of the allele frequency weighting functions can be adjusted to upweight or downweight rare or common variants. This allows the user to adjust the GenRisk scores

based on the purpose of the project and the downstream analyses.

Our second publication presents an application of GenRisk pipeline on 28 blood biomarkers from the UK biobank database. Association analyses were able to identify gene-phenotype associations between the gene-based scores and the different blood biomarkers, with distinct and well-known associations like *PCSK9* and *LDLR* association with LDL levels (Cuchel et al., 2014; Reiss et al., 2018), and *SHBG* gene association with SHBG and testosterone levels (Winters, 2020). Our association analyses were also able to detect some interesting novel associations that were not identified with other methods, such as *THRA* association with the liver function biomarkers aspartate aminotransferase and alanine aminotransferase levels (Piantanida et al., 2020). Such findings confirm that rare variants do contribute to the genetics of complex phenotypes. We further performed prediction model generation using different machine learning (ML) models to investigate the ability of rare variants to predict genetic risk and the combined effect of rare and common variants on genetic risk prediction. Our results indicated that common variants, i.e. PRS, was more informative in risk prediction and that rare variants added little to the combined rare and common variants model. This did not come as a surprise, as rare variants effects are hard to detect at a population-level and thus, are more useful on individual-level data. While many studies have conducted common and rare variant analysis on complex phenotypes separately, our paper presents the combined effect of the two variant types.

In the last publication, we investigate the contribution of rare variants on MPHL, which was also done on the UK biobank cohort. Both SKAT-O and GenRisk association analyses were performed. The analyses identified genes that were previously implicated by GWAS such as *EDA2R* and *WNT10A* (Yap et al., 2018), which indicate that both common and rare variants in these genes have effect on MPHL. *HEPH* was also detected in the association analyses and while it hasn't been previously reported, there are studies that suggest this gene plays a role in hair development (Helman et al., 2022). We further performed a GWAS conditional analysis using GenRisk's gene-based scores to investigate whether there is a dependency between the common GWAS-implicated variants and the rare variants from the GenRisk analysis. The results showed no dependency between those two variant groups. The risk prediction modeling performed similarly to the previous paper, rare variants contributed only marginally to the genetic risk prediction. This further confirms the contribution of rare variants to complex phenotypes at an individual-level, even if they provide little information at population-level. Here, we present that while rare variants do contribute to MPHL phenotype, their effect is independent of the common variants, and thus, neither factor should be ignored when accounting for the genetic risk. We also provide an example to the benefits of having a gene-based scores matrix at an individual-level,

which we were able to utilize for performing the conditional GWAS analysis.

In summary, this thesis investigates the contribution of rare variants to the genetic landscape of complex phenotypes. We applied our own pipeline, GenRisk, to multiple phenotypes, both quantitative and binary, to identify associated genes and generate genetic risk prediction models. Our findings suggest that while rare variants add little contribution to the genetic risk prediction at population-level, they contribute significantly to the etiology of complex traits at an individual-level, providing more gene targets and adding to the insight of the genetic architecture of complex phenotypes.

4.1 Limitations and future outlook

While I have conducted a thorough investigation of rare variants in complex phenotypes, it is noteworthy to mention that I have majorly worked on UK biobank data, and more specifically white British cohort of UK biobank data. This means that the findings in this thesis only account for this specific ethnicity and further studies need to be conducted on different cohorts to validate the findings. One of our recent collaborations is with Qatar biobank, in which we have obtained whole genome sequencing data for 14,000 individuals from Qatar.

It is also important to mention that this study was focused on rare pathogenic high effect variants, so rare variants with low effect sizes were most likely discarded or underpowered in the analysis. Thus, we are further working on a pathway-specific analysis that combines the gene-based scores into pathway-based scores which then can be used for further downstream analysis. This analysis can help account for the rare variants' underpower problem. Our preliminary results showed an improvement in genetic risk assessment when using pathway-based scores in ML models, but further analysis is needed to confirm these observations.

Furthermore, we were unable to include all gene-based scores in the prediction modeling because of high-dimensionality and computational power, thus, feature selection had to be made. This means that non-linear associations like gene-gene interactions might've been lost. Using the pathway-specific pipeline previously mentioned is one way to solve this problem; by combining the gene-based scores into pathway-based scores, we are significantly reducing the dimensionality of the data but we might still loss some of the gene-gene interactions specially with genes within one pathway. Another way to solve this issue is to use approaches that were adapted to deal with high-dimensional large-scale data like snpboost (Klinkhammer et al., 2023), or deep learning approaches (Sigurdsson et al., 2023).

References

- Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee (2015). Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1).
- M. Cuchel, E. Bruckert, H. N. Ginsberg, F. J. Raal, R. D. Santos, et al. (2014). Homozygous familial hypercholesterolaemia: new insights and guidance for clinicians to improve detection and clinical management. a position paper from the consensus panel on familial hypercholesterolaemia of the european atherosclerosis society. *European Heart Journal*, 35(32):2146–2157.
- David Curtis (2022). Exploration of weighting schemes based on allele frequency and annotation for weighted burden association analysis of complex phenotypes. *Gene*, 809:146039.
- Sheridan L. Helman, Jie Zhou, Brie K. Fuqua, Yan Lu, James F. Collins, et al. (2022). The biology of mammalian multi-copper ferroxidases. *BioMetals*, 36(2):263–281.
- Hannah Klinkhammer, Christian Staerk, Carlo Maj, Peter Michael Krawitz, and Andreas Mayr (2023). A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Frontiers in Genetics*, 13.
- Chang Li, Degui Zhi, Kai Wang, and Xiaoming Liu (2022). Metarnn: differentiating rare pathogenic and rare benign missense snvs and indels using deep learning. *Genome Medicine*, 14(1).
- E. Mossotto, J. J. Ashton, L. O’Gorman, R. J. Pengelly, R. M. Beattie, B. D. MacArthur, and S. Ennis (2019). Genepy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinformatics*, 20(1).
- E. Piantanida, S. Ippolito, D. Gallo, E. Masiello, P. Premoli, et al. (2020). The interplay between thyroid and liver: implications for clinical practice. *Journal of Endocrinological Investigation*, 43(7):885–899.
- Allison B. Reiss, Neal Shah, Dalia Muhieddine, Juan Zhen, Jennifer Yudkevich, Lora J. Kasselmann, and Joshua DeLeon (2018). Pcsk9 in cholesterol metabolism: from bench to bedside. *Clinical Science*, 132(11):1135–1153.
- Arnór I Sigurdsson, Ioannis Louloudis, Karina Banasik, David Westergaard, Ole Winther, et al. (2023). Deep integrative models for large-scale human genomics. *Nucleic Acids Research*, 51(12):e67–e67.
- Stephen J. Winters (2020). Shbg and total testosterone levels in men with adult onset hypogonadism: what are we overlooking? *Clinical Diabetes and Endocrinology*, 6(1).

- Chloe X. Yap, Julia Sidorenko, Yang Wu, Kathryn E. Kemper, Jian Yang, et al. (2018). Dissection of genetic variation and evidence for pleiotropy in male pattern baldness. *Nature Communications*, 9(1).
- Xiaowei Zhan, Youna Hu, Bingshan Li, Goncalo R. Abecasis, and Dajiang J. Liu (2016). Rvtests: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics*, 32(9):1423–1426.

5. Acknowledgment

I would like to express my gratitude to my advisor Prof. Dr. Peter Krawitz, for his invaluable supervision and guidance. I'm also forever grateful to my internal supervisor, Dr. Carlo Maj, for his valuable insight, support and encouragement throughout my PhD journey.

I further extend my appreciation to the rest of my thesis committee members: Prof. Dr. Holger Fröhlich, Prof. Dr. Andreas Mayr and Prof. Dr. Regina Betz, for their continuous support and feedback. Your insightful comments helped me improve and grow.

I'm grateful to all my colleagues at IGSB and human genetics department, as well as all my friends, for their support.

Finally, I would like to thank my siblings, for listening to my rants, keeping me motivated and reading my drafts even though they have minimal knowledge of my research, and my parents for their constant love and support.