THE UNIVERSITY *of* EDINBURGH

Genetic architecture of glycomic and lipidomic phenotypes in isolated populations

Arianna Landini

Doctor of Philosophy

The University of Edinburgh

2023

Abstract

Understanding how genetics contributes to the variation of complex traits and diseases is one of the key objectives of current medical studies. To date, a large portion of this genetic variation still needs to be identified, especially considering the contribution of low-frequency and rare variants. Omics data, such as proteomics and metabolomics, are extensively employed in genetic association studies as 'proxies' for traits or diseases of interest. They are regarded as "intermediate" traits: measurable manifestations of more complex phenotypes (e.g., cholesterol levels for cardiovascular diseases), often more strongly associated with genetic variation and having a clearer functional link than the endpoint or disease of interest. Accordingly, the genetics of omics have the potential to offer insights into relevant biological mechanisms and pathways and point to new drug targets or diagnostic biomarkers. The main goal of this thesis is to expand the current knowledge about the genetic architecture of protein glycomics and bile acid lipidomics, two under-studied omic traits, but which are involved in several common diseases.

First, in Chapter 2 I compared genetic regulation of glycosylation of two different proteins, transferrin and immunoglobulin G (IgG). By performing a genome-wide association study (GWAS) of ~2000 European samples, I identified 10 loci significantly associated with transferrin glycosylation, 9 of which were previously not reported as being related with the glycosylation of this protein. Comparing these with IgG glycosylation-associated genes, I noted both protein-specific and shared associations. These shared associations are likely regulated by different causal variants, suggesting that glycosylation of transferrin and IgG is genetically regulated by both shared and protein-specific mechanisms. Next, in Chapter 3 I investigated the effect of rare (MAF<5%) predicted loss-of-function (pLOF) and missense variants on the glycome of transferrin and IgG in ~3000 samples of European ancestry. Using multiple gene-based aggregation tests, I identified 16 significant gene-based associations for transferrin and 32 for IgG glycan traits,

located in 6 genes already known to have a biological link to protein glycosylation but also in 2 genes which have not been previously reported.

Finally, in Chapter 4 I applied a similar approach to bile acid lipidomics, exploring the genetic contribution of both common and rare variants. Despite more than double the sample size (N = ~5000) compared to protein glycomics analysis, I identified only 2 loci, near the *SLCO1B1* and *PRKG1* genes, significantly associated with bile acid traits., for which I noted a sex-specific effect. Further, I found 3 rare variant gene-based associations, in genes not previously reported as associated with bile acid levels. While the biological mechanisms linking these genes to levels of bile acid is not immediately clear, there is evidence in the literature of their involvement in bile acid synthesis and secretion and in liver diseases.

In summary, in my thesis I describe the genetic architecture of the protein glycome and the bile acid lipidome: the former has a higher genetic component, while the latter is largely influenced by environmental factors (e.g., sex, diet, gut flora). Despite the limited sample size, we were able to describe rare variant associations, demonstrating that isolated populations represent a useful strategy to increase statistical power. However, additional statistical power is needed to identify the possible effect of protein glycome and bile acid lipidome on complex disease. A clearer understanding of the genetic architecture of omics traits is crucial to develop informed disease screening tests, to improve disease diagnosis and prognosis, and finally to design innovative and more customised treatment strategies to enhance human health.

Lay summary

Understanding how our genes influence our risk of developing certain diseases is one of the main goals of current medical and scientific research. Many common diseases, such as type 2 diabetes or heart disease, are usually affected by the contributions of multiple genetic variants and genes, in conjunction with influences from our environment, both physical and social (e.g., pollution, diet, socio-economic status, education). Due to this complex interplay, people sharing a similar genetic make-up can have different disease outcomes, meanwhile genetically different people may still have similar disease manifestations. In this intricate scenario, the “signal” coming purely from the genetic contribution to the disease can thus get lost and become faint and hard to detect. As a result, the influence of genetics on complex diseases is not yet fully worked out.

Collections of specific molecules produced or modified by our body are usually called “omics”: for example, the complete set of proteins that can be found in our organism is called “proteome”. Omics are quantitative traits positioned in between genetic variation and complex diseases. Being under stronger genetic influence, it is usually easier to identify which genes or genetic variants influence omics rather than complex diseases themselves. These omics can thus be used as a “proxy” for the more complex diseases: for example, investigating which genes influence the level of cholesterol in blood, rather than trying to understand the genetic causes of a broad range of cardiovascular diseases.

The main goal of this thesis is to expand the current knowledge on how our DNA influences the entire set of two classes of molecules which are known to be involved in several common diseases. To do so, I studied the DNA of people coming from “isolated” populations: groups of people that, due to geographical or cultural reasons, have had limited immigration from other populations. Due to their peculiar history, they often have a unique genetic make-up and I have looked at how it impacts the two classes of molecules. The first class are glycans, which

are sugar structures that can be found attached to the surface of many proteins and are thus collectively called “glycomics”. The second is bile acids, fatty molecules produced by the liver and contributing to food digestion and nutrient absorption, which are part of “lipidomics”. I studied the DNA of thousands of Europeans to understand how it influences, on one hand, the amount of glycans that can be found attached to proteins; and on the other hand, the levels of bile acids in blood. I analysed not only DNA variation that can be easily found in many individuals, but also that present only in a few people among a population (rare variation) and is known to have a disruptive effect on the function of the genes.

Overall, I found glycan levels of transferrin and immunoglobulin G (IgG) proteins to be largely determined by both common and rarer genetic elements. I identified 10 DNA regions that varied between individuals with high and low levels of glycans attached to transferrin protein, 9 of which were not reported previously. Of these regions, some also affect the levels of glycans attached to IgG protein, while others are uniquely linked to one of the two proteins. I also identified rare, high-impact variants affecting the levels of transferrin and/or IgG glycans in both known and not previously reported genes. In contrast, levels of bile acids were less influenced by DNA variation and more likely shaped by environmental factors, such as sex, diet and composition of gut microbiota. I identified only 2 DNA regions varying between individuals with high and low levels of bile acids, which, interestingly, affect levels of bile acids differently in men and in women. Further, I identified rare, high-impact variants affecting the levels of bile acids in 3 genes, which have already been studied in relation to bile acid synthesis and secretion, and to diseases of the liver.

Overall, understanding which regions of DNA make the “omics” profile of each of us unique is crucial to develop more efficient disease screening and treatment strategies and, ultimately, improve people's health.

Declaration of originality

I declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications or collaborative contributions has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

The work presented in Chapter 2 has been published in *Nature Communications* as “Genetic regulation of post-translational modification of two distinct proteins” by **Arianna Landini**, Irena Trbojević-Akmačić, Pau Navarro (Supervisor), Yakov A. Tsepilov, Sodbo Z. Sharapov, Frano Vučković, Ozren Polašek, Caroline Hayward, Tea Petrović, Marija Vilaj, Yurii S. Aulchenko, Gordan Lauc, James F. Wilson (Supervisor) & Lucija Klarić (Supervisor). Details of each author’s contribution are listed in Chapter 2.

The work presented in Chapter 3 has been submitted to a pre-print server as “Exome sequencing reveals aggregates of rare variants in glycosyltransferase and other genes influencing immunoglobulin G and transferrin glycosylation” by **Arianna Landini**, Paul R.H.J. Timmers, Azra Frkatović-Hodžić, Irena Trbojević-Akmačić, Frano Vučković, Tea Pribić, Regeneron Genetics Center, Gannie Tzoneva, Alan R. Shuldiner, Ozren Polašek, Caroline Hayward, Gordan Lauc, James F. Wilson (Supervisor) & Lucija Klarić (Supervisor). Details of each author’s contribution are listed in Chapter 3.

The work presented in Chapter 4 has been submitted to a pre-print server as “Genome-wide association study reveals loci with sex-specific effects on plasma bile acids” by **Arianna Landini**, Dariush Ghasemi-Semeskandeh, Åsa Johansson, Shahzad Ahmad, Gerhard Liebisch, Carsten Gnewuch, Regeneron

Genetics Center, Gannie Tzoneva, Alan R. Shuldiner, Andrew A. Hicks, Peter Pramstaller, Cristian Pattaro, Harry Campbell, Ozren Polašek, Nicola Pirastu, Caroline Hayward, Mohsen Ghanbari, Ulf Gyllensten, Christian Fuchsberger, James F. Wilson (Supervisor) & Lucija Klarić (Supervisor). Details of each author's contribution are listed in Chapter 4.

Signed: Date: 05/01/2023

Acknowledgements

I would like to express my sincere gratitude to my supervisors Prof. Jim Wilson, Dr. Lucija Klarić and Dr. Pau Navarro. Not only you are excellent scientists, but most importantly you are wonderful people! You have always been there to support me, listen to me and cheer me up, believing in me when not even I could. I know I often sounded gloomy and complained about everything (especially Edinburgh's weather and queuing), but you made this PhD journey Type II fun - it was quite hard at times, but, looking back, definitely an experience to treasure. Thank you for everything.

I am deeply grateful to IMforFUTURE for funding my PhD, and especially for introducing me to a group of amazing scientists who have become close friends - Anna, Annah, Azra, Iva, Jay, Jaquie, Maarten, Samira, Shafiq, and Tamás. I will never forget the memorable experiences we shared, such as rafting for the first time, learning how to say "put down the knife!" in Croatian (disclaimer: no PhD student was harmed during this process), playing croquet, travelling, eating, and laughing together. Thank you all for being a part of this journey.

I would like to thank all of my fellow (past and present) PhD students in the Wilson Group for your support, assistance, memes and for adding also some Type I fun to the PhD experience! I would like to thank also Dr. Peter Joshi and Dr. Nicola Pirastu, both former senior members of the Wilson Group. One taught me how to adhere to the rules, while the other showed me how to creatively work around them. Thank you for your guidance and mentorship.

Finally, the biggest of the hugs to my incredibly supportive and caring partner Alessandro. Despite you still think, after nearly 10 years together, that enzymes are some sort of magic spell, your constant encouragement was essential in the completion of this thesis. Thank you for your unwavering support.

Table of Contents

Figures

Appendix

Chapter 1: Introduction

1.1 Why study the genetic architecture of human traits and diseases?

Understanding how genetics contributes to variation of phenotypes is one of the most essential and yet longest-standing questions in genetics. Extensive knowledge of genetics is fundamental for understanding how cells, organisms, populations and species live, evolve and die. Further, improving our understanding of how genetic variants contribute to and whether they cause human diseases is one of the main goals of modern medical genetics. In the context of human population studies, the characterisation of all genetic variation contributing to the heritable phenotypic variability of a given trait or disease describes its "genetic architecture". In particular, the genetic architecture of a complex trait depends both on the number of genetic variants influencing the phenotype, but also the magnitude of their effects on the phenotype, their frequency in the studied population, and the interactions between each other and/or with the environment[1]. All these elements contribute to the relationship between genotype and phenotype and provide a more complete picture of the variation of a complex trait or disease compared to the assessment of heritability alone. Here, heritability is defined as the estimated proportion of variance in a phenotypic trait that is due to additive genetic factors. Complex traits can in fact have similar heritability estimates but widely different genetic architecture. Height is an example of a highly heritable trait (estimates ranging from ~50% to ~90%)[2–5], with a polygenic architecture - it is influenced by a large number of variants scattered across multiple different genes across the genome. By contrast, phenylketonuria, a Mendelian autosomal recessive disorder, is considered a monogenic disease - mutations at a single locus, the *PAH* locus, are responsible for the impaired activity of the phenylalanine hydroxylase enzyme and the accompanying hyperphenylalaninemia phenotypes[6]. With a global disease prevalence of 0.004%[7], phenylketonuria mutations are rather rare but some of them are responsible, alone, for severe hyperphenylalaninaemia.

In summary, genetic architecture can strikingly vary from one complex trait or disease to the other. This should be taken into consideration when designing a study: traits more similar to height will require large population-based collections and genome-wide common variant data, while traits such as phenylketonuria will either need sequence data or family-based approaches to detect associated variants. Better understanding of the genetic architecture of complex traits and diseases is crucial to develop informed disease screening tests, improve disease diagnosis and prognosis, and finally design innovative and more customised treatment strategies to enhance human health[8].

1.2 Complex traits result from the contributions and interactions of multiple genetic and environmental risk factors

Complex traits result from the contributions and interactions of multiple genetic variants and environmental risk factors. Due to this relationship between multiple genetic contributors and environment, different genotypes can lead to the same phenotype, and conversely the same genotype can give rise to different phenotypic manifestations based on the effects of the environment. Clear relationships between genotype and phenotype have thus been uncovered for a few complex traits[8]. Due to the many "actors" involved and their complex interactions, genetic architecture, that is the number and attributes of genetic variants contributing to a phenotype, varies from trait to trait[9].

To exemplify this concept, the number and characteristics of associated genetic variants can be compared between diseases or other complex traits. For example, type 1 and type 2 diabetes mellitus, despite having a similar phenotype (hyperglycaemia), show a clearly different genetic architecture (Figure 1, panel a). Type 1 diabetes, an autoimmune disease causing the destruction of insulin-producing β-cells in the pancreas, is influenced, together with common variants, also by a relevant portion of rare/low-frequency genetic variants[10]. While the former generally, but not always, have a smaller effect on disease susceptibility, the latter commonly have a larger impact. On the other hand, the genetic architecture of type 2 diabetes has been found to be predominantly represented by multiple common genetic variants with small effect on disease susceptibility, with just a handful of low-frequency or rare variants being overall reported[11–13], even considering the findings of recent large-scale, multi-ancestry studies[14–17]. Moving from complex diseases to complex traits of biomedical interest, Vitamin D levels, measured as 25-hydroxyvitamin D, have, despite the large sample size tested[18], been found associated with a limited number of genetic variants, some of which, especially the rarer ones, have a large effect size[19] (Figure 1, panel b). In contrast, low-density lipoprotein (LDL) cholesterol level appears to be regulated by a larger number of genetic loci, showing a wide range of effect

sizes[20,21] (Figure 1, panel b). Despite the fundamentally different genetic architectures, the two biochemical traits have a similar heritability (~50%)[22,23].

Overall, the genetic architecture of most complex traits is characterised by many variants of varying effect size and allele frequency, similar to type 2 diabetes and LDL cholesterol (Figure 1). Accordingly, a mixed discovery strategy aimed at identifying both common and rare variants, with both small and large effect sizes, is likely needed to elucidate the genetic architecture of these traits[24].

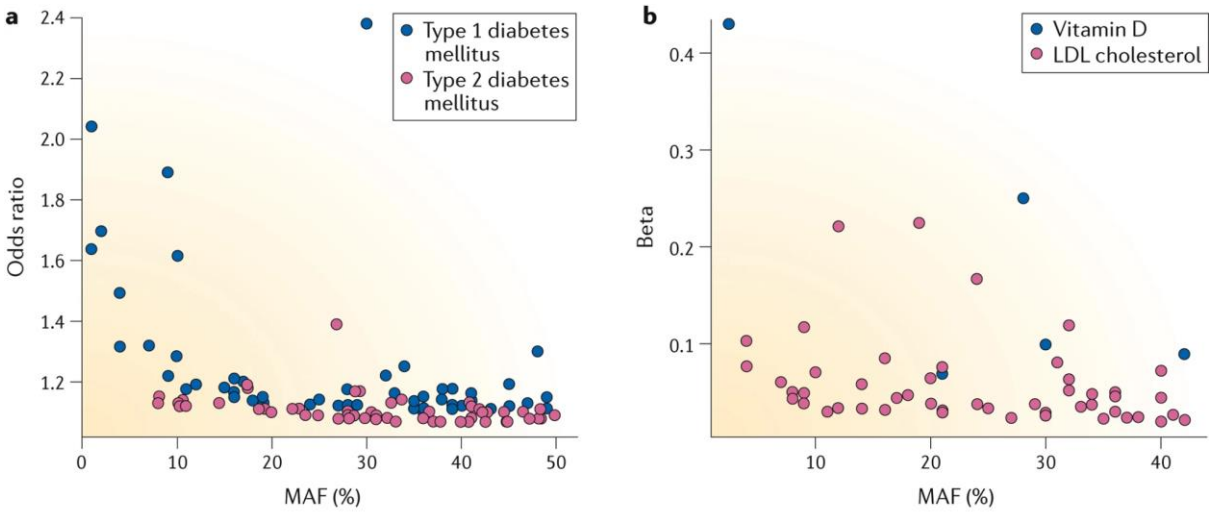


Figure 1. Examples of different genetic architectures of complex diseases and biomedical traits. a) Minor allele frequency (MAF) and effect size (measured by odds ratio) of genome-wide significant SNPs for type 1 and type 2 diabetes mellitus. Odds ratios give the odds of individuals having a phenotype (outcome) associated with a specific risk allele (exposure), compared with the odds of the same phenotype for individuals who do not have that same risk allele. Data used to generate this plot panel was taken respectively from Bradfield *et al.* (2011)[25](9,934 cases and 16,956 controls), and DIAGRAM (2012)[26](34,840 cases and 114,981 controls). b) MAF and effect size (represented by beta) of genome-wide significant SNPs for vitamin D (25-hydroxyvitamin D) and low-density lipoprotein (LDL) cholesterol. Beta quantifies in standard deviations the altering effect a reference allele has on the phenotype of interest. Data used to generate this figure was taken respectively from Manousaki *et al.* (2017)[27] and Willer et al. (2013)[20]. The figure was reproduced from Timpson *et al.* (2018)[9] with permission.

1.3 Analysis strategies to investigate the genetic architecture of complex traits

Different types of genetic architecture should be assessed by tailored study designs, as each genetic data type and analysis strategy has its own merits and disadvantages, with the only common ground of all being subject to the limitations of sample size. Genetic data include genotypes and sequence data - e.g. whole exome or genome - and analytical methods include single variant associations tests, such as genome-wide association studies (GWAS), or variant aggregating tests, such as burden tests.

GWAS tests for a difference in phenotype between individuals with different genotypes at a particular genetic variation. The most commonly used genetic variants in GWAS are single-nucleotide polymorphisms (SNPs), whose genotype is usually obtained using microarrays. Genotypes that have not been assayed directly can be statistically estimated using imputation. By comparing the haplotypes of individuals subjected to genome-wide genotyping to sequenced haplotypes, or haplotypes derived from denser genotyping arrays, observed in a reference panel, it is possible to impute missing genetic variation at millions of additional genomic sites[28]. Thanks to imputation, adequately powered GWAS are able to identify the genetic contribution of variants as rare as 0.1% MAF in samples of European ancestry[29]. Accordingly, GWAS have successfully identified a large number of significantly associated loci across numerous complex traits and diseases[30]. Nevertheless, a large portion of the genetic contribution to complex traits still remains unexplained, despite the increase in sample size and thus statistical power[31]. For example, a large GWAS of type 2 diabetes[32] identified 143 risk variants at genome-wide significance, of which only 4 are rare (0.01% ≤ MAF < 1%), in > 650k individuals of European ancestry (62,892 cases and 596,424 controls). All tested variants however explain only a small portion of the disease variability (SNP-based h^2 = ~20%), while the contribution of rare variants to type 2 diabetes still need to be fully assessed, even in larger, more recent, and multi-ancestry studies[14–17]. On the other hand, with the steady growth of GWAS sample sizes, the newly identified common variants generally have smaller

effects on risk. By continuously increasing statistical power, GWAS association signals for several complex traits thus tend to spread broadly and very densely across the genome, in what has been described as “omnigenic model”, including near many genes without an obvious connection to the trait of interest[33]. Even if the genetic make-up of some complex traits and diseases is currently far from being exhaustively explained, continuously increasing GWAS power would therefore paradoxically pose a challenge in identifying key biological pathways and causal mechanisms behind the trait of interest.

At the current statistical power, GWAS generally fails to detect the contribution not only of common variants with very small effects, but also of rare and low-frequency variants (MAF < 1%), which, by contrast, have been hypothesised to have larger effect sizes and, especially in the case of coding variants from exome sequencing, to possibly implicate “core” genes having a clear functional link to the biology of the trait of interest[34]. The impact of rare variants on a range of human phenotypes is well established, with many Mendelian disorders and rare forms of common diseases being attributed to individual, highly penetrant alleles[35]. The role that rare variation plays in common diseases and complex traits has started to be investigated fairly recently, allowing the identification of rare variants affecting trait variation or disease risk, evaluation of the relative impact of individual genes to overall phenotype variability and to further the understanding of trait genetic architecture[13,36–38]. To allow genotyping of rare variants, next-generation whole-genome (WGS) and whole-exome sequencing (WES) technologies are employed. High-throughput parallel-sequencing approaches generate billions of short sequence reads, which are then aligned to a reference genome to identify genetic sites where sequenced individuals vary. Despite sequencing providing an unprecedented opportunity to investigate the roles of low-frequency and rare variants in complex diseases, identification of these variants in sequencing-based association studies presents considerable challenges. The statistical power of classical single-variant-based association tests applied to rare variants is in fact remarkably low - the power decreases as the allele frequency decreases - unless analysing very large sample sizes or

detecting variants with large effect size[31]. With the exception of few biobank resources, such as UK Biobank[39], cohorts with sequencing data available are not usually large, due to the high cost of sequencing data[40], further decreasing power. To address these power issues, statistical methods specifically tailored for rare variant association analysis are used. Instead of testing each variant individually, as commonly done in GWAS, rare variant methods increase power by aggregating association signals across multiple rare variants included in a biologically relevant region, such as a gene. For this reason, these methods are called gene-based aggregation tests of multiple variants. Further, variants tested are chosen not only based on their MAF but usually also on their impact on amino acid sequence, predicted functional roles and deleteriousness. Burden tests represent one class of aggregation tests. They collapse information for multiple genetic variants into a single genetic score and test for association between this score, capturing the cumulative effects of rare variants in the region, and the trait of interest[41,42]. Since burden tests assume that rare variants in the region are all causal and affect the phenotype in the same direction with similar magnitudes, they suffer from a substantial loss of power if these assumptions are violated[43,44]. Variance-Component Tests, such as the sequence kernel association test (SKAT)[45], represent an alternative to burden tests. Instead of aggregating variants, these methods test for association by evaluating the distribution of the aggregated score test statistics of individual variants. SKAT performs best when the tested genetic region contains both protective and deleterious variants or many non-causal variants, and is computationally efficient for performing genome-wide sequencing association studies in large samples. Overall, the performance of gene-based aggregation tests depends on the underlying genetic architecture of the trait analysed. For regions with a large number of causal variants with the same direction of association, burden tests are likely to be more powerful. Conversely, if both risk-increasing and risk-decreasing variants are expected in a region or if the majority of variants are non-causal, variance-component tests should be the method of choice[31]. Since trait architecture is usually unknown, omnibus tests such as SKAT-O[46], ACAT-O[47] and SMMAT-E[48], combining the strength of burden and variance-component tests, have been

developed. Finally, aggregation tests of multiple variants have been extended to take into account population structure and cryptic relatedness[48].

1.4 Why use genetically isolated populations?

Describing the genetic architecture of a trait is inevitably bound to the limits of statistical power. Power required for detecting associations between a trait or disease and SNP markers is affected by sample size, allele frequency and effect size of the causal genetic variants, linkage disequilibrium with genotyped markers and inheritance models (e.g., additive, dominant, and multiplicative models)[49–51]. Together with choosing the most suitable type of data and method to better describe the genetic architecture of the trait of interest, employing genetically isolated populations can as well empower association studies.

“Population isolates” refer to population groups which derive from a small number of founding individuals and have been separated for many generations from surrounding populations. The geographical and/or cultural isolation of these populations has several genetic consequences, such as higher homogeneity than the general population and thus a reduced effective population size (i.e. the effective number of individuals required to explain the observed genetic variability)[52]. Due to the combined effect of endogamy, bottlenecks, genetic drift and selection, certain alleles at a particular locus can reach fixation or extinction in isolated populations, thus reducing the amount of genetic variability[53,54]. As a result, some variants that contribute to complex traits or diseases may be rare in the general population but have risen to higher frequency in the isolate. For example, it has been reported that about one tenth of whole genome sequencing variants from the VIKING cohort, a collection of samples from the isolated northern Scottish Shetland islands, are unique to the isolate or are seen at frequencies at least ten-fold higher than in cosmopolitan populations[55]. In addition, LD tends to extend over longer distances in isolates than in the general population[56–58], thus generating longer haplotypes which facilitate disease association studies and empower imputation approaches[59]. Reduced allelic variability combined with extended LD, can thus boost statistical power for trait association at low-frequency and rare variants in isolated populations, compared with non-isolated populations with wider allelic diversity.

Not only allele frequency, but, as a consequence, also the prevalence of diseases can be influenced by isolation. Each isolate shows a unique profile of rare disease alleles, which may be coupled with a higher prevalence of some diseases and lower incidence of others[60,61]. For example, the Pima Indians of Arizona are characterised by a very high prevalence of type 2 diabetes (~20%)[62,63], but report nearly no case of type 1 diabetes[64]. Finally, in addition to reduced genetic complexity, individuals from isolates tend to be also environmentally homogeneous, by sharing common lifestyle and cultural habits and by being exposed to similar environmental conditions. This represents another potentially advantageous property of population isolates, since the reduced noise coming from environmental confounding factors increase statistical power. While employing genetic isolated populations offers several advantages as discussed earlier, it is important to acknowledge that this approach is not without limitations or potential drawbacks. The presence of longer haplotypes, facilitating the discovery of genetic associations, is in fact a double-edge sword. On one hand, longer LD stretches can in fact increase the power to detect genetic associations, but on the other hand, they can reduce the resolution, making it more challenging to identify the specific causal variants within a broad association peak[65].

Another factor that must be considered when studying isolated populations is the complex population structure, which can cause spurious associations if not properly accounted for. A linear mixed model (LMM) is a statistical method commonly used in GWAS to account for the correlation structure among individuals and to correct for population stratification and familial relatedness. The LMM takes into account both fixed effects and random effects. Fixed effects represent systematic effects assumed to be constant across the entire population, such as the genetic variant tested and other covariates as age, sex, or environmental factors. Random effects account instead for the variation in the genetic background of individuals and the relatedness among them, capturing the effects that are specific to individual subjects and are not constant across the population. Including random effects in the LMM in the form of a kinship matrix, which captures the correlation between individuals due to shared genetic

background, helps to control for population stratification and familial relatedness. The LMM thus estimates the effect size of each genetic variant on the phenotype of interest while adjusting for both fixed and random effects[65].

In conclusion, genetically isolated populations possess several unique features, such as genetic, environmental and phenotypic homogeneity, higher frequency of otherwise rare alleles and higher disease prevalence. Despite the smaller sample size, these features can enhance the power for locus identification in genetic association studies of complex traits, especially for low-frequency and rare variants, compared to the general population. On the other hand, the study of genetic isolated populations requires statistical methods able to correct for the high levels of kinship and GWAS downstream analyses to pint-point the plausible association causal variants.

1.5 Genetic studies of complex, multifactorial diseases have so far given limited answers

As previously mentioned, complex diseases are driven by a combination of environmental and genetic factors. Due to their high prevalence, complex diseases represent a substantial burden for the public health systems[66], which will likely further increase due to the ageing population. For nearly two decades GWAS has been one of the tools of choice to identify genetic risk loci implicated in complex diseases[67]. Despite the thousands of genomic loci which have been significantly associated with human diseases so far, our understanding about the biological function of the identified variants is still limited: this is mainly due to the fact that GWAS do not necessarily pinpoint causal variants and genes.

First of all, while local correlation of multiple genetic variants due to linkage disequilibrium facilitates the identification of an associated locus, it also complicates discerning the causal variant(s)[68]. In addition, most of the association signals map to non-coding regions of the genome, for which biological interpretation in the context of the analysed disease is intrinsically challenging[69,70]. Also GWAS where too many hits are involved represents a challenge from the causal interpretation point of view. In a recent GWAS[5] (one of the largest studies so far, counting 5.4 million individuals), 12,111 independent SNPs were identified as significantly associated with height, accounting for nearly all of the common SNP-based heritability in populations of European ancestry. These SNPs are clustered within 7,209 non-overlapping loci, covering a large portion of the total human genome, about 21%. Accordingly, authors argue that, at this point, adding more data will not reveal more (common) variants in European populations. While this work obviously marks a milestone in our understanding of the contribution of genetics to complex traits, it also exemplifies how many complex traits are driven by enormously large numbers of variants of small effects and genes not directly related to the phenotype[33].

Overall, identifying and characterising “core” causal variants and genes, and their interacting pathways, in the context of the molecular pathophysiology of diseases remains a difficult task. Why do such a high number of genes outside of the pathways regarded as key to the trait of interest appear in GWAS, as the sample size increases?

In the human body, multiple biochemical networks and physiological mechanisms interact with one another in a complex coordinated fashion. These systems, which are typically under genetic control, often respond to environmental stimuli via feedback mechanisms and are packed with redundancies and compensatory mechanisms. These intricate systems can be visualised as hierarchical structures[71], with genes at the base and clinical endpoints defining the disease at the top (Figure 2). There are thus a multitude of contributing factors, positioned at different levels of these hierarchies and ultimately influenced by genetics, which may affect a certain clinical endpoint interpreted as a symptom or sign of disease. Further, different underlying pathologies, and thus different genetic causes, can lead to similar phenotypic endpoints. In such hierarchically structured physiologic systems, the effect of genes on a disease endpoint is mediated by a multitude of intermediary phenotypes. Genetic contribution gradually attenuates, becoming thus more difficult to identify and dissect, moving from one hierarchy level to a higher one. This description of the genetic contribution to complex, multifactorial disease fits well with the previously mentioned omnigenic model[33], suggesting that all genes expressed in disease-relevant cells are likely responsible to affect the functions of core disease-related genes, and ultimately, the clinical manifestation of the disease.

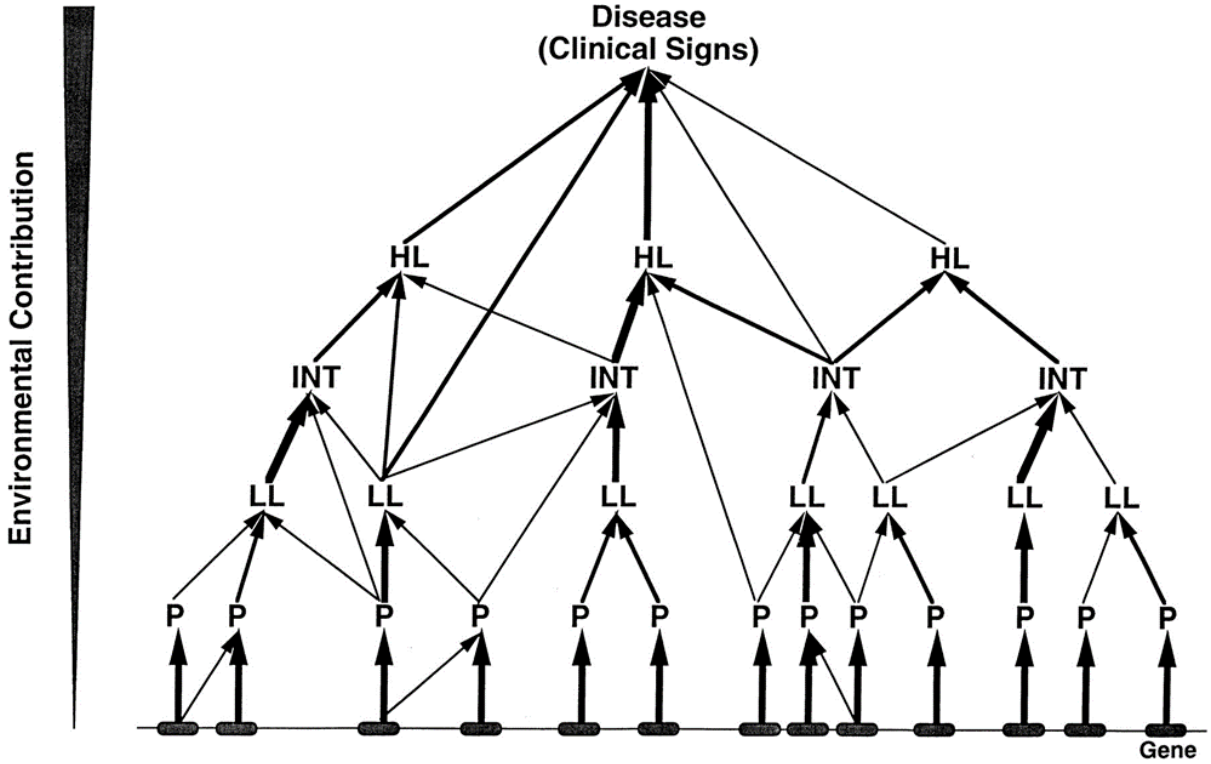


Figure 2. Schematic diagram of the hierarchical relationship linking genes to clinical manifestations of a disease. Effect of genes and their products (P) on the disease is mediated by several "intermediary" phenotypes - here indicated as low-level (LL), intermediate (INT) and high-level (HL) phenotypes - in a hierarchical fashion. The thickness of the arrows denotes the strength of the contribution of a lower-level factor to a higher-level factor. Accordingly, the genetic contribution grows more faint and subtle moving from the bottom to the top of the scheme. Conversely, the effect of environmental conditions on different levels of the hierarchy likely diminishes going from the top to the bottom of the scheme, as represented by the inverted triangle on the left-hand side of the figure. This figure was reproduced from Schork, 1997[72]. (Reprinted with permission of the American Thoracic Society. Copyright © 2022 American Thoracic Society. All rights reserved. Cite: Schork, N. J/1997/Genetics of complex disease: Approaches, problems, and solutions/Am. J. Respir. Care Med./156(4)/S103–S109. The American Journal of Respiratory and Critical Care Medicine is an official journal of the American Thoracic Society).

1.6 Employing omics techniques to reveal the molecular underpinnings of complex, multifactorial diseases

Due to the hierarchical structure characterising most physiologic systems in our bodies, linking-up a complex disease with the “core” genes that ultimately influence is a challenging task. Most importantly, associated loci detected will generally have very small effect sizes and an unclear biological role in the context of the disease of interest.

A strategy that could boost statistical power and facilitate the biological interpretation of identified genes is to switch the focus from complex diseases to intermediate phenotypes. Intermediate phenotypes are typically more proximally related to the genetic substrate than high-level, disease end-points: thanks to their position in the hierarchical structure of physiologic systems, they are in fact influenced by a smaller number of genes. Intermediate traits can thus be employed in genetic association studies as “proxies” for complex diseases of interest: they can be considered as measurable manifestations of more complex phenotypes, usually more directly linked to the underlying biology and more strongly associated with genetic variants than the complex trait or disease itself. For example, quantitative traits such as cardiac hypertrophy or cholesterol level were used as intermediate phenotypes of complex cardiovascular diseases, such as cardiac insufficiency or stroke[73]. In type 2 diabetes, insulin receptor resistance has been employed as intermediate phenotype between several susceptibility genes and clinical diabetes[74,75].

Intermediate phenotypes and diseases differ not only for their proximity to the genetic substrate, but also in the statistical models commonly used for their GWAS analysis. Linear regression is used to test associations between genetic variants and quantitative traits, such as intermediate phenotypes, which are represented by numerical values and can vary along a continuum, typically following a normal distribution. Logistic regression is instead suitable for examining the relationship between genetic variants and dichotomous outcomes,

such as disease status (affected/unaffected). The outcome of linear regression is the estimation of regression coefficients quantifying, in both direction and magnitude, the change in the mean value of the quantitative trait for each additional copy of the effect allele, assuming an additive genetic model. While the outcome of logistic regression represents the logarithm of the odds ratios (ORs), quantifying the change in odds, or likelihood, to be disease affected for each additional copy of the effect allele. Overall, quantitative traits, as intermediate phenotypes, tend to provide better statistical power to detect a genetic effect. Analysing continuous traits by linear regression gives access to a larger range of values, which increases the variability in the data and can thus improve the possibility of detecting a genetic effect, if one exists. On the other hand, studying disease end-points through case-control association studies and logistic regression necessarily use observations on the less informative observed disease scale, being these dichotomous rather than continuous[76].

Omics, representing a set of specific biological molecules that can be found in the human organism (e.g. proteomics, metabolomics, etc.), are an excellent example of intermediate phenotypes and are widely used to identify causal genes underlying common diseases[77–82]. Thanks to their intermediate position, mediating between the genetic substrate and disease endpoints, omics are more amenable to genetic mapping than complex, multifactorial diseases[83], and, at the same time, are closer to the underlying biological mechanism.

Once genomic variants have been statistically associated with a disease, omics technologies can be employed to identify the underlying molecular mechanisms. For example, GWAS of type 2 diabetes and BMI successfully identified a robust association between *FTO* gene and obesity[84,85]. Epigenomics and transcriptomics data were then employed to suggest a potential mechanism for the genetic association between *FTO* and obesity. A variant in *FTO* causes the activation of downstream targets *IRX3* and *IRX5* during early adipocyte differentiation, which results in a shift toward obesity phenotypes (e.g. lipid storage, increased fat store and body-weight gain)[86].

Omics data can be also used to categorise patients into subtypes or along a spectrum of a diseases, based on their specific molecular signatures, providing thus a more nuanced approach to patient stratification, beyond the classic binary classification of healthy and diseased. This expanded stratification can be hugely beneficial for disease diagnosis and treatment[87]. Additionally, integrating polygenic risk scores (PRS) into omics studies can enhance our understanding of disease complexity and individual risk profiles. By capturing the cumulative effect of variants across all genome on the trait or disease of interest, PRS can aid in the identification of patient subtypes with distinct molecular profiles and varying disease susceptibilities[88].

Finally, using omics as proxies of complex diseases can also provide statistical advantages in genetic association studies. It is known that inconsistency in disease diagnosis can introduce phenotyping errors reducing discovery potential for genetic associations[89–91]. And in the cases of complex, multifactorial diseases it can be difficult to determine how to measure disease outcomes, often leading to heterogenous phenotypes[89,92,93]. For example, high misdiagnosis rates have been estimated for Alzheimer's disease[94], bipolar disorder[95], migraine, fibromyalgia and psychogenic disorder[96], due to overlap of symptoms with other diseases and/or mistakes in application of diagnostic criteria. Once again, omics can be employed in lieu of complex disease to avoid the reduction in statistical power due to heterogeneity in phenotype definition. However, omics are of course not immune from measurement errors. They should undergo thorough pre-processing and quality control procedures in order to provide robust results[97].

In conclusion, intermediate phenotypes are quantitative traits positioned in between genetic variation and complex diseases. Employing omics as intermediate phenotypes have allowed identification of disease-associated variants and elucidation of molecular mechanisms behind complex diseases. They are successfully used for biomarker/drug discovery, patient stratification and disease classification. However, except rare cases, no single omic is able to

capture the whole complexity of molecular events leading to human disease. Different omics should be thus combined to create a more comprehensive picture of the mechanisms underlying human phenotypes and diseases[98].

1.7 Pleiotropy in complex traits and methods to dissect it

Given that numerous complex traits and diseases are influenced by a large number of genes and genetic variants, the phenomenon of pleiotropy is frequently observed. Pleiotropy occurs when a genetic locus has an effect on multiple phenotypes, suggesting that these traits may thus share common underlying biology. Identifying the pleiotropic effects of specific genes on complex traits and diseases may be useful for understanding their underlying causes, and provides insights into the biological mechanisms underlying the traits and their potential comorbidities[99]. It is not enough for a genetic locus to be found associated with more than one trait for it to be pleiotropic: the underlying cause for the observed pleiotropic behaviour is in fact key and many alternative models for an apparent "pleiotropic" effect can fit the observed data (Figure 3). Biological pleiotropy refers to a genetic locus having a direct biological influence on more than one phenotypic trait (Figures 3a, 3b and 3c). It can occur both at the allelic level, where a single causal variant contributes to multiple phenotypes (Figure 3a); or at the gene level, where various distinct variants in the same region are associated with different traits (Figure 3b and 3c). Mediated pleiotropy occurs when a variant directly affects one trait, which in turn affects another (Figure 3d), so that the genetic association to a phenotype is mediated by another phenotype. Finally, spurious pleiotropy includes various sources of bias causing a genetic locus to falsely appear as associated with multiple phenotypes (Figures 3e and 3f).

Since "correlation is not causation", Mendelian randomisation is the method of choice in genetic epidemiology[100] to identify cases of mediated pleiotropy, where the association between a genetic variant and a phenotype is mediated by another trait. Similar to a randomised controlled trial, Mendelian randomisation (MR) uses genetic variation, attributed to the individual randomly at conception, as a natural experiment to divide the population into subgroups and investigate the causal effect of a modifiable risk factor, also referred to as "exposure", on a health outcome in observational data[101]. MR thus uses genetic variants as

proxies to obtain valid causal inferences for the effect of the exposure on the outcome. Therefore, choosing valid genetic instrumental variables (IV) is crucial for a successful MR study. For a genetic variant to be a valid IV for causal inference in an MR test, three key assumptions must be met: (1) the genetic variant must be directly associated with the exposure; (2) the genetic variant must not be related to confounding factors obscuring the connection between the exposure and the outcome; and (3) the genetic variant affects the outcome only through the exposure[102,103]. Colocalisation analysis is another method that has been developed to distinguish different pleiotropy scenarios. In particular, it allows us to assess whether two traits are regulated by the same underlying causal variant or by two distinct causal variants, possibly in linkage disequilibrium, in the same gene or region[104]. As good practice, colocalisation is often used jointly with Mendelian randomisation to assess the validity of instrumental variables for a given genetic region. For example, a genetic instrument of the exposure could be in linkage disequilibrium with another variant that independently influences the outcome, either directly or via an alternative exposure. This would represent a violation of one of the three assumptions of Mendelian randomisation. If there is strong evidence that the association signals of exposure and outcome do not colocalise, meaning that the two traits are influenced by distinct causal variants, then it is implausible that variants in that region represent valid instrumental variables for the exposure[105]. Other than validation for MR, colocalisation methods are widely used to assess whether disease endpoints and potential biological mediators might share one or more causal variants[104,106], in order to favour mechanistic interpretation of disease endpoints and associated genetic variants identified by GWAS.

In conclusion, Mendelian Randomisation uses genetic variants associated with modifiable traits (exposures) to identify causal associations with diseases (outcomes). Colocalisation instead is used to discern between two possible underlying situations at a genetic region: distinct causal variants, possibly in linkage disequilibrium, or a single shared casual signal influencing two traits.

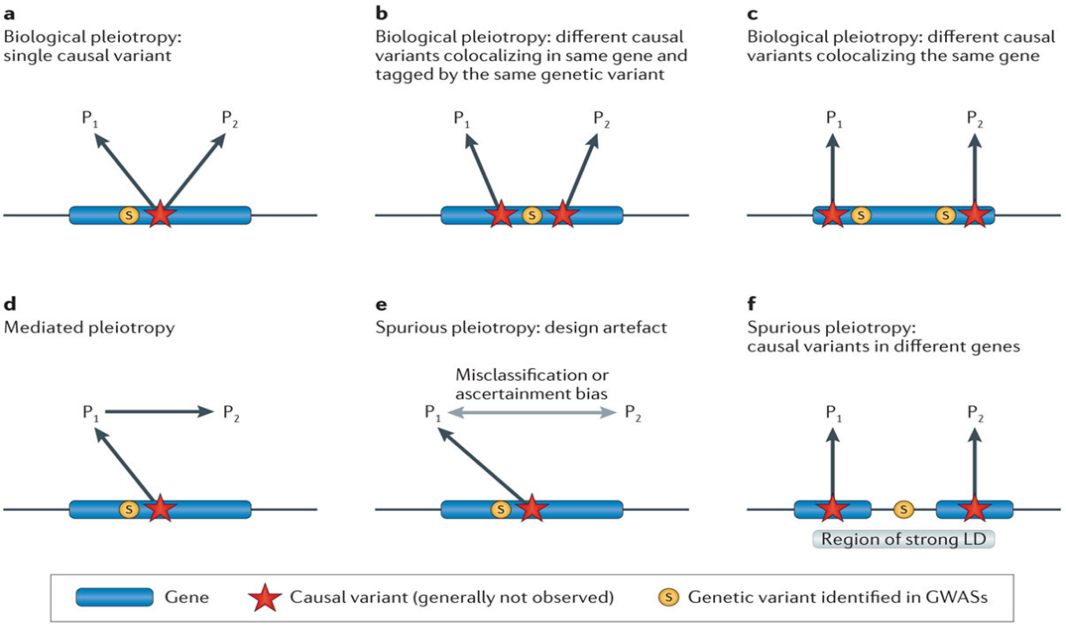


Figure 3. Different types of underlying pleiotropic structures. In each scenario, the observed genetic variant (S) is found associated with phenotypes 1 and 2 (P1 and P2) by GWAS. The observed variant S is in linkage disequilibrium (LD) with a causal, generally unobserved, variant (red star) that affects one or more phenotypes. In some cases, the causal variant may be identified directly by GWAS and the figures can be simplified accordingly. a) Biological pleiotropy at the allelic level, with a single causal variant affecting both phenotypes. b) Biological pleiotropy with colocalising association: the observed genetic variant is in strong LD with two causal variants in the same gene, each one affecting a different phenotype. c) Biological pleiotropy at the genic level, with two independent causal variants in the same gene affecting different phenotypes. d) Mediated pleiotropy, where the causal variant affects P2 through P1: P1 lies on the causal pathway to P2, and thus the association occurs between the observed variant and both phenotypes. e) Spurious pleiotropy: the causal variant affects only P1, but P2 has been linked to P1 due to misclassification or sampling bias, causing a spurious association between the observed variant and the P2. f) Spurious pleiotropy, where the observed variant is in LD with two causal variants in different genes, each one affecting a different phenotype. The figure was reproduced from Solovieff *et al.,* 2013 with permission.

1.8 Thesis aim: investigating the genetic architecture of understudied molecular traits with known involvement in health

The aim of this thesis is to investigate the genetic architecture of two classes of omics, known for being involved in a great number of complex diseases, but that have thus far received limited attention by the scientific community: protein glycomics and bile acid lipidomics. While the biochemical pathways involved in the synthesis of these two omic classes are well characterised, there is currently an insufficient understanding of their genetic regulation. In this thesis therefore I have taken advantage of different genetic data available, namely imputed genotypes and whole exome sequencing data, and different statistical methods for identification of genetic associations, such as GWAS and gene-based aggregation tests, to investigate both common and rare variants contributing to variation of protein glycomics and bile acid lipidomics. Especially for rare and low frequency variants, I leveraged the increased statistical power provided by genetically isolated populations. Using statistical methods designed for identifying causation and pleiotropy, such as Mendelian randomisation and colocalisation analysis, I then assessed the potential impact of protein glycome- and bile acid lipidome-associated variants on health-related traits and diseases, to gain insights of the underlying molecular mechanisms.

In the following paragraphs I briefly describe the structure and function of the omics studied in this thesis - protein glycome and bile acid lipidome - highlighting their known involvement in diseases and what has been reported so far regarding their genetic architecture. I then describe in more detail the work that has been carried out in each results chapter of this thesis. In brief, the results chapters reflect three underlying aims of this thesis - 1) understanding the differential regulation of glycosylation of two different proteins, 2) understanding the contribution of low frequency and rare variants to protein glycosylation and 3) understanding the genetic architecture of bile acid lipidome. In the concluding discussion chapters I compare similarities and differences in genetic regulation of two different omics intermediate phenotypes.

1.8.1 Protein glycome

N-glycosylation is one of the most common protein post-translational modifications (PTMs), where carbohydrate structures called glycans are covalently attached to an asparagine (Asn or N) residue of a polypeptide backbone. Glycosylated proteins, referred to as glycoproteins, are known for performing a plethora of different relevant functions in the cell, with mainly structural, modulatory and signalling properties[107]. Protein glycosylation can be evaluated at three distinct levels: released glycans, glycopeptides, or whole glycoproteins. The most common high-throughput approach involves analysing released glycans, which are obtained by cleaving glycan structures from the protein carrier. This method allows for detailed structural characterisation of glycan species and is largely independent of the original glycoprotein source[108]. Although adaptable to various glycoproteins, current high-throughput analysis of glycopeptides is primarily used to examine immunoglobulin glycosylation. Due to the increased quantity and complexity of analytes in comparison to released glycans, this approach requires analytical techniques with high sensitivity and resolution. Finally, analysis of intact glycoprotein is emerging in the field of high-throughput glycomics, enabling the characterisation of the proteoform distribution of an isolated protein; however, it faces limitations in sensitivity and glycoform resolution[108]. Several high-throughput approaches for the analysis of released glycans are available, hydrophilic interaction liquid chromatography with fluorescence detection using ultrahigh-performance liquid chromatography (HILIC-UHPLC-FLD) being the most used[109]. Accordingly, prior to glycome measurement, the glycoprotein of interest must be purified and attached glycans must be cleaved of their protein carrier. Large-scale studies of individual glycoprotein glycomes still face challenges due to the absence for several glycoproteins of affordable, fast, and efficient purification methods in a large samples[110]. HILIC-UHPLC-FLD is then used to separates glycans based on their size, structure and charge. This technology allows for complete characterisation of complex glycan mixtures in a relatively short time and is the method of choice

for routine analysis of protein glycosylation with previously characterised glycan structures[110]. Integration of the resulting chromatogram peaks can be performed manually, which represent one of the most time-consuming tasks in analysis of large number of individuals, or using an automatic semi-supervised method[111]. For annotation of novel glycan structures, UHPLC may be coupled with other methods that can provide further structural information, especially mass spectrometry (MS) with Liquid Chromatography with tandem mass spectrometry (LC-MS) techniques. This synergetic approach can be used to annotate glycan structures in samples that are representative of the larger sample set[110].

Thanks to improvements of available high-throughput analytical methods, glycomics is an emerging omics field, studying the whole collection of glycans in biological systems[112]. N-glycans are built from only ten monosaccharides (i.e. fucose, galactose, glucose, N-acetylgalactosamine, N-acetylglucosamine, glucuronic acid, iduronic acid, mannose, sialic acid and xylose) but are characterised by vast structural diversity and high complexity. N-glycan structures result in fact from a sophisticated interplay of glycosyltransferases, glycosidases, transporters, transcription factors and other proteins[113], with a number of epigenetic and environmental factors contributing to N-glycome composition as well[114,115]. Accordingly, the glycome is expected to be several orders of magnitude more complex than the proteome itself[116]. Protein N-glycosylation is involved in a multitude of biological processes, including receptor interaction, immune response, protein secretion and transport[117].,Changes in N-glycosylation patterns have been associated with the ageing process[118] and a wide range of diseases, including Parkinson's disease[119], low back pain[120], rheumatoid arthritis[121], ulcerative colitis[122], Crohn's disease[122], type 2 diabetes[123] and cancer[124–126]. In addition, N-glycans are considered as potential therapeutic targets[127] and biological markers for ageing[118] and disease prognosis[123,128–130].

Nevertheless, genetic regulation of N-glycosylation is not yet fully understood. Previous GWAS have focused either on the N-glycome of total blood plasma proteins or on glycosylation of one specific protein - immunoglobulin G (IgG).

Thus far, sixteen genomic regions have been associated with N-glycosylation of total blood plasma proteins[131–134], simultaneously assaying glycosylation of all proteins in plasma, without the possibility to distinguish protein-specific glycosylation. On the other hand, protein-specific GWAS of IgG N-glycans identified 27 associated genomic regions[135–138]. IgG antibodies are one of the most abundant proteins in human serum, and their alternative N-glycosylation is suggested to trigger different immune responses and thus impacts the action of the immune system[139]. Overall, the majority of loci identified as associated with N-glycosylation of proteins mapped onto genes encoding glycosyltransferases, enzymes directly involved in the biochemical pathway of N-glycosylation, or, in the specific case of IgG, involved in differentiation and maturation of B cells (since IgG is produced by fully differentiated B-lymphocytes called plasma cells). Protein N-glycosylation-associated loci were found also around genes associated with transcription factor activity, corroborating the idea that protein N-glycosylation is a complex process, genetically regulated by an intricate interplay of multiple genes[133,135].

In Chapter 2, I investigate the genetic regulation of glycosylation of transferrin protein. This represents the first protein-specific study for a protein other than IgG. Further, this allows for the first time to compare the genetic regulation of glycosylation of two different proteins - transferrin and IgG - potentially giving insight also on their protein-specific contribution to glycosylation of total plasma protein. After expanding the current knowledge of genetic regulation of protein glycosylation by including a new protein, in Chapter 3 I extend instead the number of variants assayed for association with protein glycosylation, by focusing on rare and low frequency pLOF and missense variants in exome sequencing. Further, I investigate the potential impact of rare variants associated with glycosylation on health-related traits.

In addition to the directly measured glycan structures, defined by the number of their glycan peak (GP for IgG, TfGP for transferrin), which were analysed in both Chapter 2 and 3, in Chapter 3 I assayed the genetic association also of several

glycan derived traits. These derived traits represent common biologically meaningful features shared among several measured glycans or the overall presence of a certain sugar structure on the totality of glycan traits measured (e.g. percentage of fucosylated glycans, triantennary glycans, monogalactosylated glycans, etc.). The structural characterization of actually measured transferrin (TfGP) and IgG (GP) glycan traits is available respectively at Supplementary Table 2 of Trbojević-Akmačić et al.[140] and Supplementary Table 1 of Huffman et al.[141] The description and computing formulas of derived transferrin glycan traits representing the relative abundance of a certain sugar structure are available at Supplementary Table 15 of the Chapter 3 of this thesis. The description and computing formulas of derived IgG glycan traits capturing shared biological features are available at Supplementary Table 1 of Huffman et al.[141], while those representing the relative abundance of a certain sugar structure are available at Supplementary Table 16 of the Chapter 3 of this thesis. Details about the number of transferrin and IgG glycan traits analysed, the different cohorts used, and their sample size are also visually summarised in Supplementary Figure 1. Finally, to point out that certain glycan structures tend to co-occur or exhibit consistent patterns, I show in Supplementary Figure 2 and Supplementary Figure 3 the inter-trait correlation structure of transferrin and IgG glycans respectively in VIKING cohort, as example.

1.8.2 Bile acid lipidome

Bile acids (BA), together with cholesterol, phospholipids and bilirubin, comprise the principal constituents of bile. They are synthesised from cholesterol in the liver and subsequently stored in the gallbladder. After ingestion of food, bile flows into the duodenum, where, thanks to BA, it contributes to the digestion of lipid-soluble nutrients[142]. BA are then absorbed from the terminal ileum and transported back to the liver via the portal vein - a process termed “enterohepatic circulation”[143]. Lipid molecules, including bile acids, can be quantified using two primary methods: direct infusion mass spectrometry analysis, also known as shotgun lipidomics, and liquid chromatography-mass spectrometry (LC-MS) analysis. Infusion-based approaches are advantageous for their simplicity and straightforward lipid quantification but have limitations in terms of sensitivity and the ability to distinguish between isomeric lipid species[144]. LC–MS is a powerful analytical technique used for separation, identification, and quantification of both unknown and known compounds as well as to elucidate the structure and chemical properties of different molecules. LC–MS typically provides a higher sensitivity than shotgun, allowing for the detection and quantification of a wide range of lipid species, including the low abundant ones[145]. Additional benefits of LC–MS include confident identification of lipid structures, enabling identification and differentiation of lipid isomers[146].

BA biosynthesis is the primary pathway for cholesterol catabolism, which occurs via two different pathways: the classical (or neutral) pathway and the alternative (or acidic) pathway, both involving several members of the cytochrome P450 enzyme superfamily (i.e. CYP7A1, CYP8B1, CYP7B1 and CYP27A1)[147]. Primary BA cholate (CA) and chenodeoxycholate (CDCA) represent the two end products of these pathways. In the intestinal lumen, especially in the colon, gut flora deconjugates, oxidates and dehydroxylates the primary BA produced in the liver to generate secondary BA, with deoxycholate (DCA), a CA derivative, and lithocholate (LCA), a CDCA derivative, being the most prevalent[143]. Once transported back to the liver, these secondary BA can be further processed to

form tertiary BA (e.g. tauroursodeoxycholate and ursodeoxycholate)[148], which represent only marginal BA species under normal conditions[149]. Overall, these syntheses and metabolic pathways allow the generation of several BA species, which ensures the perfect solubilisation and absorption of a broad range of lipophilic molecules in the intestine, but also perform a multitude of signalling activities in the body.

BA have in fact emerged as versatile hormone-like signalling molecules endowed with systemic endocrine functions[149], serving as ligands for G protein-coupled receptors, such as Takeda G-protein-coupled receptor 5 (TGR5)[150], and for nuclear hormone receptors, such as farnesoid X receptor (FXR)[151]. Through activation of these signalling pathways, BA have been shown to regulate not only their own synthesis and enterohepatic recirculation, but also triglyceride, cholesterol, energy and glucose homeostasis[149]. Accordingly, dysregulation of bile acid homeostasis has been linked to cholestatic liver disorders, cholesterol gallstone disease and other gallbladder-related conditions[147], inflammatory bowel disease[152], type 2 diabetes[153], obesity[154] and non-alcoholic fatty liver disease[155], but also to carcinogenesis in several tissues or organs[156].

Despite the recognised role of BA in human health and the detailed characterisation of BA's biochemical pathways, their genetic regulation is poorly understood. While several studies investigating the genetic contribution to the variability of blood metabolites[77,157–163] have been published so far, research focusing specifically on BA in a large sample from the general population is still lacking.

In Chapter 4, similarly to the previous chapters, I investigate the contribution of both common and low-frequency/rare variants to variability of bile acids, also reporting sex-specific associations. Further, I explore whether complex traits or diseases influence bile acids variability.

Chapter 2: Genetic regulation of transferrin and IgG glycome

2.1 Introduction

Glycosylation is a common post-translational modification that involves the attachment of sugar structures, called glycans, to the surface of human proteins. While the genetics of proteins themselves have been extensively studied, the discovery of genetic factors that contribute to the glycome - the set of glycans present on a protein - is still lagging behind. Thus far, three GWASs have been performed on the N-glycome of total human plasma proteins, which involved around 3500 individuals in the discovery set and have collectively identified 16 distinct genetic regions[131–133]. Out of these, 15 genetic regions have been confirmed in independent sample sets through validation within the original studies or a subsequent investigation by Sharapov *et al.*[134]. While these studies simultaneously assayed glycosylation of all proteins in plasma, protein-specific glycosylation GWAS have been so far limited to IgG protein. These publications have discovered 33 significantly associated genetic regions and replicated 29 of them in separate sample sets. The first GWAS found genetic variations near four glycosyltransferase genes and five other regions not previously linked to protein glycosylation[132]. Shen *et al.*[164] expanded this list by five regions, including one containing a glycosyltransferase gene. Wahl *et al.*[138] replicated several associations previously discovered and identified a novel one, unrelated to glycosyltransferase genes. Klarić *et al.*[135] drastically increased the sample size of the previous studies from around 2000 participants in the discovery stage to over 8000 participants, identifying 27 genetic regions associated with IgG glycosylation, 22 of which were consistently replicated in independent datasets.

In this chapter, I investigate genes and genetic variants influencing glycan traits of transferrin. Transferrin is a protein produced in the liver and released into the bloodstream, where it binds to iron and transports it to tissues and cells that require it. It plays an important role in iron metabolism, by maintaining adequate iron levels in the body and preventing iron overload. Human plasma transferrin

N-glycome was analysed at the level of released glycans by HILIC-UHPLC-FLD and separated into 35 glycan peaks, whose structural characterization is reported by Trbojević-Akmačić et al.[140] Since the intensity of chromatogram peaks obtained from glycan quantification can vary largely, raw data have undergone total area normalisation to transform measurements to comparable scales. Total area normalisation was performed by dividing the area of each of the 35 chromatographic peaks by the total area of the corresponding chromatogram. Resulting measures are therefore relative abundances of each glycan structure in the overall transferrin glycosylation profile. Normalised glycan traits were then log transformed to reduce the skewness of their distribution. Finally, batch effect creating sub-groups of measurements due to factors unrelated to biological variation was removed by using empirical bayes batch correction. The efficiency of this procedure in removing experimental noise was assessed by comparing, in raw and processed data, the variation of standard samples (6 pooled samples previously quantified, thus not affected by experimental errors from the current study) and the correlation of duplicated samples (9 samples from the current study which were analysed twice), present on each of the 12 96-well plates. As expected, correlation of duplicated samples overall increased thanks to the applied procedure (Supplementary Figure 4), while also batch effect across different plates was reduced (Supplementary Figure 5).

This study represents the first exploration of the genetic factors influencing the glycome of transferrin. Until now, it has not been possible to study protein-specific pathways regulating glycosylation for proteins other than IgG, due to technical challenges that have hindered the quantification of glycosylation of individual glycoproteins in large samples. Here, I report genetic variants associated with variation of 35 transferrin glycan traits (TfGP), and also compare these variants with those associated with 24 glycan traits (GP) of the IgG protein, whose structures have been described by Huffman et al.[141] For loci associated with the glycosylation of both transferrin and IgG, I assess whether the underlying causal variants are specific to each protein or rather shared between the two proteins. To the best of my knowledge, this is the first study investigating whether the same

post-translational modification of two proteins is regulated by the same genes and whether it is driven by the same underlying causal variants.

2.2 Published article

This work was published as an article in the journal Nature Communications on 24 March 2022 after completing formal peer review. A copy of the Author Accepted Manuscript prior to proofing is included below, provided under the terms of the Creative Commons Attribution License CC BY 4.0. The formatted article, detailed methods, and supplementary information are available at: https://doi.org/10.1038/s41467-022-29189-5.

In this study, I conducted single-cohort GWAS and performed meta-analysis of the transferrin and IgG glycome, starting from clean genotypic and phenotypic data. I also carried out down-stream analyses, with the exception of SMR-HEIDI analysis of the transferrin glycome and gene expression/complex traits, which was conducted by Yakov A. Tsepilov. I used my own scripts to perform pairwise conditional and colocalisation analysis of the transferrin and IgG glycome. Sodbo Z. Sharapov obtained gene expression data for selected genes and created Figure 5. Lucija Klarić helped formalise the likelihood ratio test for evaluating the impact of transferrin protein levels on transferrin glycome associations. I wrote the initial draft of the manuscript with the assistance of Irena Trbojević-Akmačić (for methods on glycan measurements and more technical description of glycan structures and binding sites in the introduction), Yakov A. Tsepilov (for SMR-HEIDI methods) and Sodbo Z. Sharapov (for technical details regarding data retrieval of gene expression data toward Figure 5). The full list of author contributions can be found in the “Author contributions” section of this article.

Genetic regulation of post-translational modification of two distinct proteins

Arianna Landini*[1], Irena Trbojević-Akmačić*[2], Pau Navarro[3], Yakov A. Tsepilov[4,5], Sodbo Z. Sharapov[4], Frano Vučković[2], Ozren Polašek[6,7] Caroline Hayward[3], Tea Petrović[2], Marija Vilaj[2], Yurii S. Aulchenko[4], Gordan Lauc*[2,8], James F. Wilson*[1,3], Lucija Klarić*[3]

1 Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
2 Genos Glycoscience Research Laboratory, Zagreb, Croatia
3 MRC Human Genetics Unit, Institute for Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom
4 Laboratory of Glycogenomics, Institute of Cytology and Genetics, Novosibirsk, Russia
5 Laboratory of Theoretical and Applied Functional Genomics, Novosibirsk State University, Novosibirsk, Russia
6 Department of Public Health, School of Medicine, University of Split, Split, Croatia
7 Algebra University College, Zagreb, Croatia
8 Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb, Croatia
* Authors contributed equally.
Correspondence to: J.F.W ( ) or L.K. ( )

Abstract

Post-translational modifications diversify protein functions and dynamically coordinate their signalling networks, influencing most aspects of cell physiology. Nevertheless, their genetic regulation or influence on complex traits is not fully understood. Here, we compare the genetic regulation of the same PTM of two proteins – glycosylation of transferrin and immunoglobulin G (IgG). By performing genome-wide association analysis of transferrin glycosylation, we identify 10 significantly associated loci, 9 of which were not reported previously. Comparing these with IgG glycosylation-associated genes, we note protein-specific associations with genes encoding glycosylation enzymes (transferrin - *MGAT5*, *ST3GAL4*, *B3GAT1*; IgG - *MGAT3*, *ST6GAL1*), as well as shared associations (*FUT6*, *FUT8*). Colocalisation analyses of the latter suggest that different causal variants in the FUT genes regulate fucosylation of the two proteins. Glycosylation of these proteins is thus genetically regulated by both shared and protein-specific mechanisms.

Introduction

Post-translational modifications (PTMs) are essential mechanisms used by cells to diversify and extend their protein functions beyond what is dictated by protein-coding sequences in the genome. These chemical reactions range from the addition of small moieties, such as phosphate (phosphorylation), complex biomolecules, as in glycosylation, to proteolytic cleavage[165]. PTMs alter the structure and properties of proteins and are thus involved in the dynamic regulation of most cellular events. It is common for a PTM enzyme to target multiple substrates or interact with multiple sites. For example, only 18 histone deacetylases target more than 3600 acetylation sites on 1750 proteins[166]. Environmental or pathological conditions can lead to dysregulation of PTM activities, which has been related to aging[167] and several diseases, including cancer, diabetes, and neurodegeneration[168–170]. Despite their importance, little is known about genetic regulation of post-translational modifications.

N-glycosylation is one of the most common protein PTMs, where carbohydrate structures called glycans are covalently attached to an asparagine (Asn) residue of a polypeptide backbone. N-glycans are characterised by vast structural diversity and high complexity. While polypeptides are encoded by a single gene, N-glycan structures result from a sophisticated interplay of glycosyltransferases, glycosidases, transporters, transcription factors, and other proteins[113]. Protein N-glycosylation is involved in a multitude of biological processes[117]. Accordingly, changes in N-glycosylation patterns have been associated with aging[118] and a wide range of diseases, including Parkinson's disease[119], lower back pain[120], rheumatoid arthritis[121], ulcerative colitis[122], Crohn's disease[122], type 2 diabetes[123] and cancer[124–126]. However, for most of these conditions it still remains to be clarified whether the disease causes changes in N-glycosylation or vice-versa. In addition, N-glycans are considered as potential therapeutic targets[127] and prognostic biological markers[123,128–130].

As with other PTMs, genetic regulation of N-glycosylation is not yet fully understood. Previous genome-wide association studies (GWAS) have so far focused either on the N-glycome of total blood plasma proteins as a whole or on glycosylation of one specific protein - immunoglobulin G (IgG)[131–138]. IgG antibodies are one of the most abundant proteins in human serum, and their alternative N-glycosylation is suggested to trigger different immune response and thus impacts the action of the immune system[139]. N-glycan structures are predominantly of the biantennary complex type and vary due to additions of core fucose, galactose, sialic acid, and bisecting N-acetylglucosamine (GlcNAc), with disialylated digalactosylated biantennary glycan with core fucose and bisecting GlcNAc being the most complex N-glycan structure on IgG[171]. While a clear overlap in genetic control between total plasma proteins and IgG N-glycosylation was highlighted by previous studies[133], it was not possible, until recently, to identify protein-specific N-glycosylation pathways for glycoproteins other than IgG due to technical challenges of their isolation in large cohorts.

Here we investigate whether the same PTM of two proteins is regulated by the same genes and whether they are driven by the same causal genetic variants. We report genes associated with the regulation of transferrin N-glycosylation and compare these with the genetic regulation of glycosylation of a different protein (IgG). Transferrins are blood plasma glycoproteins regulating the level of iron in an organism. Iron plays a central role in many essential biochemical processes of human physiology: the cells' need for iron in the face of potential danger as an oxidant has given rise to a complex system that tightly regulates iron levels, tissue distribution, and bioavailability[172]. Human transferrin has two N-glycosylation sites – at the N432 and N630 residues, with biantennary disialylated digalactosylated glycan structure without fucose being the most abundant glycan attached[173,174]. We performed genome-wide association meta-analysis (GWAMA) of 35 transferrin N-glycan traits (N=1890) and compared it with GWAMA of 24 IgG N-glycan traits (N=2020) in European-descent cohorts, discovering both protein-specific and shared associations. For loci associated with the N-glycosylation PTM of both transferrin and IgG, we used colocalisation

analysis to assess whether the underlying causal variants are protein-specific or rather shared between these proteins. We then suggested a molecular mechanism by which these independent causal variants could regulate the expression of glycosylation related genes in different tissues.

Results

Loci associated with transferrin N-glycosylation

To investigate the genetic control of transferrin N-glycosylation and assess whether the same genes and underlying causal variants are associated with N-glycosylation of both transferrin and IgG, we first performed GWAS of glycosylation for each protein (i.e. transferrin and IgG). A more extensive GWAS on the genetic regulation of IgG glycosylation has already been published[135], so we focus here on glycosylation of transferrin. We performed GWAS of 35 ultra-high-performance liquid chromatography (UHPLC)-measured transferrin N-glycan traits and Haplotype Reference Consortium (HRC) r1.1[28]-imputed genetic data in two cohorts of European descent, CROATIA-Korcula (N=948) and VIKING (N=952). Overall, we identified 8 loci genome-wide significantly associated (p-value ≤ 1.43×10^{-9}) with transferrin N-glycans in the CROATIA-Korcula cohort (Supplementary Figure 1, Supplementary Data 1), 6 of which replicate in the VIKING cohort (p-value ≤ 0.00625) (Supplementary Figure 2, Supplementary Data 2). Replicated loci contained genes encoding glycosyltransferases, enzymes directly involved in the biochemical pathway of N-glycosylation (*MGAT5*, *ST3GAL4*, *B3GAT1*, *FUT8* and *FUT6*) and the transferrin (*TF*) gene. Cohort-specific heritability estimates for each transferrin glycan trait (Supplementary Data 3) ranged from 0% (VIKING TfGP2 and TfGP12) to 67% (CROATIA-Korcula TfGP23) and were high overall (>40% for the majority of the traits), similar to heritabilities previously reported for the total plasma glycome[175] as well as immunoglobulin G glycosylation[176]. To further increase the power of

our analyses, we performed fixed-effect inverse-variance meta-analysis of the discovery and replication cohort, discovering 2 additional loci (*FOXI1* and *HNF1A*) (Table 1). To identify secondary association signals at each genomic region, we performed approximate conditional analysis on transferrin N-glycan traits using GCTA-COJO software[177]. Overall, we identified 15 independently contributing variants, located in 10 genomic loci significantly associated (p-value ≤ 1.43×10^{-9}, Bonferroni adjusted for the number of glycan traits) with at least one of the 35 transferrin N-glycan traits (Table 1, Figure 1, complete list of all associations in Supplementary Data 4). Multiple SNPs independently contributed to transferrin N-glycan variation in 4 out of 10 loci, all mapping to glycosyltransferase genes. The highest number of independently associated SNPs (3) was observed for the glucuronyltransferase locus, *B3GAT1,* while two SNPs contributed to transferrin N-glycan levels in the acetylglucosaminyltransferase locus (*MGAT5*), the fucosyltransferase locus (*FUT8*) and the sialyltransferase locus (*ST3GAL4*) (Supplementary Data 5).

To assess the potential impact of transferrin protein levels on the reported transferrin glycome associations, we utilised the transferrin *cis*-protein quantitative trait locus (pQTL), rs8177240[178]. This variant is associated with transferrin protein abundance and so can act as a proxy for protein levels and is not in linkage disequilibrium with glycan QTL (glyQTL) rs6785596, the sentinel glycosylation-associated SNP in *TF* (LD r^2= 0.02). Two glycans, TfGP3 and TfGP9, were significantly associated with transferrin *cis*-pQTL. However, both *cis*-pQTL and the glyQTL (rs6785596) contribute to the variation of TfGP3 levels, while only the *cis*-pQTL contributes to levels of TfGP9. Overall, this suggests that glycan associations with the *TF* gene are only completely accounted for by the transferrin protein levels in the case of TfGP9 (Supplementary Data 6). Further details about the potential effects of transferrin gene expression and protein levels can be found in Supplementary Results and Supplementary Data 19.

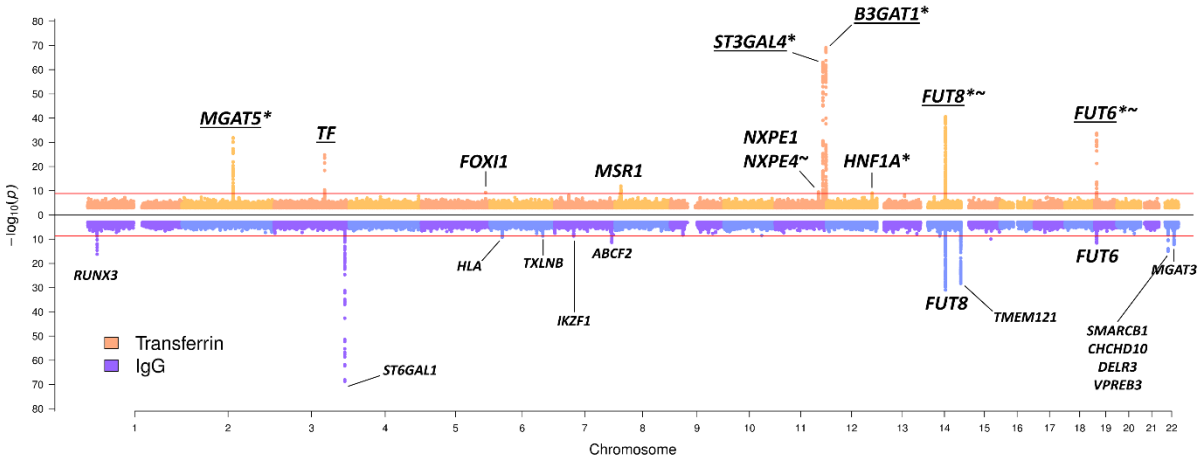


Figure 1. Transferrin and IgG N-glycome GWAMA summary Miami plot. Miami plot pooling together meta-analysis results obtained across all 35 transferrin glycan traits at the top in orange, and across all 24 IgG glycan traits at the bottom in blue. The pooling was performed by selecting the lowest p-value (y-axis) from all 35 (TF) or 24 (IgG) glycan traits for every genomic position (x axis). For transferrin N-glycome associations, "*" marks loci previously reported in total blood plasma N-glycome GWAS[131–134], while "~" marks loci previously reported in IgG N-glycome GWAS[135–138]. Bonferroni-corrected genome-wide significance threshold for the transferrin N-glycome meta-analysis (horizontal red line in the top part of the plot) corresponds to 1.43×10^{-9}, while Bonferroni-corrected genome-wide significance threshold for the IgG N-glycome meta-analysis (horizontal red line in the bottom part of the plot) corresponds to 2.08×10^{-9}. For simplicity, SNPs with p-value > 1×10^{-3} are not plotted. Gene or sets of genes annotated for transferrin N-glycome loci have been prioritised in this study; gene or sets of genes annotated for IgG N-glycome loci are those prioritised by Klarić et al.[135]. P-values are derived from two-sided Wald test with one degree of freedom.

Table 1. Loci genome-wide significantly associated with at least one of the 35 transferrin N-glycan traits in GWAMA. Glycosyltransferase loci are reported at the top of the table, while other loci are listed at the bottom of the table. Each locus is represented by the SNP with the strongest association in the region, according to the p-value rejecting the null hypothesis of no association with at least one of 35 transferrin glycan traits. An association was considered significant if the p-value was lower than or equal to 1.43×10^{-9}, the genome-wide significance threshold Bonferroni-corrected for the number of glycan traits.

Locus	Gene	SNP	EA	OA	EAF	No. of SNPs	Lead glycan	Phe. var.	No. of glycans	Beta	SE	P
2:134839539-135024803	*MGAT5*	rs2442046	C	G	0.747	2	TfGP23	0.071	4	-0.44	0.037	$1.38x10^{-32}$
11:126052988-126312874	*ST3GAL4*	rs4055121	T	C	0.12	2	TfGP17	0.131	9	0.782	0.046	$9.67x10^{-64}$
11:133906302-134613230	*B3GAT1*	rs74622686	A	G	0.905	3	TfGP21	0.144	3	0.931	0.053	$8.53x10^{-70}$
14:65751627-66281192	*FUT8*	rs2411815	A	T	0.306	2	TfGP20	0.092	3	-0.469	0.035	$2.69x10^{-41}$
19:5813766-5841356	*FUT6*	rs12019136	A	G	0.039	1	TfGP32	0.079	5	-1.016	0.083	$2.00x10^{-34}$

3:133433470-133499063	*TF*	rs6785596	A	T	0.047	1	TfGP3	0.065	3	0.787	0.075	$1.57x10^{-25}$
5:169535155-169535155	*FOXI1*	rs115399307	T	C	0.018	1	TfGP23	0.031	1	0.941	0.152	$5.18x10^{-10}$
8:15831868-16623073	*MSR1*	rs41341748	A	G	0.027	1	TfGP35	0.031	1	0.778	0.109	$1.16x10^{-12}$
11:114381448-114384985	*NXPE1/ NXPE4*	rs1671819	A	G	0.454	1	TfGP14	0.02	1	-0.2	0.032	$3.32x10^{-10}$
12:121420263-121424861	*HNF1A*	rs2393775	A	G	0.638	1	TfGP28	0.019	1	-0.203	0.033	$8.97x10^{-10}$

Locus - coded as “chromosome: locus start–locus end” (GRCh37 human genome build); Gene - suggested candidate gene; SNP - variant with the strongest association in the locus; EA - SNP allele for which the effect estimate is reported; OA - other allele; EAF - frequency of the effect allele; No. of SNPs - number of SNPs in the locus independently contributing to trait variation according to GCTA-COJO; Lead glycan - glycan trait with the strongest association to the reported SNP; Phe. var. - proportion of variance in phenotype explained by the strongest associated SNP; No. of glycans - number of glycan traits significantly associated with variants at the given locus; Beta - effect estimate for the SNP and glycan with the strongest association in the locus; SE - standard error of the effect estimate, P - p-value of the effect estimate (two-sided Wald test with one degree of freedom).

Prioritising candidate genes associated with transferrin N-glycosylation

For the 10 loci associated with the transferrin N-glycome, we identified plausible candidate genes following multiple lines of evidence, such as evaluating the biological role of the candidate gene in the context of protein N-glycosylation, assessing colocalisation with eQTL, and investigating variant effects on the coding sequence or on putative transcription factor binding sites.

The majority of genes that were closest to variants associated with transferrin N-glycosylation had a clear biological link to protein N-glycosylation. In particular, for 5 out of 10 loci, the closest genes (i.e. *MGAT5*, *ST3GAL4*, *B3GAT1*, *FUT8,* and *FUT6*) encode glycosyltransferases, key enzymes in protein glycosylation, that have been previously associated with IgG and/or total plasma protein glycosylation (Supplementary Data 7). Another gene closest to variants associated with transferrin N-glycosylation and with a validated functional role in plasma protein glycosylation is *HNF1A*, a transcription factor previously associated with protein fucosylation (Supplementary Data 7). On the other hand, we also identified 3 loci that had not been associated with N-glycosylation. A locus on chromosome 3 contains the transferrin (*TF*) gene, which encodes the transferrin glycoprotein. A locus on chromosome 5 containing *FOXI1* encodes a member of the forkhead family of transcription factors (Forkhead box I1). Finally, a locus on chromosome 8 contains the *MSR1* gene, encoding the class A macrophage scavenger receptor, a trimeric integral membrane glycoprotein. Another gene of potential biological relevance at the chromosome 8 locus is the tumour suppressor candidate 3 (*TUSC3*), which encodes a protein localised to the endoplasmic reticulum and acting as a component of the oligosaccharyltransferase complex, responsible for N-linked protein glycosylation.

Using eQTL analysis in PhenoScanner, variants associated with transferrin N-glycosylation (and their proxies, LD $r^2 > 0.8$) were identified to be significantly associated with the expression of multiple genes in several human tissues

involved in transferrin metabolism (Supplementary Data 8a). For example, variants associated with transferrin glycosylation were associated with *ST3GAL4* expression in liver and whole blood, with *B3GAT1* expression in visceral adipose omentum, liver, and whole blood, with *TF* expression in several adipose tissues and with *HNF1A*, *FUT8,* and *MGAT5* expression in whole blood. The majority of these genes were also the closest to the strongest association in the locus. We next used Summary data-based Mendelian Randomization (SMR) analysis followed by the Heterogeneity in Dependent Instruments (HEIDI) test[179] to assess whether expression of these genes colocalises with transferrin glycosylation (TfGP) traits. SMR-HEIDI provided evidence of colocalisation, suggesting that the same underlying causal SNPs are likely to regulate both transferrin glycosylation traits and gene expression, for *B3GAT1* in liver and peripheral blood and *ST3GAL4* in liver (Supplementary Data 8b).

We next explored whether any of the SNPs independently contributing to transferrin glycosylation (or their proxies) result in a change of amino acid sequence using the Ensembl Variant Effect Predictor (VEP) v97[180]. While the majority of associated variants (> 60%) were classified as intronic, several SNPs were identified as missense variants: rs115399307 (chr5:169535155-T/C) causes the substitution of the non-polar, aliphatic amino acid isoleucine (I) to the polar, hydrophilic amino acid threonine (T) in the FOXI1 transcription factor. Similarly, *NXPE4* variant rs550897 (chr11:114442103-A/G, r^2=0.94 with rs1671819) causes an amino acid substitution from tyrosine (Y) to histidine (H). Genetic variant rs41341748 (chr8:16012594-A/G) disrupts a stop codon sequence in *MSR1*, causing an elongated transcript with the amino acid arginine (Arg) added to the protein chain (Supplementary Data 9). The *FUT6* variant rs17855739 (chr19:5831840-T/C, r^2=0.95 with rs12019136) maps to the enzyme's catalytic domain and the allele T results in a change from negatively charged glutamic acid (E) to positively charged lysine (K), which leads to a full-length, but inactive, enzyme[181]. While the effect of reduced enzymatic activity on fucosylation of transferrin glycans needs to be experimentally validated, we observed that levels of TfGP32 are significantly lower in individuals carrying the

T allele at rs17855739, compared to those with two C alleles (Supplementary Figure 3). The structure of TfGP32 is currently not known, but its genetic association signal colocalises with two plasma glycan traits containing antennary fucose (A4F1G3S[3,3+6,3+6]3, A4F1G4S[3,3,3,6]4) and overall plasma antennary fucosylation (Supplementary Figure 4, Supplementary Results). Overall, this suggests that transferrin might contribute to these plasma glycan peaks and that TfGP32 might contain antennary fucose and could therefore be a proxy for *FUT6* activity. However, these inferences need to be further experimentally validated.

Finally, we used the regulatory sequence analysis tools (RSAT)[182] to assess if variants associated with transferrin N-glycosylation overlap transcription factor-binding sites and hence may be hypothesised to affect transcription factor binding. From the list of prioritised genes, we selected the two encoding transcription factors, *FOXI1* and *HNF1A,* and checked whether associated variants in the remaining 8 loci were likely to affect binding of these transcription factors. Overall, binding of both FOXI1 and HNF1A transcription factors might be affected by the sentinel variant (the SNP with lowest p-value in the region for the given glycan trait) in the *FUT8* gene. Similarly, binding of HNF1A might be affected by the sentinel variants in the *TF* and *ST3GAL4* loci (Supplementary Data 10).

Shared genetic associations with complex traits and diseases

To assess whether variants associated with transferrin glycosylation were also associated with complex traits and diseases we used PhenoScanner[183], followed by SMR-HEIDI to determine whether the shared associations are caused by the same underlying causal variant (colocalisation). We observed an overlap of transferrin N-glycan-associated SNPs and their proxies with variants associated with complex trait- and disease-associated variants for 5 out of 10 glycosylation loci (Supplementary Data 11a). For the remaining shared associations, we had no power to assess colocalisation (Supplementary Results for further details).

Interestingly, variants at the *TF* locus have been previously associated with serum concentration of carbohydrate-deficient transferrins (CDT) (Supplementary Data 11a), less glycosylated transferrin isoforms traditionally used as a biomarker of excessive alcohol consumption[184], thus corroborating our finding for a related trait. We then assessed SMR-HEIDI findings (Supplementary Data 11b) using bi-directional Mendelian Randomisation (MR) to infer the causal direction between glycan traits and complex traits, and further validated the colocalisation results using a Bayesian approach. After Bonferroni correction (p-value < $0.05/8 = 6.25\times10^{-3}$), there was no evidence of complex traits having an effect on glycan traits. However, we found positive associations of levels of TfGP14 and ulcerative colitis, and levels of TfGP28 and C-reactive protein levels, LDL and total cholesterol (Supplementary Data 12), although these results relied on few instrumental variables and were driven by associations in a single locus (Supplementary Figure 5). Contrary to the SMR-HEIDI analysis, Bayesian colocalisation analysis suggested that the association of ulcerative colitis and TfGP14 levels at the *NXPE1/NXPE4* locus are driven by independent, trait-specific causal variants. However, colocalisation confirmed that the associations between TfGP28 and C-reactive protein levels, LDL and total cholesterol are driven by a shared causal variant at the *HNF1A* locus (Supplementary Data 13, Supplementary Figure 5).

Comparison of genetic regulation of glycosylation of transferrin and immunoglobulin G

One of the main aims of this study is to understand if the N-glycosylation of two proteins is regulated by the same enzymes and if so, whether the same underlying genetic variant or a set of variants are driving the process. To address this question, in addition to the already described GWAMA of transferrin glycosylation, we performed a GWAMA of 24 UHPLC IgG N-glycan traits in the same individuals (N=2020), following the same protocol. 13 loci were significantly associated with at least one of the 24 IgG N-glycan traits (Figure 1, Supplementary Data 14). The IgG N-glycome GWAS was annotated using genes

or sets of genes prioritised by Klarić et al.[135] By comparing the two GWAS we discovered mainly protein-specific associations, but also two genomic regions that were associated with glycosylation of both proteins (Figure 1). The protein-specific associations were with genes encoding known glycosylation enzymes (transferrin - *MGAT5*, *ST3GAL4*, *B3GAT1*; IgG - *ST6GAL1*, *MGAT3*), but also with transcription factors (transferrin - *HNF1A, FOXI1*; IgG - *IKZF1*, *RUNX3*), the protein itself (transferrin - *TF*; IgG - *TMEM121*, gene in proximity of *IGH* genes encoding immunoglobulin heavy chains) as well as other genes (transferrin - *MSR1*; IgG - *TXLNB*, *ABCF2*, *SMARCB1* region, HLA-region). Interestingly, the regions containing *FUT8* and *FUT6,* genes encoding fucosyltransferases, enzymes adding core and antennary fucose, respectively, to the synthetized glycan, were associated with glycosylation of both proteins (Figure 1). We then proceeded to assess whether the same underlying causal variants in these regions are controlling glycosylation for both proteins using colocalisation analysis.

Given that multiple glycan traits of the same protein can be associated with the same locus, we first asked whether all glycan traits of the same protein associated with a certain locus, colocalise (Supplementary Figure 6). Indeed, we found strong support for colocalisation (PP.H4 > 80 %, where PP.H4 represents the posterior probability for the same underlying causal variant contributing to trait variation), suggesting that for a given protein, all glycan traits associated with these loci are regulated by the same underlying causal variant (Supplementary Data 15, Supplementary Figures 7-9). One example of within-protein colocalisation can be seen in Figure 2. We next tested whether at the same genomic region, glycosylation of two different proteins is regulated by the same underlying causal variants. For this, we selected as the protein-representative glycan trait the one with the lowest p-value in the given region (one pair for each locus - transferrin TfGP20 and IgG GP7 for the *FUT8* locus and transferrin TfGP32 and IgG GP20 for the *FUT6* locus) and proceeded to test for colocalisation between glycosylation of the two proteins. We found strong support against colocalisation in both genomic regions (PP.H3 = 100% at *FUT8* locus,

PP.H3 = 99.71% at *FUT6* locus, where PP.H3 represents the posterior probability for different underlying causal variants contributing to trait variation) (Figure 3 and Figure 4, Supplementary Data 16). Since colocalisation methods are sensitive to multiple independent variants in the region contributing to the trait variation, which was the case here, we validated our findings with the PwCoCo approach[100] (Methods) and again, obtained robust evidence against the colocalisation hypothesis for all tested traits in both loci (Supplementary Data 16 and Supplementary Results for further details).

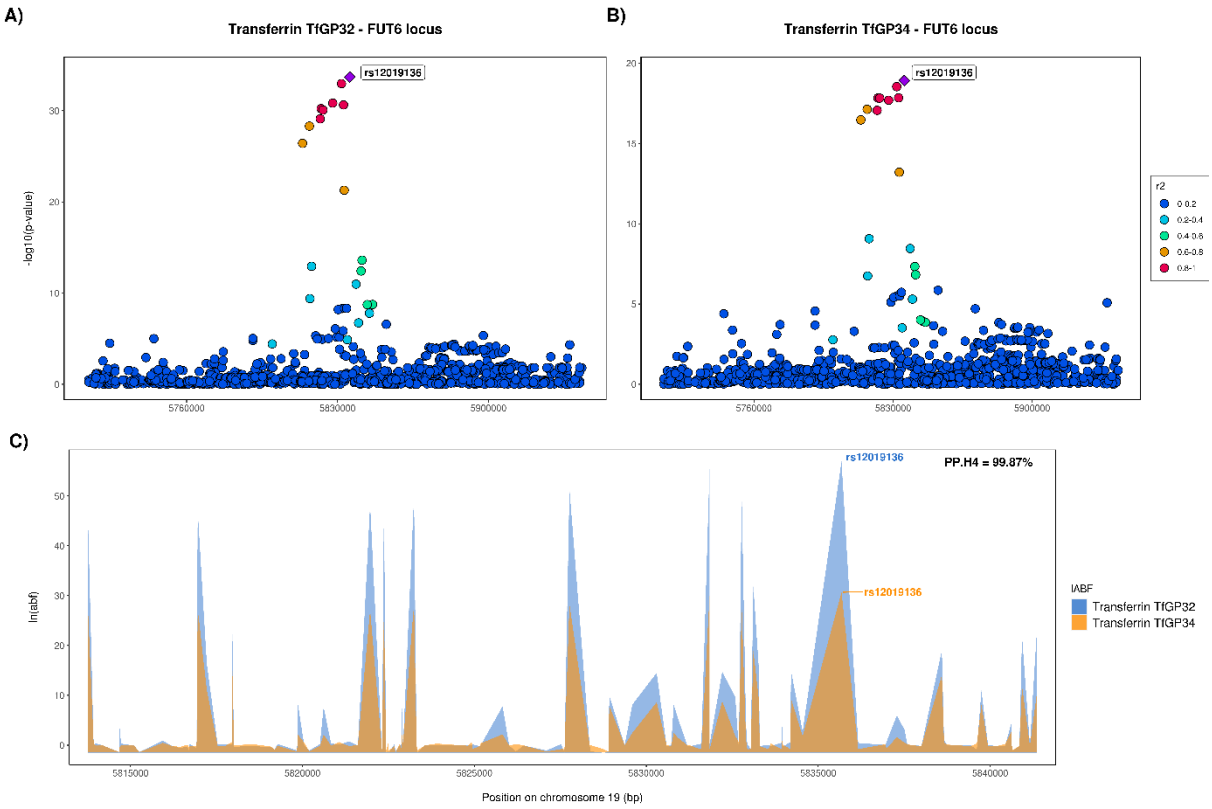


Figure 2. Local association patterns for transferrin (A) TfGP32 and (B) TfGP34 glycans, and (C) their colocalisation pattern at the *FUT6* locus. TfGP32 and TfGP34 association patterns colocalise, with PP.H4 (posterior probability for hypothesis 4, of colocalisation) of 99.87%. The natural logarithm of Approximate Bayes Factor (ABF) of each SNP for transferrin TfGP32 and transferrin TfGP34 in the *FUT6* region shows that TfGP32 and TfGP34

associations are concordant (the patterns of ln(ABF) calculated for each SNP of both traits overlap), suggesting that the same underlying causal variant is associated with both traits. SNP most strongly associated in the region with the listed glycan trait is reported in bold and labelled.

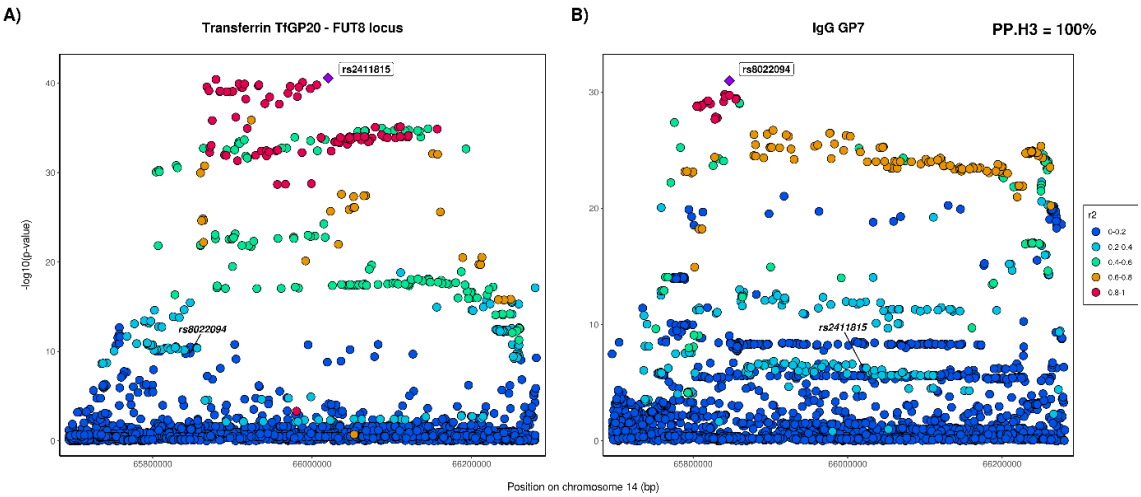


Figure 3. Local association patterns for (A) transferrin TfGP20 and (B) IgG GP7 glycans at the *FUT8* locus. TfGP20 and IgG GP7 association patterns do not colocalise, with PP.H3 (posterior probability for hypothesis 3, of different causal variants) of 100%. Colocalisation patterns are not reported since the width of the *FUT8* region makes the plot non-informative. SNP most strongly associated in the region with the listed glycan trait is reported in bold and labelled. For comparison, SNP most strongly associated with the other listed glycan trait is reported in italic, in the same panel.

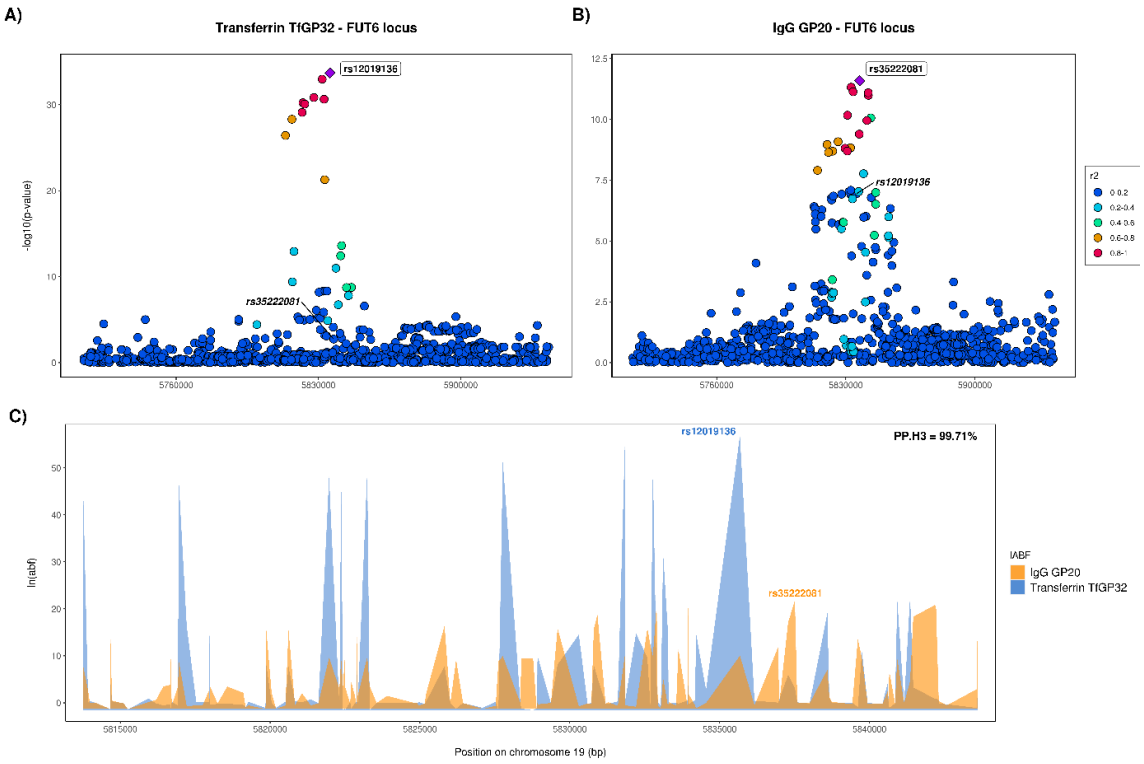


Figure 4. Local association patterns for (A) transferrin TfGP32 and (B) IgG GP20 glycans, and (C) their colocalisation pattern at the *FUT6* locus. TfGP32 and IgG GP20 association patterns do not colocalise, with PP.H3 (posterior probability for hypothesis 3, of different causal variants) of 99.7%). The natural logarithm of Approximate Bayes Factor (ABF) of each SNP for transferrin TfGP32 and IgG GP20 in the *FUT6* region shows that TfGP32 and GP20 associations are not concordant (the patterns of ln(ABF) calculated for each SNP of both traits do not overlap), suggesting that two different underlying causal variants in this region regulate glycosylation of these two proteins. SNP most strongly associated in the region with the listed glycan trait is reported in bold and labelled. For comparison, SNP most strongly associated with the other listed glycan trait is reported in italic, in the same panel.

Having established that different underlying causal variants regulate glycosylation at the *FUT6* and *FUT8* loci, we next explored the potential mechanisms behind these associations. The RSAT analysis suggests that the sentinel transferrin glycosylation SNP in the *FUT8* region might be affecting binding of the HNF1A transcription factor (Supplementary Data 10). Similarly, it was previously shown that the sentinel IgG glycosylation SNP in the same *FUT8* region potentially affects binding of the IKZF1 transcription factor[135]. In addition, we observed protein-specific associations with two transcription factors: transferrin glycosylation was associated with variants in the *HNF1A* locus and IgG glycosylation was associated with variants in the *IKZF1* locus (Figure 1). We therefore checked expression of these genes in tissues where the two proteins are predominantly expressed. It is known that plasma transferrin, encoded by *TF* gene, is mostly secreted by hepatocytes[185], while IgG, the heavy chain constant region of which is encoded by *IGHG* gene, is predominantly synthesised by the antibody-secreting plasma cells, the fully differentiated form of B-lymphocytes[186]. Indeed, we see that *IGHG1* (encoding the most prevalent IgG1 subclass) is highly expressed in plasma cells and has low expression in hepatocytes, while the converse is true for *TF* (Figure 5). Similarly, the transcription factor encoded by *HNF1A* is predominantly expressed in the hepatocytes, while *IKZF1* is mainly expressed in plasma cells (Figure 5). Altogether these suggest that two distinct causal variants regulating glycosylation of transferrin and IgG in the *FUT8* locus might have tissue-specific effects, where the transferrin-associated variant affects the binding of HNF1A in liver and the IgG-associated variant affects the binding of IKZF1 in plasma cells, with both influencing expression of the *FUT8* gene and therefore affecting fucosylation of the two proteins.

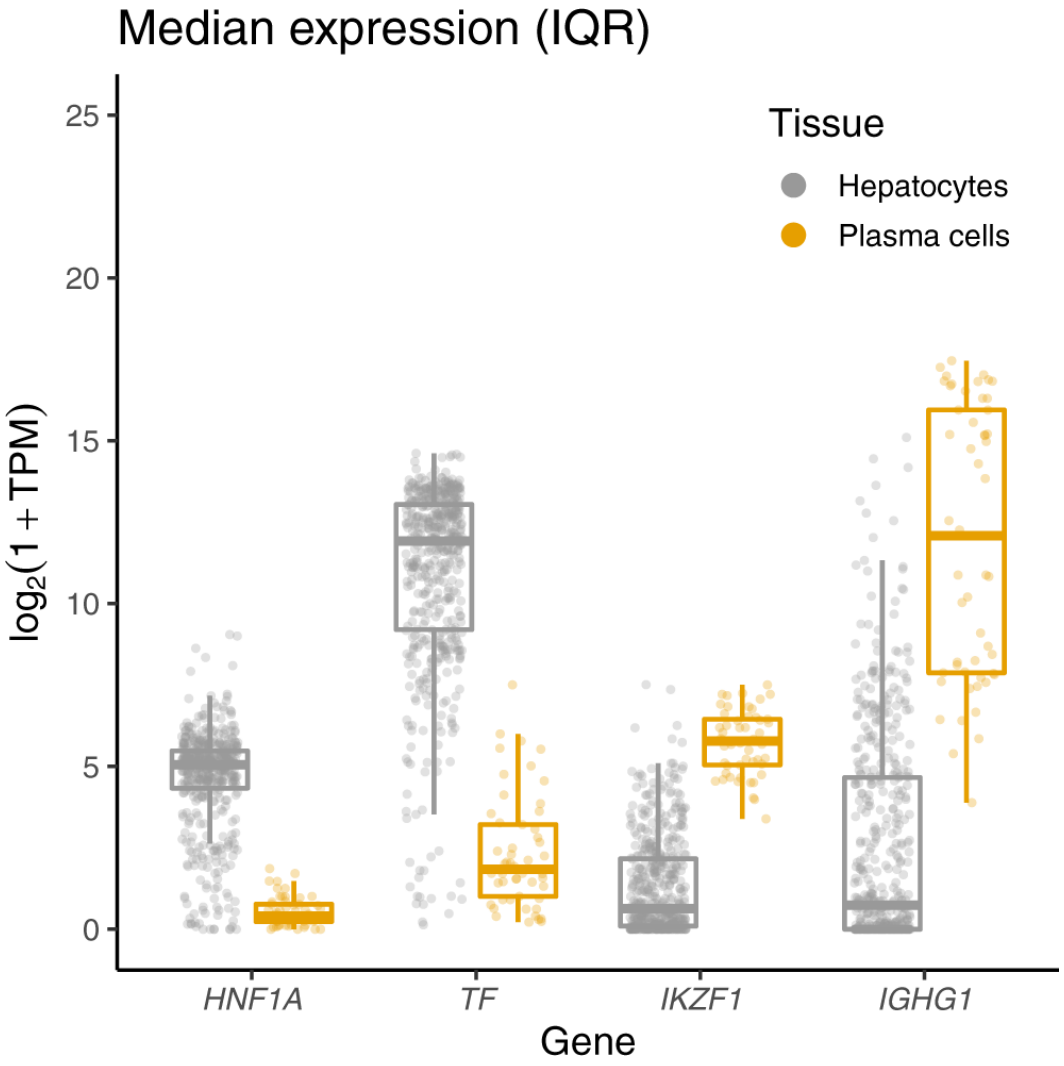


Figure 5. Expression of *TF, IGHG1, HNF1A* and *IKZF1* in main tissues of transferrin and IgG synthesis (liver and plasma cells). Gene expression data, expressed in gene counts, was scaled to transcripts per million (TPM) and $\log_2$(1+TPM) transformed. The data for hepatocyte (N=513) and plasma (N=53) cell samples were obtained from the ARCHS4 portal[187]. *TF* encodes transferrin protein, *IGHG1* encodes the constant region of immunoglobulin heavy chains, *HNF1A* and *IKZF1* encode two transcription factors involved in glycosylation of transferrin and IgG respectively. In the plot, the middle line represents the median, lower and upper limits of the box represent 1st and 3rd quartile, whiskers represent 1.5 interquartile range. All individual data points are overlapped to the box plot.

Discussion

Post-translational modifications (PTMs) are essential mechanisms that dynamically regulate a large portion of cellular events by altering the structure and properties of proteins[165]. In common with other PTMs, genetic regulation of protein N-glycosylation has not been extensively investigated. Here, we performed genome-wide association meta-analysis of glycosylation of two proteins - transferrin and IgG - and compared how their glycosylation is genetically regulated. In the GWAS of the transferrin N-glycome, (N=1890), we identified 10 significantly associated loci, two of which (near *FOXI1* and *MSR1*) were never previously associated with the glycome of any protein. The other eight have been previously associated with glycosylation of transferrin, total plasma proteins and/or IgG (Supplementary Data 7). The previous study on carbohydrate-deficient transferrin (CDT)[188], a composite measure that gives partial insight into the sialylation status of the protein, reported two genetic regions associated with the trait, near *PMG1* and *TF*, one of which we also found in this study (*TF*). Here, we were able to measure 35 different transferrin glycan traits, providing higher resolution of underlying structures and insight into the overall transferrin *N*-glycome. The total plasma glycome quantifies the glycome of all proteins in plasma, but without information on which glycan was bound to which protein. Given that IgG and transferrin are among the most abundant plasma glycoproteins[117], an overlap in genetic control of transferrin and IgG N-glycomes with that of total plasma proteins is to be expected. Sharapov et al.[133] previously indicated that some of the genomic loci associated with the plasma glycome overlap with loci associated with IgG N-glycosylation. The present work suggests that the *MGAT5, ST3GAL4,* and *B3GAT1* loci, that were also observed in the total plasma protein GWAS, might be capturing a signal within plasma protein glycosylation that comes from transferrin N-glycosylation.

We then compared the genetic architecture underlying glycosylation of transferrin and IgG proteins. Using the GWAS from this study we showed that there are both protein-specific and shared genetic loci. Looking specifically at

glycosyltransferase enzymes, the main “drivers” of this post-translational modification, that catalyse the transfer of saccharide moieties from a donor to an acceptor molecule, *MGAT5, ST3GAL4,* and *B3GAT1* were only associated with transferrin while *ST6GAL1* and *MGAT3* were only associated with glycosylation of IgG. On the other hand, two fucosyltransferase genes, *FUT8* and *FUT6,* were associated with both proteins. Since antennary fucose (produced by the FUT6 enzyme) is not typically found on IgG, we hypothesise that IgG glycosylation might be indirectly associated with *FUT6* through antennary fucosylation of other enzymes or proteins involved in glycosylation of IgG. Even though the genes encoding FUT6 and FUT8 enzymes were associated with glycosylation of both proteins, using Approximate Bayes Factor colocalisation analysis, we showed that associations with transferrin and IgG N-glycosylation at these genomic regions is driven by independent underlying causal variants, where one variant regulates fucosylation of transferrin and the other of IgG. Our results suggest that while the same fucosyltransferase enzymes are involved in N-glycosylation of both transferrin and IgG proteins, the process is independently regulated by protein-specific causal variants.

There are at least two mechanisms that could explain how different variants in an enzyme-coding gene could have distinct effects on two different substrates. If the two variants were in the coding region of the gene and affected the amino-acid sequence of the enzyme, they could affect the enzyme’s specificity for binding each protein. However, none of the sentinel variants in the *FUT8* and *FUT6* loci were in strong linkage disequilibrium (LD) with coding variants from the enzymes’ active sites, suggesting that this is likely not the mechanism of regulation of fucosylation of the two proteins. In addition, overall, SNPs associated with transferrin glycosylation predominantly mapped to regulatory rather than coding regions of the genome (Supplementary Data 9). The other hypothesis is that these two variants affect the expression of enzymes in different tissues. In common with all other antibodies, most of IgG found in blood plasma is produced by bone marrow plasma cells, the fully differentiated form of B-cells[186]. The transferrin found in blood plasma is mostly produced by liver hepatocytes[185]. In

addition, the glycomes of the two proteins were also associated with different transcription factor genes, namely, variants in the *IKZF1* region were associated with IgG glycosylation, and variants in the *HNF1A* region with transferrin glycosylation. IKZF1, a transcription factor predominantly expressed in immune cells and tissues, has been functionally validated as a regulator of IgG core fucosylation in lymphoblastoid cells: IKZF1 binds to regulatory regions of *FUT8* and, in turn, knockdown of *IKZF1* results in increased expression of *FUT8* and increased core fucosylation of IgG[135]. On the other hand, we showed that transferrin glycosylation-associated variants in the *FUT8* region might affect the binding of HNF1A, a transcription factor predominantly expressed in the liver. Lauc *et al*.[132] have shown that *HNF1A* knockdown results in down-regulation of *FUT6* and up-regulation of *FUT8* in the HepG2 hepatocyte cell line. While it might be expected that a change in levels of FUT6 and FUT8 enzymes would impact levels of antennary and core fucosylation (their enzymatic products), this link, especially in the context of transferrin glycosylation, has yet to be experimentally proven. Overall, our data could suggest that the two different causal variants may affect the binding of different transcription factors in different tissues and therefore regulate the glycosylation of the two plasma proteins in a tissue-specific manner. However, the effect of specific SNPs on binding of the two transcription factors and their downstream effect on expression of fucosyltransferases in a tissue-specific manner, still needs to be functionally validated.

In addition to HNF1A, variants in the *FUT8* locus associated with transferrin glycosylation might also be affecting the binding of the FOXI1 transcription factor. However, unlike HNF1A, possible involvement of FOXI1 in the regulation of the transferrin fucosylation is to date unknown and would require functional validation. We also found that HNF1A binding could also be affected by variants associated with glycosylation in the *TF* and *ST3GAL4* genes. While these relationships were hitherto undocumented and need further supporting evidence, they may suggest that HNF1A might regulate multiple genes associated with transferrin N-glycosylation.

The most strongly N-glycosylation-associated variant for the *TF* gene, rs6785596, can be considered an example of a "*cis*-glyQTL": a genomic locus that explains variation in glycosylation levels and is local to the gene encoding the protein being glycosylated. Similar was observed for IgG glycosylation, where associated variants mapped to the *IGH* locus[137], a genetic region encoding the heavy chain of immunoglobulin G. The transferrin glycosylation "*cis*-glyQTL" is an eQTL for expression of transferrin in adipose tissue, but not in liver, where transferrin is predominantly expressed. The variant is also in middling LD (r^2 = 0.57) with a missense variant, rs179989, providing potential alternative explanation for the association. Altogether, the exact mechanism of how these "*cis*-glyQTL" could be affecting glycosylation levels remains unclear. Considering causal relations between the transferrin glycome and complex traits and diseases, we found associations between levels of TfGP28 and C-reactive protein levels, LDL and total cholesterol. These associations were, however, driven by a single locus encoding the transcription factor *HNF1A*, suggesting that the locus might be pleiotropic and has an impact on both transferrin glycan levels and complex traits.

In conclusion, by performing the GWAS of the plasma transferrin N-glycome and comparing it with that of the IgG N-glycome, we were able to describe similarities and differences in the genetic regulation of post-translational modification of two different proteins. When focusing on glycosyltransferases, the main enzymes of this PTM, we showed that there are associations specific to each protein, but also those that are involved in glycosylation of both proteins. For the latter, we showed that fucosylation of transferrin and IgG are regulated by independent, protein-specific variants in the FUT8 and FUT6 genes. In the FUT8 region, these variants are likely to regulate fucosylation of transferrin and IgG in a tissue-specific manner, potentially acting through tissue-specific transcription factors. Additional studies, with larger sample sizes and focusing on other non-IgG proteins, will be necessary to further unravel the genetic architecture of N-glycosylation and to understand its relationship with human diseases and complex traits. While PTMs involved in intracellular signalling (e.g. phosphorylation) remain difficult to

quantify in a high-throughput manner, here we investigated glycosylation of two plasma proteins, constraining the analysis to one type of PTM in the extracellular space. The impact of genetics on other, both intra- and extra-cellular post-translational modifications will be an interesting area of future research.

Methods

Population cohorts

The CROATIA-Korcula isolated population cohort includes samples of blood DNA, plasma and serum, anthropometric and physical measurements, information related to general health, medical history, lifestyle, and diet for ~3000 residents of the Croatian island of Korčula[189]. Written informed consent was given and the study was approved by the Ethics Committee of the Medical School, University of Split (approval id: 2181-198-03-04/10-11-0008). The Viking Health Study - Shetland (VIKING) is a family-based, cross-sectional study that seeks to identify genetic factors influencing cardiovascular and other disease risk in the population isolate of the Shetland Isles in northern Scotland[190]. Genetic diversity in this population is decreased compared to mainland Scotland, consistent with the high levels of endogamy. 2105 participants were recruited between 2013 and 2015, most having at least three grandparents from Shetland. Fasting blood samples were collected and many health-related phenotypes and environmental exposures were measured in each individual. All participants gave written informed consent and the study was approved by the South East Scotland Research Ethics Committee, NHS Lothian (reference: 12/SS/0151). Details of cohort-specific demographics, genotyping, quality control, and imputation performed before GWAS can be found in Supplementary Data 17.

Phenotypic data

Transferrin and total IgG N-glycome quantification for CROATIA-Korcula and VIKING samples was performed at Genos Glycobiology Laboratory. Isolation of the protein of interest and N-glycan quantification is described in more detail in Supplementary Materials and Methods and in Trbojević-Akmačić et al.[191] for transferrin and by Trbojević-Akmačić et al.[192] for IgG. Briefly, proteins were first isolated from blood plasma (IgG depleted blood plasma in the case of transferrin) using affinity chromatography binding respectively to anti-transferrin antibodies plates for transferrin and protein G plates for IgG. The proteins isolation step was followed by enzymatic release and labelling of N-glycans with 2-AB (2-aminobenzamide) fluorescent dye. IgG N-glycans have been released from total IgG (all subclasses). N-glycans were then separated and quantified by hydrophilic interaction ultra-high-performance liquid chromatography (HILIC-UHPLC). As a result, transferrin and total IgG samples were separated into 35 (transferrin: TfGP1− TfGP35) and 24 (IgG: GP1−GP24) chromatographic peaks. It is worth noting that there is no correspondence structure-wise between transferrin TfGP and IgG GP traits labelled with the same number.

Prior to genetic analysis, raw N-glycan UHPLC data was normalised and batch corrected to reduce the experimental variation in measurements. Total area normalisation was performed by dividing the area of each chromatographic peak (35 for transferrin, 24 for IgG) by the total area of the corresponding chromatogram. Resulting measures are therefore relative abundances of each glycan structure in the overall glycosylation profile. Due to the multiplicative nature of measurement error and right-skewness of glycan data, normalised glycan measurements were $\log_{10}$-transformed. Batch correction was then performed using the empirical Bayes approach implemented in the “ComBat” function of the sva 3.34.0 R package[193], modelling the technical source of variation (96-well plate number) as batch covariate. Batch corrected measurements where then exponentiated back to the original scale.

Genome-wide association analysis

Genome-wide association analyses (GWAS) were performed in the two cohorts of European descent, CROATIA-Korcula and VIKING. Associations with 35 transferrin N-glycan traits were performed in 948 samples from CROATIA-Korcula and 959 samples from VIKING. Associations with 24 IgG N-glycan traits were performed in 951 samples from CROATIA-Korcula and 1086 samples from VIKING. The sample size of the same cohort differs between transferrin and IgG due to the different number of samples successfully measured for each protein. Prior to GWAS, each glycan trait was rank transformed to normal distribution using the "rntransform" function from the GenABEL 1.1-6 R package[194] and then adjusted for age and sex, as fixed effects, and relatedness (estimated as the kinship matrix calculated from genotyped data) as random effect in a linear mixed model, calculated using the "polygenic" function from the GenABEL R package[194]. Residuals of covariate and relatedness correction were tested for association with HRC (Haplotype Reference Consortium)[28] imputed SNP dosages using the RegScan v0.5 software[195], applying an additive genetic model of association.

Meta-analysis

Prior to meta-analysis the following quality control was performed on cohort-level GWAS summary statistics. We removed all SNPs with a difference in allele frequency between the two cohorts higher than +/- 0.3 (~37,000 SNPs in total), as well as variants showing a minor allele count (MAC) lower or equal to 6 (~6 million SNPs in total). Cohort-level GWAS were then meta-analysed (N=1890 for transferrin and N=2020 for IgG N-glycans, for ~10.7 million SNPs) using METAL v2011-03-25 software[196], applying the fixed effect inverse-variance method, followed by genomic control correction. Mean genomic control inflation factor (λ_{GC}) was 0.997 (range 0.982-1.011) for transferrin N-glycans and 0.995 (range 0.981-1.008) for IgG N-glycans meta-analysis, showing that the confounding effects of family structure were correctly accounted for. The standard genome-

wide significance threshold was Bonferroni corrected for the number of N-glycan traits analysed: variants were considered statistically significant if their p-value was lower than $5\times10^{-8}/35 = 1.43\times10^{-9}$ for transferrin and $5\times10^{-8}/24 = 2.08\times10^{-9}$ for IgG N-glycan traits.

We used a positional approach to define genomic regions (loci) significantly associated with transferrin N-glycan traits, following the procedure adopted by Sharapov et al.[133] For each glycan trait, we grouped all genetic variants located within a 500 kb window (+/- 250 kb) from the sentinel SNP in the same locus. To obtain a unique list of loci that are independent of the specific glycan trait, we then merged this list of sentinel SNP-glycan trait pairs for all 35 glycan traits and applied a similar procedure - all SNP-glycan trait pairs within a 1000 kb window (+/-500 kb from sentinel SNP) were grouped in the same locus, resulting in a unique list of sentinel SNP-top glycan trait pairs, summarising the genomic regions most strongly associated with N-glycans across all traits. A visual representation of the procedure can be seen in Supplementary Figure 10. For all sentinel SNP-top glycan trait pairs, regional association plots were created with LocusZoom[197] and visually checked - in case of overlapping patterns of association, only the sentinel SNP-top glycan trait pair showing the lowest p-value was selected as a locus representative.

Impact of transferrin protein levels on transferrin glycome associations

To assess the potential impact of transferrin protein levels on transferrin glycome associations and to check whether the associations in the region of the *TF* gene are driven by protein levels, we tested association of transferrin *cis*-pQTL rs8177240[178] with transferrin glycosylation using the likelihood ratio test implemented in the lmtest 0.9-38 R package[198] between the following models:

M0: glycan ~ age + sex

M1: glycan ~ age + sex + pQTL (rs8177240)

M2: glycan ~ age + sex + glyQTL (rs6785596)

M3: glycan ~ age + sex + pQTL (rs8177240) + glyQTL (rs6785596)

where pQTL is the SNP most strongly associated with transferrin levels and here used as proxy for the protein levels, and glyQTL is the SNP most strongly associated with transferrin glycan levels in the *TF* gene region.
The likelihood ratio tests were performed between:

- M0 and M1 to assess associations of glycans and pQTL (rs8177240)
- M1 and M3 to assess whether glyQTL contributes to glycan levels even when the pQTL is included in the model
- M2 and M3 to assess whether pQTL contributes to glycan levels even when the glyQTL is included in the model

To control for increased levels of relatedness between subjects in our studies, the models were fitted using linear mixed models as implemented in the lme4qtl 0.0.2 R package[199], with age, sex, pQTL and glyQTL as fixed effects and kinship matrix as a random effect. The kinship matrix was estimated from the genotyped data using the "ibs" function from GenABEL[194] R package.

Transferrin N-glycan traits post-meta-analysis follow-up

The meta-analysis follow-up analyses were performed only for the transferrin N-glycans meta-analysis, since genetic regulation of IgG N-glycosylation has already been explored in a larger, IgG-specific study[135] and is beyond the scope of the present work.

Conditional analysis and phenotypic variance explained

To capture the overall contribution to phenotypic variation at each genomic region and identify secondary association signals at a locus, we performed approximate conditional analysis using the GCTA-COJO[177] 1.91.4beta stepwise model selection, "cojo-slct", with the IgG and transferrin N-glycan meta-analysis summary statistics and genotypes of 10,000 unrelated individuals of white British ancestry from UK Biobank[200] as independent LD reference panel. Collinearity was restricted to 0.9 and the p-value threshold was set to 1.43×10^{-9} for transferrin

and to 2.08×10^{-9} for IgG. Reported joint p-values were then adjusted by the genomic control method[201]. The list of samples for the independent LD reference panel was created with R 3.6.0, while the panel itself was generated using Plink 2.0[202]. After sample extraction from the UK Biobank full dataset, SNP deduplication was performed both by position (removing all SNPs not carrying a unique position on the chromosome) and marker name (--rm-dup exclude-all function). We acknowledge that UK Biobank might not be a perfect reference population for the CROATIA-Korcula cohort, however there are no other reference panels with suitable ancestry and sample size (>4,000)[177]. The proportion of variance (var) in phenotype (*Y*) explained by independently associated SNPs at each transferrin N-glycans associated locus was calculated with the following formula

$$\mathrm{var}(Y) = \frac{2 * \mathrm{freq} * (1 - \mathrm{freq}) * \beta^2}{\mathrm{var}(Y \text{ covariates adjusted residuals})} \quad (1)$$

where freq represents the frequency of the SNP's effect allele, β is the effect estimate for the SNP and phenotype association at the locus, *Y* covariates adjusted residuals are the residuals resulting from the adjustment of the phenotype by age and sex, as fixed effects, and relatedness (estimated as the kinship matrix calculated from genotyped data) as random effect in a linear mixed model. The "polygenic" function from the GenABEL R package was used also to estimate cohort-specific heritability for each transferrin glycan trait.

Gene prioritisation

For all genome-wide significant loci we suggested plausible candidate genes combining different evidence, namely evaluating biological role in the context of protein N-glycosylation of genes nearest to sentinel variants (positional mapping), assessing colocalisation with gene expression (expression quantitative trait loci, eQTL) or investigating associated variant's predicted effects on the protein

sequence or on putative transcription factor binding sites. Positional gene mapping was performed using FUMA v1.3.5e SNP2GENE function[203]. Genes having a clear biological link to protein N-glycosylation (e.g. genes coding for enzymes involved in biochemical pathway of protein glycosylation) and genes previously associated with IgG and/or total blood plasma proteins N-glycome were given a priority. The overlap of independent significant SNPs identified by COJO with eQTL was investigated using PhenoScanner v1.1 database[183], taking into account significant genetic association (p-value < 5×10^{-8}) at the same or strongly (LD r^2 > 0.8) linked SNPs in populations of European ancestry. The Ensembl Variant Effect Predictor (VEP v 97) tool[180] was used to determine putative functional effect and impact on a transcript or protein of independent significant SNPs and their strongly (LD r^2 > 0.8) linked SNPs in populations of European ancestry. Among genes prioritised so far, two were transcription factors (i.e. *HNF1A* and *FOXI1*), while the remaining were non transcription factor protein-coding genes (i.e. *MGAT5*, *TF*, *MSR1*, *NXPE1/NXPE4*, *ST3GAL4*, *B3GAT1*, *FUT8* and *FUT6*). Using the Regulatory sequence analysis tools (RSAT) v2018-08-04 program *matrix-scan*[182], we applied a pattern-matching procedure to search for sequences recognized as binding sites for *HNF1A* and *FOXI1* transcription factors in associated regions of the other 8 prioritised genes. Position-specific scoring matrices (PSSMs), representing the frequency of each nucleotide at each position of the transcription factor motif, were downloaded for *HNF1A* and *FOXI1* from the JASPAR[204] database. For each of the 8 genomic regions explored for possible transcription factor binding sites, we included the most strongly associated SNP and a 60 bp surrounding sequence (30 bp either side of the sentinel SNP). The significance threshold was set to the p-value ≤ 0.003, Bonferroni corrected for 16 tests performed (8 putative transcription factor binding sites tested for 2 transcription factors).

Overlap and colocalization analysis with gene expression levels and complex traits

The PhenoScanner v1.1 database[183] was used to investigate the overlap of significant transferrin glycosylation SNPs with gene expression levels and complex human traits. As previously described, we considered traits with genome-wide significant association (p-value < 5×10^{-8}) at the same or strongly (LD $r^2 > 0.8$) linked SNPs in populations of European ancestry. We then used Summary data-based Mendelian Randomization (SMR) analysis followed by the Heterogeneity in Dependent Instruments (HEIDI) test[179] to assess whether overlapping expression and complex traits, identified by PhenoScanner, were also colocalising with transferrin glycosylation (TfGP) traits. The SMR test indicates whether two traits are associated with the same locus, and HEIDI test specifies whether both traits are affected by the same underlying functional SNP. Each of 10 sentinel SNPs – TfGP pair (Table 1) was used for SMR/HEIDI analysis with gene expression levels and several complex traits. Summary statistics for gene expression levels in tissues/cell types were obtained from the Blood eQTL study[205] (http://cnsgenomics.com/software/smr/#eQTLsummarydata), the CEDAR project[206] (http://cedar-web.giga.ulg.ac.be/), and the GTEx project version 7[207] (https://gtexportal.org/home/). Summary statistics for complex traits were obtained from various resources. In total, we used data for 3 tissues/cell types: CD19+ B lymphocytes (CEDAR), GTEx liver (GTEx) and peripheral blood (the Blood eQTL study) and 8 complex traits. Full list of GWAS collections, tissues and complex traits see in Supplementary Data 18. SMR/HEIDI analysis was performed according to the protocol described by Zhu et al.[179] We used sets of SNPs having the following properties: (1) being located within ± 250 kb from the sentinel SNPs identified in the present study; (2) being present in both the primary GWAS and eQTL data/GWAS for the complex trait; (3) having MAF ≥ 0.03 in both datasets; (4) having squared Z-test value ≥ 10 in the primary GWAS. Those SNPs that met criteria (1), (2), (3), (4), had the lowest p-value in the primary GWAS and were in high LD ($r2 > 0.8$) with the sentinel SNPs were used as instrumental variables to elucidate the relationship between gene expression/disease and TfGP (we define them as “top SNPs”). It should be noted that SMR/HEIDI analysis

does not identify a causative SNP affecting both traits. It can be either the top SNP or any other SNP in strong LD. After defining the set of eligible SNPs for each locus, we made the “target” and “rejected” SNP sets and added the top SNP to the “target” set. Then we performed the following iterative procedure of SNP filtration: if the SNP from the eligible SNP set with the lowest PSMR had r2 > 0.9 with any SNP in the “target” SNP set, it was added to the “rejected” set; otherwise, it was added to the “target” set. The procedure was repeated until eligible SNP set was exhausted, or the “target” set had 20 SNPs. If we were unable to select three or more SNPs, the HEIDI test was not conducted. HEIDI statistics was calculated as

$$T_{HEIDI} = \sum_{i}^{m} z_{d(i)}^{2}, \quad (2)$$

where m is the number of SNPs selected for analysis, $z_{d(i)} = {d_i}/{SE_{(d_i)}}$ and $d_i = \beta_{SMR_i} - \beta_{SMR\ (top\ SNP)}$.

The results of the SMR test were considered statistically significant if PSMR < 1.7×10^{-4} (0.05/302, where 302 is a total number of tests corresponding to analysed loci and gene expression/disease traits). Inference of whether a functional variant may be shared between the TfGP and gene expression/disease were made based on the HEIDI test: $P_{HEIDI} \geq 0.001$ (possibly shared), and $P_{HEIDI} < 0.001$ (sharing is unlikely).

We then proceeded to further explore SMR-HEIDI significant findings using bi-directional Mendelian Randomisation (MR), as implemented in the TwoSampleMR 0.5.6 R package[208]. MR uses genetic variants as instrumental variables to investigate the effects of one trait (exposure) on another trait (outcome), assuming that the instrumental variables associate with the outcome only through the exposure. GWAS summary statistics for complex traits were obtained from the IEU GWAS database[209] and their references are listed in Supplementary Data 18. For each glycan and complex trait, we selected as instruments for the exposure genetic variants associated with the trait at genome-

wide significance (p-value < 5×10^{-8}) and independent (r^2 = 0.001, using the European population from the 1000 Genomes Project reference panel). To distinguish causal relationships from confounding by LD, we followed-up significant MR tests (p-value $\leq$ 0.05/8 = 6.25×10^{-3}, Bonferroni corrected for the number of tests) with approximate Bayes factor colocalisation analysis, developed by Giambartolomei et al.[106] and implemented in the "coloc.abf" function from the coloc 4.0-6 R package, using default priors of 10^{-4} for prior probability of SNP being associated with trait 1 or trait 2 (p1 and p2) and 10^{-5} for prior probability of a SNP being associated with both traits (p12). To further assess the robustness of our findings, where available, we performed the "coloc" analysis using a different complex-trait GWAS dataset compared to the SMR-HEIDI analysis (listed in Supplementary Data 18). Colocalisation analysis tests whether local genetic association signals for different traits are driven by the same shared causal variant or distinct variants. This Bayesian method provides posterior probabilities (PP) for 5 different hypotheses: the null hypothesis of no association with either of the traits (H0) and four alternative hypotheses of either association with only the first or the second of the traits (H1, H2), or association of both traits via distinct underlying causal variants (H3), or association of both traits through a shared causal variant (H4) i.e. trait colocalisation. A posterior probability >80% was considered as robust evidence supporting the tested hypothesis.

Colocalisation analysis for transferrin and IgG N-glycan traits

The *FUT8* and *FUT6* genomic regions were significantly associated with both transferrin and IgG N-glycans. To investigate a possible overlap in genetic control of glycosylation between the two proteins, we used the approximate Bayes factor colocalisation analysis implemented in coloc R package[106], followed by pairwise conditional and colocalization analysis (PwCoCo)[100] in case of multiple independent variants contributing to the trait variation. A posterior probability > 80% was considered as robust evidence supporting the tested hypothesis.

Overview of the overall procedure can be seen in Supplementary Figure 6. First, we assessed whether for one protein all glycans that are associated with the same genomic region (p-value $\leq 5x10^{-8}$) are regulated by the same underlying variants. For each protein (i.e. transferrin and IgG) and each genomic region (i.e. *FUT8* and *FUT6*), we tested separately the group of glycans carrying only one independent association signal at locus and the group of glycan traits showing multiple independent signals of association (Supplementary Figure 6). Pairs of glycan traits obtaining a PP.H4 > 80% (suggestive of colocalisation) were pooled in the same colocalisation group, following the principle that if trait A colocalises with trait B and trait B colocalises with trait C, thus also trait A and trait C colocalise. For each within-protein colocalisation group identified, the glycan trait with the lowest p-value was selected as group representative and carried on to the next step, where traits with single and multiple independent associations for each protein were tested for colocalisation. Similar to previous steps, glycan traits were grouped together on the basis of their colocalisation analysis results and the lowest p-value representative was chosen for the next step, where finally representative transferrin and IgG glycans were tested for between-protein colocalisation.

For glycan traits with multiple independent association signals and lacking strong evidence for colocalisation, we applied PwCoCo[100] approach. Briefly, the PwCoCo approach tests not only the traits' full, complete GWAS association statistics for colocalisation, but also summary statistics conditioned for the top primary association, testing whether any of the underlying causal variants between traits colocalise. For example, assuming that each trait is carrying two conditionally independent association signals in the tested region, colocalisation analysis will be conducted between both full and conditioned association statistics (conditioned for each independent variable), for a total of nine pairwise combinations. Secondary association signals at *FUT8* and *FUT6* loci for both transferrin and IgG N-glycans were assessed using GCTA-COJO approximate conditional analysis stepwise model selection[177] and an LD reference panel of 10,000 unrelated, white British ancestry individuals from UK Biobank[200]. We then

performed the association analysis conditional on identified secondary association signals at *FUT8* and *FUT6* loci using GCTA-COJO[177] "cojo-cond" and the same 10,000 UK Biobank samples LD reference panel, with 5×10^{-8} p-value threshold and used those for pairwise colocalisation analyses.

Expression of N-glycome associated genes in transferrin and IgG-relevant tissues

Gene expression data for *TF, IGHG1, HNF1A* and *IKZF1*, expressed in gene counts, for hepatocytes (529 samples) and plasma cells (648 samples) was obtained from ARCHS4 portal[187]. Samples with total number of gene counts less than 5,000,000 were filtered out, leaving 513 hepatocyte and 53 plasma cell samples for the analysis. Gene counts were scaled to transcripts per million (TPM) and $\log_2(1+TPM)$ transformed.

Data availability

The full summary statistics from the GWAS of 35 transferrin glycan traits and 24 IgG glycan traits generated in this study have been deposited in the DataShare repository (https://datashare.ed.ac.uk/handle/10283/4088). There is neither Research Ethics Committee approval, nor consent from individual participants, to permit open release of the individual level research data underlying this study. The datasets analysed during the current study are therefore not publicly available. Instead, the research data and/or DNA samples are available from accessQTL@ed.ac.uk on reasonable request, following approval by the QTL Data Access Committee and in line with the consent given by participants. Each approved project is subject to a data or materials transfer agreement (D/MTA) or commercial contract. The UK Biobank genotypic data used in this study were approved under application 19655 and are available to qualified researchers via the UK Biobank data access process (http://www.ukbiobank.ac.uk/register-

apply/). The position-specific scoring matrices (PSSMs) for *HNF1A* and *FOXI1* genes used in this study are available in the JASPAR[204] database under the accession code MA0046.2 (http://jaspar.genereg.net/api/v1/matrix/MA0046.2/?format=transfac) and MA0042.1 (http://jaspar.genereg.net/api/v1/matrix/MA0042.1/?format=transfac), respectively. The summary statistics for gene expression levels in tissues/cell types used in this study are available in the Blood eQTL study (http://cnsgenomics.com/software/smr/#eQTLsummarydata), in the CEDAR project (http://cedar-web.giga.ulg.ac.be/), in the GTEx project version 7 (https://gtexportal.org/home/) and in the eQTLGen consortium (https://www.eqtlgen.org/). The summary statistics for complex traits are available in various publicly available resources, as detailed in Supplementary Data 18.

Code availability

The following software packages were used in this study: Ensembl variant effect predictor (VEP):
https://www.ensembl.org/info/docs/tools/vep/script/vep_options.html;
phenoscanner: https://github.com/phenoscanner/phenoscanner;
GCTA-COJO: https://yanglab.westlake.edu.cn/software/gcta/#COJO;
coloc: https://github.com/chr1swallace/coloc;
TwoSampleMR: https://mrcieu.github.io/TwoSampleMR/.
The remaining code used in this paper may be requested from the authors.

Acknowledgements

We thank Dr Nicola Pirastu for sharing his knowledge and experience with Mendelian Randomisation and Jelena Šimunović for her technical assistance in the laboratory work. The CROATIA-Korcula study was funded by grants from the MRC (United Kingdom), European Commission Framework 6 project EUROSPAN (contract number LSHG-CT-2006-018947), Croatian Science Foundation (grant 8875), and the Republic of Croatia Ministry of Science, Education and Sports (216-1080315-0302). Genotyping was performed in the Genetics Core of the Clinical Research Facility, University of Edinburgh. We would like to acknowledge all the staff of several institutions in Croatia that supported the CROATIA-Korcula fieldwork, including, but not limited to, the University of Split and Zagreb Medical Schools, Institute for Anthropological Research in Zagreb, and the Croatian Institute for Public Health in Split. The Viking Health Study – Shetland (VIKING) was supported by the MRC Human Genetics Unit quinquennial programme grant "QTL in Health and Disease". DNA extractions and genotyping were performed at the Edinburgh Clinical Research Facility, University of Edinburgh. We would like to acknowledge the invaluable contributions of the research nurses in Shetland, the administrative team in Edinburgh and the people of Shetland. Finally, we thank the UK Biobank Resource, approved under application 19655. We acknowledge support from the European Union's Horizon 2020 research and innovation program IMforFUTURE (A.L.: H2020-MSCA-ITN/721815); the RCUK Innovation Fellowship from the National Productivity Investment Fund (L.K.: MR/R026408/1); the Russian Science Foundation (RSF) (Y.S.A. and S.Z.S.: 19-15-00115); and the MRC Human Genetics Unit programme grant, "QTL in Health and Disease" (J.F.W. and C.H.: MC_UU_00007/10).

Author contributions

A.L.: Data analysis and interpretation, Visualization, Writing—Original draft preparation, Writing—Review and editing. I.T.-A.: Quantification of transferrin and IgG N-glycans, Data interpretation, Writing—Original draft preparation, Writing—Review and editing. P.N.: Supervision, Data interpretation, Writing—Review and editing. Y.T.: Data analysis and interpretation, Writing—Original draft preparation. S.Z.S.: Visualization, Writing—Original draft preparation. F.V.: Glycan data quality control. O.P.: Genomic and demographic data provider for CROATIA-Korcula cohort. C.H.: Genomic and demographic data provider for CROATIA-Korcula cohort. T.P.: Quantification of transferrin and IgG N-glycans. M.V.: Quantification of transferrin and IgG N-glycans. Y.S.A.: Writing—Review and editing. G.L.: Conceptualisation, Glycan data provider for CROATIA-Korcula and VIKING cohorts, Writing—Review and editing. J.F.W.: Conceptualisation, Genomic and demographic data provider for VIKING cohort, Supervision, Data interpretation, Writing—Original draft preparation, Writing—Review and editing. L.K.: Conceptualisation, Supervision, Data interpretation, Writing—Original draft preparation, Writing—Review and editing.

Competing interests

G.L. is the founder and owner of Genos Ltd, a private research organization that specializes in high-throughput glycomic analysis and has several patents in this field. I.T.-A., F.V., T.P., and M.V. are employees of Genos Ltd. Y.S.A. is a founder and a co-owner of PolyOmica and PolyKnomics, private organizations providing services, research and development in the field of computational and statistical genomics. The remaining authors declare no competing interests.

2.3 Conclusion

I conducted the first GWAS of UHPLC-measured transferrin glycome (35 glycan traits, N = 1890), and identified ten significantly associated loci, two of which (near *FOXI1* and *MSR1* genes) had not previously been associated with the glycosylation of any protein.

By comparing the glycan-associated loci of transferrin with those of IgG (24 glycan traits, N = 2020), I was able to describe for the first time similarities and differences in the genes and variants contributing to the glycome of these two proteins. When focusing on loci containing genes encoding glycosyltransferase, enzymes that play a key role in the glycosylation process, we found both associations specific to each protein (transferrin - *MGAT5, ST3GAL4* and *B3GAT1*; IgG - *ST6GAL1* and *MGAT3*) and associations shared by the two proteins (*FUT8* and *FUT6*). For the shared associations, we showed that glycosylation of transferrin and IgG is regulated by independent, protein-specific variants in *FUT8* and *FUT6* genes. In the *FUT8* region, these independent variants likely regulate the glycosylation of transferrin and IgG in a tissue-specific manner, acting through tissue-specific transcription factors HNF1A and IKZF1.

While some low frequency associations were identified (MAF = 0.047 for rs6785596 - *TF*; MAF = 0.018 for rs115399307 - *FOXI1*; MAF = 0.027 for rs41341748 - *MSR1*), this chapter mainly described common variation associated with the transferrin glycome. Similarly, while a stop lost (rs41341748 - *MSR1*) and several missense variants (rs115399307 – *FOXI1*, rs550897 - *NXPE4* and rs17855739 - *FUT6*) (Supplementary Data 9) were identified and described in this chapter, the majority of transferrin glycans-associated variants (>60%) were classified as intronic by VEP[180].In the next chapter, I expand our current knowledge on the genetic architecture of the protein glycome by focusing on the contribution of low frequency and rare variants having an impact on the gene coded product.

Chapter 3: Rare and low frequency variants contributing to variation in the protein glycome

3.1 Background

GWAS studies have successfully identified numerous genetic loci associated with complex traits and diseases, but as sample sizes have increased, the effects of newly identified variants on trait measurements or disease risk have become smaller[33]. This "omnigenic model" suggests many common variant associations emerging from GWAS may not provide a mechanistic understanding of the studied trait[210]. By contrast, rare-variant studies usually detect larger-effect variants that can implicate genes whose function is core to the studied trait, providing more direct insights into its biology. The contribution of rare variants to several human traits is well established, with many disorders being explained by individual, highly penetrant alleles[35,211]. Recent rare-variant studies, aided by advances in statistical methods and improved accessibility to sequencing data, have not only pinpointed genes with significant association with diseases but have also showed that rare variants play an important role in the genetic architecture of complex traits and diseases[212–214].

All protein glycome studies so far, including the one reported in Chapter 2, have used single-variant GWAS tests to investigate the genetic regulation of glycosylation, succeeding in identifying mainly common associated variants, located in non-coding regions of the genome[131–138,215]. Therefore, the contribution of low frequency and rare variants (MAF < 5%) on glycan variation, and their possible impact on human health, has not been thoroughly explored. In this chapter, I employ multiple gene-based aggregation tests (i.e. burden test, SKAT and omnibus tests such as SKAT-O) to examine the effect of rare, predicted loss-of-function (pLoF) and missense variants from whole exome sequencing (WES) on transferrin and IgG glycan traits. Details about the number of transferrin and IgG glycan traits analysed, the different cohorts used, and their sample size are

reported in Supplementary Figure 1. In addition to the directly measured glycan structures, (35 for transferrin and 24 for IgG) which were analysed in Chapter 2, in this Chapter I also include several glycan derived traits. These derived traits (16 for transferrin as described in Supplementary Table 15 of this chapter, 54 as defined by Huffman et al.[141] plus 16 as detailed in the Supplementary Table 16 of this Chapter for IgG) represent common biologically meaningful features shared among several measured glycans or the overall presence of a certain sugar structure on the totality of glycan traits measured (e.g. percentage of fucosylated glycans, triantennary glycans, monogalactosylated glycans, etc.). For transferrin glycan traits, the individuals studied in this Chapter are the same analysed in Chapter 2 (VIKING N=952, batch 2 of CROATIA-Korcula N=938). For IgG glycans instead, additional cohort/sample size are added (ORCADES N=1959, batch 1 of CROATIA-Korcula N=849) to the individuals assayed in Chapter 2 (VIKING N=1079, batch 2 of CROATIA-Korcula N=941).

In addition to increasing statistical power by testing the cumulative effects of multiple rare variants in genetic regions, I also benefited from the favourable features of genetically isolated populations - the increase in frequency of some otherwise rare variants, due to high genetic drift in these populations, resulting in increased power to identify rare variant associations. Further, WES studies allow for the identification of rare variants with a large impact on the coded product, which often have a more direct role in biological mechanisms. I also investigated the potential impact of rare variants associated with glycosylation on health-related traits. Chapter 2 expanded the current knowledge of genetic architecture underlying human glycomic variation by identifying genetic variants contributing to glycosylation of transferrin, a previously unstudied protein in this sense. In this chapter I explore the effect of rare genetic variation on protein glycosylation.

3.2 Manuscript pre-print

The following chapter is based on a manuscript submitted to a preprint server. A copy of the manuscript (available at https://www.medrxiv.org/content/10.1101/2022.12.26.22283911v1) is included below, with permission from the co-authors.

Exome sequencing reveals aggregates of rare variants in glycosyltransferase and other genes influencing immunoglobulin G and transferrin glycosylation

Arianna Landini[1,2], Paul R.H.J. Timmers[1,2], Azra Frkatović-Hodžić[3], Irena Trbojević-Akmačić[3], Frano Vučković[3], Tea Pribić, Regeneron Genetics Center[4], Gannie Tzoneva[4], Alan R. Shuldiner[4], Ozren Polašek[5,6], Caroline Hayward[1], Gordan Lauc[3,7], James F. Wilson*[1,2] & Lucija Klarić*[1]

1 MRC Human Genetics Unit, Institute for Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom
2 Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
3 Genos Glycoscience Research Laboratory, Zagreb, Croatia
4 Regeneron Genetics Center, Tarrytown, NY, USA
5 Department of Public Health, School of Medicine, University of Split, Split, Croatia
6 Algebra University College, Zagreb, Croatia
7 Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb, Croatia
* Authors contributed equally.
Correspondence to: J.F.W ( ) or L.K. ( )

In this study, I conducted gene-based aggregation analysis, GWAS and ExWAS of the transferrin and IgG glycome, and gene-based aggregation analysis of health-related traits. I used pre-processed glycan data and formulas for calculating derived glycan traits developed by Irena Trbojević-Akmačić and Azra Frkatović-Hodžić, as well as clean exome sequence data provided by Regeneron and a gene-based aggregation analysis pipeline created by Paul R.H.J. Timmers.

I wrote the first draft of the manuscript, with significant contributions from Lucija Klarić on the interpretation of rare variant associations independent of ExWAS signals and link between transferrin glycosylation and human health.

Introduction

Genome-wide association studies (GWAS) have so far identified thousands of loci associated with human complex traits and diseases. However, the large majority of these variants are found in noncoding regions of the genome[216], posing a challenge when attempting to uncover their functional impact on the phenotype. On the contrary, whole-exome sequencing (WES) studies offer the opportunity to identify rare variants of larger effect in the encoded protein, such as predicted loss of function (pLoF) and missense variants, for which causal biological mechanisms are generally easier to elucidate[217]. Methods for exome-wide rare variant analysis have been successfully employed to discover variants and genes associated with both complex molecular traits[157] and diseases[12,36]. While single-variant tests, such as GWAS, are largely adopted to explore associations of common genetic variants with phenotypes of interest, they have little power to identify rare variant associations, due to the low number of observations. Therefore, a set of methods testing cumulative effects of multiple rare variants in genetic regions, where rare variants are grouped at the gene level (also known as 'masks') via a collapsing test, such as burden tests, or variance-component tests (e.g. sequence kernel association test, SKAT[218]) were developed. In addition to increasing the statistical power by aggregating multiple rare-variants, using genetically isolated populations can provide unique opportunities for novel discovery in an association study[219]. Recent bottlenecks, restricted immigration and limited population size lead to increased genetic drift. Consequently, in such populations some otherwise rare variants can substantially increase in frequency compared to the general population, therefore increasing association power for these variants.

Glycosylation is one of the most common post-translational modifications, where sugar residues, called glycans, are attached to the surface of proteins. Changes in protein N-glycosylation patterns have been described in the ageing process[220,221] and in a wide variety of complex diseases, including autoimmune diseases[222], diabetes[123], cardiovascular diseases[223], neurodegenerative diseases[224] and cancer[225]. Despite glycans having an important role in human

health and serving as potential biomarkers in clinical prognosis and diagnosis[129], we have just started scratching the surface of the complex network of genes regulating protein glycosylation. All studies published to date exploring the genetic regulation of total plasma protein, immunoglobulin G (IgG) and transferrin N-glycosylation have employed single variant-based GWAS tests, mostly uncovering common variants located in non-coding regions of the genome[131–133,135–138,215]. Rare variants contributing to glycan variation, and their impact on human health, thus remain unexplored.

To address this knowledge gap, we used multiple gene-based aggregation tests to investigate how rare (MAF<5%) pLoF and missense variants from whole exome sequencing affect 51 transferrin (N = 1907) and 94 IgG (N = 4912) glycan traits in European-descent cohorts. IgG is both the most abundant antibody and one of the most abundant proteins in human serum. It contains evolutionary conserved N-glycosylation sites in the constant region of each of its heavy chains, occupied by biantennary, largely core-fucosylated and partially truncated glycan structures, that may carry a bisecting N-acetylglucosamine and sialic acid residues[226,227]. Transferrin is a blood plasma glycoprotein that binds iron (Fe) and consequently mediates its transport through blood plasma. Human transferrin has two N-glycosylation sites, with biantennary disialylated digalactosylated glycan structure without fucose being the most abundant glycan attached[173,174].

In this study, we used gene-based aggregation of rare variants to identify several genes associated with transferrin and IgG glycosylation traits. Significant genes include known protein glycosylation genes as well as novel genes with no previously known role in post-translational modification. Importantly, several associations would not have been detectable by single-point analysis and one association was detected as result of enrichment of rare variants in population isolates. Finally, we highlight the impact of rare variation in these genes on health-related traits by performing gene-based aggregation tests of 116 health-related traits together with gene lookups in public repositories of gene-based association tests.

Results

Exome variant annotation

To assess the effect of rare genetic variants on glycosylation of two proteins, we sequenced the exomes of 4,801 participants of European ancestry. After quality control, a total of 233,820 distinct autosomal coding genetic variants were available in the ORCADES cohort (N=2090), 244,649 in the VIKING cohort (N=2106) and 340,203 in the CROATIA-Korcula cohort (N=2872). Percentages of variants for each effect category in the total sequenced coding variation are similar across the three cohorts (Table 1). More than half (~53%) of the sequenced coding variants are missense variants, of which nearly half (~28% of total coding variation) are classified as likely or possibly deleterious by multiple variant effect predictor algorithms (see Methods). The second most represented effect category is synonymous mutations (~33%), followed by variants in splice regions (~8%), predicted loss of function (pLoF) (~4%) and in-frame insertions/deletions (~1.5%). Around one quarter of coding variants in the ORCADES and VIKING cohorts are singletons (minor allele count, MAC=1); this percentage is instead higher in CROATIA-Korcula cohort (~35%), possibly due to the larger sequenced sample size.

Table 1. Number of coding exome variants sequenced in the complete sample of 3 isolated cohorts. Counts and prevalence of autosomal variants observed in WES-targeted regions across all individuals in the ORCADES, CROATIA-Korcula and VIKING cohort, by type or functional class for all and for singleton variants (MAC= 1).

	ORCADES (N=2090)			CROATIA-Korcula (N=2872)			VIKING (N=2106)		
Variant category	**No. of variants**	**% of total coding variants**	**Variants % with MAC=1**	**No. of variants**	**% of total coding variants**	**Variants % with MAC=1**	**No. of variants**	**% of total coding variants**	**Variants % with MAC=1**
coding variants	233,820		25.1%	340,203		35.5%	244,649		28.9%
pLOF	8639	3.69%	37.1%	12,970	3.81%	47.2%	9025	3.69%	41.4%
Splice acceptor	872	0.37%	37.8%	1309	0.38%	45.5%	945	0.39%	42.7%
Splice donor	1042	0.45%	37.5%	1506	0.44%	48.1%	1079	0.44%	40.4%
Stop gained	2833	1.21%	36.8%	4171	1.23%	47%	2879	1.18%	41.8%
Frameshift	3401	1.45%	37.8%	5274	1.55%	47.9%	3583	1.46%	42.1%
Stop lost	151	0.06%	32.5%	244	0.07%	43%	182	0.07%	39%
Start lost	340	0.15%	30%	466	0.14%	44.2%	357	0.15%	30.3%
Missense	124,416	53.2%	27%	183,056	53.8%	37.6%	130,299	53.3%	30.6%
Likely benign (0-1)	56,366	24.1%	21.8%	80,777	23.7%	31.6%	59,141	24.2%	25.4%
Possibly deleterious (2-3)	31,693	13.5%	28.1%	47,235	13.9%	39.1%	33,138	13.5%	31.6%
Likely deleterious (4-5)	35,728	15.3%	34.2%	54,187	15.9%	45.4%	37,384	15.3%	38.2%
Unclassified missense	629	0.27%	23%	857	0.25%	29.3%	636	0.26%	23.1%
Splice region	18,580	7.95%	24.7%	26660	7.84%	32.4%	19,297	7.89%	27.8%
In-frame indel	3383	1.45%	21.5%	5244	1.54%	29.5%	3606	1.47%	26%
Protein altering	3	0%	33.3%	4	0%	25%	2	0%	0%
Stop retained	1	0%	0%	2	0%	50%	1	0%	25.2%
Synonymous	78,798	33.7%	21.1%	112,267	33%	31.6%	82,419	33.7%	100%

Exome-wide aggregated rare variant analysis of transferrin and IgG glycomes

We performed exome-wide gene-based tests across 51 transferrin traits (glycome subset of CROATIA-Korcula N = 948, VIKING N = 959) and 94 IgG glycan traits (glycome subset of ORCADES N = 1960, CROATIA-Korcula N = 1866, VIKING N = 1086), testing low frequency and rare (MAF <5%) pLoF and missense variants. In total, we identified 16 significant associations for transferrin- (Supplementary Table 1) and 32 significant associations for IgG- (Supplementary table 2) glycan traits, at Bonferroni-corrected p-values of 8.06×10^{-8} and 1.19×10^{-7}, respectively (Figure 1, Table 2). Most gene-aggregated rare variants were associated with protein-specific glycans (transferrin: variants in *FUT6, TIRAP, MSR1* and *FOXI1* genes, IgG: variants in *MGAT3, ST6GAL1* and *RFXAP* genes); only *FUT8* was associated with glycans from both proteins (Table 2, Supplementary Tables 1 and 2). Almost all identified genes encode key enzymes in protein glycosylation (*MGAT3, ST6GAL1, FUT6, FUT8*) or have been previously associated with transferrin and IgG glycan traits in GWAS analysis (*MSR1, FOXI1*)[135,215]. The exceptions are *TIRAP* and *RFXAP*, which have no previously known link to protein glycosylation. We successfully replicated (p-value < 3.2×10^{-4} for transferrin, p-value < 5.9×10^{-4} for IgG) associations of glycans with low-frequency and rare variants from 4 genes - *FUT6* and *TIRAP* with transferrin glycans, and *FUT8* and *MGAT3* for IgG glycans (Table 2) - as frequencies of variants in these genes are similar across the studied cohorts (Supplementary Table 3). While the associations of IgG glycans and variants from *FUT8* replicated, the association of transferrin glycans with variants from the same gene did not reach the significance threshold for replication (p-value in VIKING = 1.7×10^{-3}), likely because of the 7-fold decreased frequency of the rs2229678 variant in the VIKING (MAF = 0.0056) compared to CROATIA-Korcula (MAF = 0.049) cohort (Supplementary Table 4). However, given the known biological role of *FUT8* in protein glycosylation as a fucosyltransferase (one of the enzymes involved in the synthesis of glycans), we believe this association to be real. Associations of rare variants from the CROATIA-Korcula cohort in the *MSR1* gene with transferrin glycosylation also did not formally replicate in the VIKING

cohort (p-value = $8.6x10^{-4}$) (Table 2). However, the cumulative allele count of rare variants in this gene is different between CROATIA-Korcula (MAC=46) and the VIKING cohort (MAC=38) (Supplementary Table 3), decreasing the power to replicate. We also detected a couple of isolate-specific associations that are driven by variants increased in frequency compared to publicly accessible biobanks and variant repositories. Namely, the rs750567016 variant in *ST6GAL1* that affects IgG glycosylation is more than 300 times more common in ORCADES (MAF = $3.3x10^{-3}$) than in UK Biobank (MAF = $1.0x10^{-5}$) or gnomAD (MAF = $9.0x10^{-6}$) and is absent from CROATIA-Korcula and VIKING cohorts. The rs115399307 variant in *FOXI1,* associated with transferrin glycosylation, is seven times more common in VIKING (MAF = $2.1x10^{-2}$) than in CROATIA-Korcula cohort (MAF = $2.7x10^{-3}$), UK Biobank (MAF = $8.5x10^{-3}$) and gnomAD (MAF = $7.1x10^{-3}$) (Supplementary Table 4). While the role of sialyltransferase *ST6GAL1* in IgG glycosylation is well described, the roles of the transcription factor *FOXI1* and the regulatory factor X-associated protein *RFXAP* still need to be confirmed and investigated.

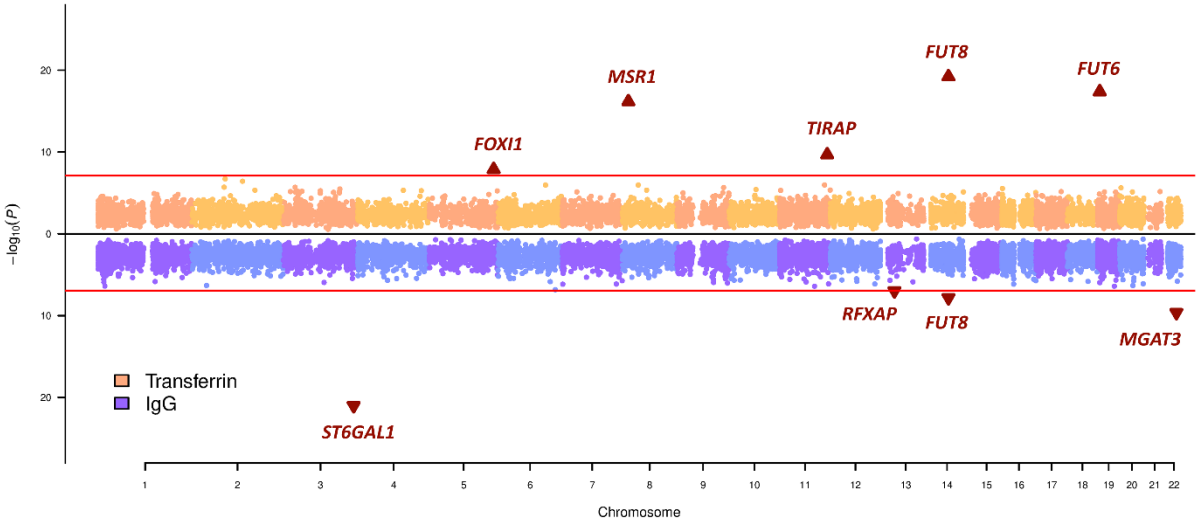


Figure 1. Miami plot summarising the results from exome-wide gene-based tests for transferrin and IgG glycan traits. Genomic positions of the genes, calculated as the mean position of variants included in the reported mask, are labelled on the x-axis and the $-\log_{10}$ of the p-value for each rare-variants aggregating test on the y-axis. For each gene-glycan association, the lowest p-value across multiple masks, multiple variant aggregate tests and cohorts was selected for plotting. The Bonferroni-corrected significance threshold for transferrin glycan traits (horizontal red line in the top part of the plot) corresponds to 8.06×10^{-8}, while Bonferroni-corrected threshold for the IgG glycan traits (horizontal red line in the bottom part of the plot) corresponds to 1.19×10^{-7}. Genes significantly associated with transferrin/IgG glycan traits are indicated with a triangle and labelled, while genes not passing the significance threshold are indicated with dots.

Table 2: Gene-based rare variants associations of transferrin and IgG glycosylation.

Lead glycan	Gene	MAF	Variants	N variants	Discovery cohort	Discovery P	Assoc. test	Discovery MAF	Discovery AC	Repl. cohort	Repl. P	Repl. MAF	Repl. AC	No. of glycans
Transferrin														
TfGP20	*FUT8*	<0.05	pLoF and deleterious (1/5)*	6	CROATIA-Korcula	$6.29x10^{-20}$	Burden	0.0111	124	VIKING	$1.73x10^{-3}$	0.0042	8	3
TfGP32	*FUT6*	<0.05	pLoF and deleterious (1/5)*	5	CROATIA-Korcula	$4.31x10^{-18}$	SKAT	0.0097	90	VIKING	$1.56x10^{-14}$	0.0072	96	8
TfGP35	*MSR1*	<0.05	pLoF	3	CROATIA-Korcula	$6.93x10^{-17}$	Burden	0.0083	46	VIKING	$8.64x10^{-4}$	0.01	38	2
TfGP17	*TIRAP*	<0.05	pLoF and deleterious (1/5)*	3	VIKING	$2.17x10^{-10}$	SKAT-O	0.0077	44	CROATIA-Korcula	$8.12x10^{-9}$	0.0076	98	2
TfGP23	*FOXI1*	<0.05	pLoF and missense	3	VIKING	$1.37x10^{-8}$	SKAT-O	0.0074	42	CROATIA-Korcula	$1.56x10^{-2}$	0.0014	8	1
IgG														
FG2S1/ (FG2+FG2S1+FG2S2)	*ST6GAL1*	<0.01	pLoF and missense	2	ORCADES	$9.82x10^{-22}$	Burden	0.0019	15	-	-	-	-	9
Fn/(Bn+FBn)	*MGAT3*	<0.01	pLoF and deleterious (1/5)*	4	ORCADES	$2.31x10^{-10}$	SMMAT-E	0.0021	33	CROATIA-Korcula	$6.57x10^{-9}$	0.0012	29	17

FG1n total/G1n	*FUT8*	<0.05	pLoF and missense	7	CROATIA-Korcula	6.74×10^{-8}	Burden	0.0072	177	ORCADES	3.04×10^{-6}	0.0037	43	5
GP21	*RFXAP*	<0.01	pLoF and missense	2	ORCADES	1.04×10^{-7}	SMMAT-E	0.0033	26	VIKING	6.29×10^{-2}	0.0009	2	1

Lead glycan - glycan trait reporting the strongest rare-variants association at the gene. Gene – gene for which rare variants are grouped. MAF - the highest allele frequency of variants aggregated for the mask. Variants - functional consequence of variants in the mask, aggregated in the given gene (pLoF – predicted loss of function: * denotes the number of algorithms predicting the missense variant to be deleterious). N variants - number of variants included in the mask. Discovery cohort - cohort reporting the lower p-value for the glycan-gene association. Discovery/Repl. P - p-value of the gene-based association with the lead glycan trait in the discovery/replication cohort. Assoc. test – gene-based association test for which p-value is reported. Discovery/repl. MAF - mean minor allele frequency of variants from the mask in the discovery/replication cohort. Discovery/repl. AC - sum of allele counts of variants from the mask in the discovery/replication cohort. Repl. Cohort - cohort reporting the higher p-value for the glycan-gene association. No. of glycans - number of other glycan traits associated with variants from the same mask. Results for transferrin glycome are reported at the top of the table, while results for IgG glycome are reported at the bottom of the table. Discovery significance threshold is 8.06×10^{-8} for transferrin and 1.19×10^{-7} for IgG glycans. Replication significance threshold is 3.23×10^{-4} for transferrin and 5.95×10^{-4} for IgG glycans.

IgG glycans gene-based aggregation meta-analysis

To further increase statistical power, we performed gene-based aggregation meta-analysis of IgG glycan traits for ORCADES and VIKING cohorts. In addition to two genes already found to be associated with IgG in the cohort-specific analysis (*MGAT3* and *ST6GAL1*), the combined analysis of VIKING and ORCADES cohort added *FUT6* to the list of genes whose rare variants are significantly (p-value<1.19×10^{-7}) associated with IgG glycan traits (Supplementary Table 5). *FUT6* is another gene known to be involved in glycosylation[135–138], encoding a glycosyltransferase enzyme that catalyses the transfer of fucose moieties to a growing glycan chain.

Genetic architecture of aggregated effects of rare-variants

To better understand the genetic architecture of identified associations, we next assessed whether our findings could be discoverable within a GWAS framework and whether they are driven by single variants or multiple rare variants working in concert to affect levels of transferrin/IgG glycosylation.

We first performed GWAS on imputed genotypes for each glycan trait and then repeated the rare variant association tests incorporating the dosages of GWAS sentinel SNPs as additional covariates. Genome-wide significant (transferrin p-value<1.61×10^{-9}, IgG p-value<2.38×10^{-9}) associations reported in this study (Supplementary Table 6 for transferrin glycans and Supplementary Table 7 for IgG glycans) have been described in further details in Landini *et al*.[215] and Klarić *et al*.[135]. Overall, aggregated associations with variants from 3 out of 8 genes, *FUT8*, *ST6GAL1* and *MGAT3*, remained significant (p-value <8.06×10^{-8} for transferrin and p-value<1.19×10^{-7} for IgG) after conditioning on sentinel GWAS associations (Table 3). For one gene, *RFXAP*, there were no significant associations in the GWAS analysis, while the remaining four gene-based associations (*FUT6*, *MSR1*, *FOXI1* and *TIRAP*) were explained by sentinel GWAS variants. For three of these genes, *FUT6, MSR1* and *FOXI1*, the GWAS

sentinel variants are low frequency (0.02<MAF<0.05; Supplementary Table 8). More specifically, for transferrin glycans, 14 out of the 16 glycome-gene aggregate pairs fail to reach genome-wide significance (p-value $<8.06 \times 10^{-8}$) after conditioning on GWAS sentinel SNPs (Supplementary Table 9), meaning that a considerable part of the rare variant signal was being tagged by variants identifiable by GWAS. In contrast, for IgG glycans, 24 of the 32 glycome-gene aggregate pairs remained significant (p-value$<1.19 \times 10^{-7}$), even after adjusting for GWAS sentinel SNPs (Supplementary Table 10).

Next, we performed single-point exome-wide association analysis (ExWAS) and repeated the aggregated rare variant association tests while conditioning on the sentinel ExWAS associations. In this way we tested whether the rare-variants associations with glycosylation were driven by a single variant (i.e. showed an attenuated signal after conditioning on the sentinel ExWAS variant) or were actually affected by multiple rare variants in concert (i.e. associations remain significant after the conditioning). Two of the associations, between variants in the *FUT8* gene and IgG glycans, and variants in the *MGAT3* gene and transferrin glycans, remain significant after conditioning on the sentinel ExWAS variant (Table 3). Upon closer inspection, for both of these genes, the sentinel ExWAS variant was a common variant that is also an eQTL for the gene in blood (eQTLGen[228]: rs35949016, *FUT8* eQTL, p-value = 6.5×10^{-159}; rs6001566, *MGAT3* eQTL p-value = 3.9×10^{-230}) (Supplementary Table 8). Hence, it appears that glycosylation is affected by common variants and independently by aggregates of rare variants in these two genes. Indeed, by looking at the single-point effects of each rare variant from the mask, we can see that multiple independent rare variants contribute to the effect on glycosylation levels (Supplementary Table 11).

In summary, four of the identified associations, three with low-frequency variants from *FUT6*, *MSR1* and *FOXI1* and one with a common variant from the *TIRAP* gene, could have been discovered using a GWAS of imputed genotype data. On the other hand, the rare variant association at *ST6GAL1* gene could only have

been discovered using an ExWAS, as it is too rare to be imputed well. Finally, associations with variants from two genes, *FUT8* and *MGAT3*, are driven by multiple rare variants simultaneously contributing to glycosylation of IgG and transferrin. Also the rare variant association at *RFXAP* gene could not have been discovered by either GWAS or ExWAS as there were no significant single-point associations. However, it is important to note that we could not replicate this association because the variants from its mask are depleted in the other studied cohorts (Supplementary Table 4).

Table 3: Genetic architecture of aggregated effect of rare variant associations when conditioning on sentinel variants from GWAS or ExWAS analysis. Two associations, those with variants from the *FUT8* and *MGAT3* regions remain significant after conditioning on GWAS/ExWAS sentinel variants. Associations with variants from the *MSR1* gene are dependent on both GWAS and ExWAS sentinel variants. The association with variants from the *ST6GAL1* gene is driven by the sentinel ExWAS variant, which was not present in the imputed GWAS.

Lead glycan	Gene	MAF	Variants	Association test	cohort	Discovery P	GWAS adj p	ExWAS adj p
Transferrin								
TfGP20	*FUT8*	<0.05	pLoF and deleterious (1/5)*	Burden	CROATIA-Korcula	$6.29x10^{-20}$	$2.75x10^{-12}$	$2.70\ x10^{-15}$
TfGP32	*FUT6*	<0.05	pLoF and deleterious (1/5)*	SKAT	CROATIA-Korcula	$4.31x10^{-18}$	$3.07x10^{-1}$	$2.94\ x10^{-1}$
TfGP35	*MSR1*	<0.05	pLoF	Burden	CROATIA-Korcula	$6.93x10^{-17}$	$6.62x10^{-7}$	$3.70\ x10^{-3}$
TfGP17	*TIRAP*	<0.05	pLoF and deleterious (1/5)*	SKAT-O	VIKING	$2.17x10^{-10}$	$8.61x10^{-1}$	$1.38\ x10^{-1}$
TfGP23	*FOXI1*	<0.05	pLoF and missense	SKAT-O	VIKING	$1.37x10^{-08}$	$6.85x10^{-1}$	$6.38\ x10^{-1}$

IgG								
FG2S1/(FG2+FG2S1+FG2S2)	*ST6GAL1*	<0.01	pLoF and missense	Burden	ORCADES	9.82×10^{-22}	6.99×10^{-19}	1.44×10^{-2}
Fn/(Bn+FBn)	*MGAT3*	<0.01	pLoF and deleterious (1/5)*	SMMAT-E	ORCADES	2.31×10^{-10}	5.47×10^{-10}	5.68×10^{-10}
FG1n total/G1n	*FUT8*	<0.05	pLoF and missense	Burden	CROATIA-Korcula	6.74×10^{-8}	2.31×10^{-6}	8.4×10^{-6}
GP21	*RFXAP*	<0.01	pLoF and missense	SMMAT-E	ORCADES	1.04×10^{-7}	1.04×10^{-7}**	1.04×10^{-7}**

Lead glycan - glycan trait reporting the strongest rare-variants association at the gene. Gene – gene for which rare variants are grouped. MAF - the highest allele frequency of variants aggregated for the mask. Variants - functional consequence of variants in the mask, aggregated in the given gene (pLoF – predicted loss of function: * denotes the number of algorithms predicting the missense variant to be deleterious). Association test – gene-based association test for which p-value is reported. Cohort - cohort reporting the lower p-value for the glycan-gene association. Discovery P - p-value of the gene-based association test with the lead glycan trait in the cohort. GWAS adj p – p-value of association test when conditioning on the significant variants from the GWAS analysis; ** no significant GWAS variants were found. ExWAS adj p – p-value of association test when conditioning on the significant variants from the ExWAS analysis; ** no significant ExWAS variants were found. Results for transferrin glycome are reported at the top of the table, while results for IgG glycome are reported at the bottom of the table. P-value significance threshold is 8.06×10^{-8} for transferrin and 1.19×10^{-7} for IgG glycans.

Links to health-related traits

We next wanted to assess the potential impact of rare protein glycosylation variants on health. Since some of the gene-glycan associations are population-specific, stemming from the genetic drift in isolated populations, we first performed “gene-level PheWAS” with quantitative health-related traits measured in studied cohorts. At the same time, since these cohorts, because of their sample size, might be underpowered to detect associations with common diseases, we queried public repositories of aggregated rare-variants associations for these genes.

We performed “gene-level PheWAS” with 116 quantitative health-related traits measured in the ORCADES, CROATIA-Korcula and VIKING cohorts, limited to the genes containing exonic pLoF and missense variants that were associated with transferrin or IgG glycome variation (Table 2). When possible, we sought to perform the analysis in the same cohort where the glycan-gene association was discovered. The only significant (p-value<5.4×10^{-5}) association was with transferrin glycosylation-associated rare variants from the *MSR1* gene and blood levels of HbA1c in the VIKING cohort (Supplementary Table 12). However, the association with HbA1c levels is not significant in CROATIA-Korcula, the cohort where we discovered the connection between *MSR1* and transferrin glycosylation, and it also does not replicate in ORCADES, suggesting that it might be a false positive association. We next checked whether any of the glycome-associated genes were significantly associated with health-related traits in UK Biobank. We used two repositories of aggregated rare-variants associations: Genebass[229] and the AstraZeneca PheWAS portal[230]. Missense variants from the *MSR1* gene were significantly associated with insulin-like growth factor 1 levels (*IGF1*) in both Genebass (SKAT-O p-value = $4.6\text{x}10^{-10}$) and the PheWAS portal (p-value = $1.6\text{x}10^{-24}$, for the “ptv5pcnt” collapsing model).

Discussion

Statistical power to detect associations with rare genetic variants can be increased by aggregating the association signals across multiple rare variants in a gene[231], or by using genetically isolated populations where, due to genetic drift, some variants are increased in frequency compared to a general population[219]. Further, intermediate phenotypes, more proximal to the genes and consequently more strongly influenced by them, can be used as “proxies’ of complex diseases to boost power. Glycosylation, one of the most common post-translational modifications, is one such intermediate phenotype and has been implicated in many diseases[222,224,225]. Here, we utilised the power of genetic isolates, aggregation of multiple rare variants and intermediate phenotypes to study the effect of rare variants on glycosylation of two proteins and their effect on disease.

We performed multiple gene-based aggregation tests to assess associations with transferrin (N = 1907) and IgG (N = 4912) glycan traits in three isolated cohorts of European descent, testing rare (MAF<5%) pLoF and missense variants from whole exome sequencing. We found rare variants from 8 genes contributing to glycan levels of either IgG or transferrin. As previously observed in GWAS using imputed genotypes, transferrin and IgG glycans showed mostly protein-specific gene-based associations[215], including genes encoding known glycosylation enzymes (transferrin - *TIRAP,* a gene in the proximity of *ST3GAL4*; IgG - *ST6GAL1* and *MGAT3*), transcription factors (transferrin - *FOXI1*), as well as other genes (transferrin - *MSR1*; IgG - *RFXAP*). On the other hand, rare variants in *FUT8* and *FUT6*, genes encoding fucosyltransferase enzymes adding core and antennary fucose structures to the synthesised glycan, were associated with glycosylation of both proteins. Previously we showed that, while glycosylation of both transferrin and IgG proteins is associated with genes encoding FUT6 and FUT8 fucosylation enzymes, these associations are driven by independent, protein-specific variants mapped to the regulatory region of the two genes[215]. Accordingly, here we identified rare variants in the exonic portions of *FUT8* and *FUT6*, acting independently or in concert with GWAS-identifiable variants.

We successfully replicated 4 gene-glycan associations (*FUT6*, *FUT8*, *TIRAP* and *MGAT3*); however, noting variants in certain genes were lower in frequency (*MSR1* and *FOXI1*) or completely absent (*ST6GAL1* and *RFXAP*) in replication cohorts, we were underpowered to replicate the glycan associations with the remaining four genes. Two of the 8 identified associations, the ones with variants from the *FUT8* and *MGAT3* genes, were driven by multiple rare variants simultaneously contributing to protein glycosylation. The association with variants from the *ST6GAL1* gene would have been discovered using single-point ExWAS (but not GWAS). Interestingly, for all three of these genes, we have also detected common variants independently affecting IgG and transferrin glycans. While four associations (*TIRAP*, *FUT6*, *MSR1* and *FOXI1*) could have been discovered using a GWAS of imputed genotype data, three of them (*FUT6*, *MSR1* and *FOXI1*) were with low frequency variants ($0.02 < MAF < 0.05$). The associations with the *RFXAP* gene could not have been discovered by either GWAS or ExWAS single-point analysis.

Except for *RFXAP* and *TIRAP*, all of 8 identified genes have already been associated with IgG and transferrin glycosylation in previous GWAS studies[135–138,215]. The novel gene *TIRAP* is located in close proximity to *ST3GAL4*, another glycosyltransferase-coding gene known to be associated with transferrin glycosylation. *TIRAP* has a function in the innate immune system, where it is involved in cytokine secretion and the inflammatory response[232,233]. The lead rare variant in the mask, rs8177399 (Supplementary Table 3), in addition to being an expression QTL (eQTL) for *TIRAP* and several other genes, is also a splicing QTL (sQTL) for *ST3GAL4* in whole blood (GTEx[234], p-value = 1.9×10^{-8}). The regulatory factor X-associated protein encoded by *RFXAP* gene, whose variants are associated with IgG glycans, is part of a multimeric complex, called the RFX DNA-binding complex, that binds to certain major histocompatibility (MHC) class II gene promoters and activates their transcription. MHC-II molecules are transmembrane proteins, found on the surface of professional antigen-presenting cells (including B cells)[235], which have a central role in development and control of the immune response. While the mechanism of *TIRAP*'s influence on the

glycome could be through controlling the splicing of the known glycosyltransferase enzyme ST3GAL4, the precise role of *RFXAP* in protein glycosylation still needs to be established.

Changes in the glycosylation patterns are often observed in a wide range of pathological states, such as cancer, inflammatory, autoimmune, neurodegenerative and cardiovascular diseases[236–239]. We thus assessed the potential involvement of glycome-associated genes in health, by performing, in the same three cohorts, gene-based association tests of 116 quantitative health-related traits, limited to genes whose rare variants we found associated with the protein glycomes. However, given the likely small effect-size of variants on complex diseases, we did not find any significant associations. On the other hand, using publicly available repositories of gene-based associations in the UK Biobank data, we found that the low-frequency, stop-gained variant rs41341748 from *MSR1* (associated with transferrin glycosylation) is also associated with blood levels of insulin-like growth factor 1 (IGF1). IGF1 is a hormone with significant structural and functional similarities to insulin: lower levels of IGF1 are associated with higher risk of Type 1 and 2 diabetes mellitus[240,241]. Recently, a rare deleterious missense variant in *IGF1* receptor (*IGF1R*) was found to be significantly associated with Type 2 diabetes in UK Biobank, further corroborating the link between IGF1 and diabetes[242]. In addition, genetic variants in *MSR1* have been previously associated with plasma levels of the galectin-3-binding protein[81]. Similarly to IGF1, galectin-3 has been identified as a marker and a pathogenic factor in type 2 diabetes, with the serum protein levels increased in type 2 diabetes patients[243–247]. An important part of iron delivery depends on recycling transferrin via clathrin-mediated endocytosis. Interestingly, binding of galectin-3 to transferrin can affect its intracellular trafficking[248,249]. Based on the glycosylation profile, galectin-3 was found bound only to a select, minor fraction (~5%) of transferrin, while interestingly none or little was bound to IgG[249]. Overall, variants from the *MSR1* gene seem to have a pleiotropic effect on transferrin glycosylation, and, based on literature, on galectin-3 and IGF1. In turn, both galectin-3 and IGF1 are reported to be associated with type 2 diabetes. The

potential role of glycosylation of transferrin in these processes still needs to be established.

In conclusion, we identified rare pLoF and missense variants associated with transferrin and IgG N-glycome, in both known and not previously reported genes (*TIRAP, RFXAP*). By utilising the power of genetic isolates and aggregated effects of rare variants, we discovered biologically relevant associations with a 300-fold up-drifted variant in the ORCADES cohort (in the sialyltransferase gene, *ST6GAL1*, affecting levels of sialylation of IgG) and associations independent of single-point GWAS and ExWAS analyses (in glycosyltransferase genes *FUT8* and *MGAT3*). Interestingly, many of glycan traits are influenced both by common and rare variants, revealing a complex genetic architecture of these intermediate phenotypes. While we did not find any robust links between glycome-associated genes and diseases in studied cohorts, we discover a potential link between transferrin glycosylation, galectin-3, IGF1 and diabetes. The exact mechanism behind these connections still needs to be confirmed and further explored. This study shows that, utilising the power of genetic isolates, gene-based aggregation tests and intermediate phenotypes such as glycosylation, rare variant associations are detectable even in relatively small sample sizes (low thousands). However, larger cohorts would be required to identify the contribution of rare variants to multifactorial, complex diseases.

Methods

Ethics

All studies were approved by local research ethics committees and all participants have given written informed consent. The ORCADES study was approved by the NHS Orkney Research Ethics Committee and the North of Scotland REC. The CROATIA-Korcula study was approved by the Ethics Committee of the Medical School, University of Split (approval ID: 2181-198-03-04/10-11-0008). The VIKING study was approved by the South East Scotland Research Ethics Committee, NHS Lothian (reference: 12/SS/0151).

Genotypic data

Exome sequencing

The “Goldilocks” exome sequence data for ORCADES, CROATIA-Korcula and VIKING cohorts was prepared at the Regeneron Genetics Center, following the protocol detailed in Van Hout *et al.*[217] for the UK Biobank whole-exome sequencing project. In summary, the multiplexed samples were sequenced on the Illumina NovaSeq 6000 platform using S2 flow cells. The raw sequencing data was processed by automated analysis using the DNAnexus platform[250], where files were converted to FASTQ format, and then aligned to GRCh38 genome reference using the BWA-mem[251]. Duplicated reads were identified and flagged by the Picard tool[252]. Genotypes for each individual sample were called using the WeCall variant caller[253]. During quality control, samples genetically identified as duplicates, showing disagreement between genetically determined and reported sex, high rates of heterozygosity or contamination, low sequence coverage (less than 80% of targeted bases achieving 20X coverage) or discordant with genotyping chip were excluded. The number of samples removed after quality control are listed in Supplementary Table 13 for each cohort. Finally, the “Goldilocks” dataset was generated by (i) filtering out genotypes with read depth

lower than 7 reads, (ii) keeping variants having at least one heterozygous variant genotype with allele balance ratio greater than or equal to 15% (AB ≥ 0.15) or at least one homozygous variant genotype, and (iii) filtering out variants with more than 10% of missingness and HWE $p<10^{-6}$. Overall, a total of 2,090 ORCADES (820 male and 1,270 female), 2,872 CROATIA-Korcula (1,065 male and 1,807 female) and 2,108 VIKING (843 male and 1,265 female) participants passed all exome sequence and genotype quality control thresholds. A pVCF file containing all samples passing quality control was then created using the GLnexus joint genotyping tool.[254]

Variant annotation

Exome sequencing variants were annotated as described in Van Hout, *et al.*[217] In brief, each variant was labelled with the most severe consequence across all protein-coding transcripts, implemented using SnpEff[255]. Gene regions were defined according to Ensembl release 85. Variants annotated as stop gained, start lost, splice donor, splice acceptor, stop lost and frameshift were considered as predicted LOF variants. The deleteriousness of missense variants was assessed using the following algorithms and classifications (based on dbNSFP 3.2): (1) SIFT: “D” (Damaging), (2) Polyphen2_HDIV: “D” (Damaging) or “P” (Possibly damaging), (3) Polyphen2_HVAR: “D” (Damaging) or “P” (Possibly damaging), (4) LRT[256]: “D” (Deleterious) and (5) MutationTaster[257]: “A” (Disease causing automatic) or “D” (Disease causing). Missense variants were considered “likely deleterious” if predicted as deleterious by all five algorithms, “possibly deleterious” if predicted as deleterious by at least one of the algorithms and “likely benign” if not predicted as deleterious by any of the algorithms.

Generation of gene burden masks

For each gene, we grouped the variants in the gene in four categories (masks), based on severity of their functional consequence. Mask 1 included only predicted loss-of-function (pLoFs) variants, mask 2 consisted of pLoF variants and all missense variants, and masks 3 and 4 contained pLoF and predicted

deleterious missense variants ("possibly deleterious" and "likely deleterious" for mask 3 and mask 4, respectively). We considered two separate variations of each mask based on the frequency of the minor allele of the variants that were screened in that group: MAF ≤ 5% and MAF ≤ 1%. Overall, up to 8 burden tests were performed for each gene (Supplementary Table 14). Consequently, the masks are not independent - certain masks will include the variants listed in a different mask and additional, less severe or more frequent variants.

Phenotypic data

Transferrin and IgG N-glycome quantification

Transferrin and total IgG N-glycome quantification for ORCADES, VIKING and CROATIA-Korcula samples was performed at Genos Glycobiology Laboratory, following the protocol described in Trbojević-Akmačić *et al.*[191] for transferrin, in Pučić *et al.*[258] for IgG in ORCADES cohort and batch 1 of CROATIA-Korcula cohort, in Trbojević-Akmačić *et al*.[259] for IgG in VIKING cohort and batch 2 of CROATIA-Korcula cohort. In summary, proteins of interest were first isolated from blood plasma (IgG depleted blood plasma, in the case of transferrin) using affinity chromatography binding to anti-transferrin antibodies plates for transferrin and protein G plates for IgG. The protein isolation step was followed by release and labelling of N-glycans and clean-up procedure. IgG N-glycans have been released from total IgG (all subclasses). N-glycans were then separated and quantified by hydrophilic interaction ultra-high-performance liquid chromatography (HILIC-UHPLC). As a result, transferrin and total IgG samples were separated into 35 (transferrin: TfGP1 − TfGP35) and 24 (IgG: GP1 − GP24) chromatographic peaks. It is worth noting that there is no correspondence structure-wise between transferrin TfGP and IgG GP traits labelled with the same number.

Normalisation and batch correction

Prior to genetic analysis, raw N-glycan UHPLC data was normalised and batch corrected to reduce the experimental variation in measurements. Total area normalisation was performed by dividing the area of each chromatographic peak (35 for transferrin, 24 for IgG) by the total area of the corresponding chromatogram. Due to the multiplicative nature of measurement error and right-skewness of glycan data, normalised glycan measurements were log10-transformed. Batch correction was then performed using the empirical Bayes approach implemented in the "ComBat" function of the "sva" R package[193], modelling the technical source of variation (96-well plate number) as batch covariate. Batch corrected measurements were then exponentiated back to the original scale. Prior to further analysis, each glycan trait was rank transformed to normal distribution using the "rntransform" function from the "GenABEL" R package[194].

Derived glycan traits

IgG derived traits analysed included those defined by Huffman *et al*.[141], and were calculated using the glycanr R package. In addition, new derived traits were calculated for both transferrin and IgG, representing the overall presence of a certain sugar structure on the totality of transferrin/IgG N-glycan traits measured (e.g. percentage of fucosylation). These newly generated traits are expected to give a direct insight in the biological pathway involved in the addition of the sugar moiety to glycan structures. Exact formulas used for defining transferrin and IgG newly derived traits can be found in Supplementary Tables 15 and 16 respectively.

Health-related quantitative traits

To evaluate the potential effect of rare variants affecting glycome on health-related phenotypes, in the same cohorts we collected 148 health-related, quantitative traits (e.g. anthropological measurements, blood levels of proteins, metabolites and biomarkers). Excluding traits with fewer than 800 samples, a total

of 116 traits were considered for analysis (75 traits for ORCADES, 79 for VIKING and 47 for CROATIA-Korcula cohort). Each health-related trait was rank transformed to normal distribution using the "rntransform" function from the "GenABEL" R package[194], followed by applying the rare-variants association pipeline described below.

Gene-based aggregation analysis

We performed variant Set Mixed Model Association Tests (SMMAT)[48] on rank-transformed glycan traits, fitting a generalised linear mixed model (GLMM) adjusting for age, sex, sampling batch in the case of CROATIA-Korcula IgG glycan traits, and familial or cryptic relatedness by kinship matrix. The kinship matrix was estimated from the genotyped data using the 'ibs' function from GenABEL R package[194]. The SMMAT framework includes 4 variant aggregate tests: burden test, sequence kernel association test (SKAT), SKAT-O and SMMAT-E, a hybrid test combining the burden test and SKAT. The 4 variant aggregate tests were performed on 8 different pools of genetic variants, called "masks", described above (Supplementary Table 14).

Discovery significance threshold was Bonferroni corrected for the approximate number of genes in the human genome, 20,000, and the number of independent glycan traits, 21 for IgG and 31 for transferrin ($0.05/20000/31 = 8.06 \times 10^{-8}$ for transferrin, $0.05/20000/21 = 1.19 \times 10^{-7}$ for IgG). The number of independent glycan traits was estimated as the number of principal components that jointly explained 99% of the total variance of transferrin/IgG glycan traits in each cohort (Supplementary Tables 17 and 18). PCA was calculated on rank-transformed glycan traits, separately for each cohort, using the "prcomp" function from "factoextra" R package[260]. A gene association was considered significant if it passed the above-described Bonferroni corrected significance threshold in at least one of the 4 performed variant aggregate tests and if the cumulative allele count of the variants included in the gene was equal or higher than 10. Replication significance threshold was defined as P = 0.05 divided by the number of genes

and independent glycans to be replicated. For IgG glycans, this threshold was P = $5.95x10^{-4}$ (P = 0.05/4 genes/21 glycans) and for transferrin glycans, this threshold was P = $3.23x10^{-4}$ (P = 0.05/5 genes/31 glycans).

A similar analysis plan was applied to the health-related phenotypes analysed. Variant Set Mixed Model Association Tests (SMMAT)[48] was performed on rank-transformed traits, fitting a GLMM adjusting for age, sex, first 20 ancestral principal components (PCs), batch covariates when available (e.g. season, time of the day and batch/subcohort) and familial or cryptic relatedness.

IgG glycome gene-based aggregation meta-analysis

Gene-based aggregation analysis of IgG glycan traits for ORCADES and VIKING cohorts was repeated following the same approach as previously described, except for the restriction that masks included only variants present in both cohorts. Since IgG GP3 was not quantified in ORCADES cohort, this glycan was excluded from the meta-analysis, bringing the total number of IgG glycan traits considered to 93. We then used the “SMMAT.meta” function of “SMMAT” R package[48] to meta-analyse, for each trait, the two studies. To identify significant results we filter results by the previously described Bonferroni-corrected significance threshold of $1.19x10^{-7}$ and by the cumulative allele count of variants included in the gene equal or higher than 10.

Genome-wide association analysis

Genome-wide association analyses (GWAS) between HRC-imputed genotypes and 51 transferrin N-glycan traits were performed in 948 samples from CROATIA-Korcula and 959 samples from VIKING. GWAS with 94 IgG N-glycan traits were performed in 1960 samples from ORCADES, 1866 samples from CROATIA-Korcula and 1086 samples from VIKING. The sample size of the same cohort differs between transferrin and IgG due to the different number of samples successfully measured for glycosylation of each protein. Transferrin N-glycan

measurements were not available in ORCADES. Rank-transformed glycan traits were adjusted for age and sex, as fixed effects, and relatedness (estimated as the kinship matrix calculated from genotyped data) as random effect in a linear mixed model, calculated using the "polygenic" function from the "GenABEL" R package[194]. Since IgG N-glycan traits for the CROATIA-Korcula cohort were measured at two separate occasions, the two were considered as separate cohorts. Therefore, for CROATIA-Korcula, rank transformation was performed separately in each subcohort. Samples were then merged together for GWAS but adding batch (subcohort number - 1 or 2) as fixed effect covariate. Residuals of covariate and relatedness correction were tested for association with Haplotype Reference Consortium (HRC) r1.1-imputed SNP dosages using the RegScan v. 0.5 software, applying an additive genetic model of association.

The genomic control inflation factor (λ_{GC}) was calculated for each glycan and health-related trait. The mean genomic control inflation factor (λ_{GC}) for IgG glycan traits was 1.002 (0.982-1.026) in ORCADES, 1 in CROATIA-Korcula (0.971-1.031) and 0.993 in VIKING cohort (0.972-1.017) cohort; for transferrin glycan traits λ_{GC} was 1.002 in CROATIA-Korcula (0.982-1.026) and 0.998 in VIKING (0.974-1.021) cohort. Overall, the confounding effects of the family structure were correctly accounted for in our analyses.

Identification of rare variant associations independent of GWAS and ExWAS signals

To ensure that the rare variant associations identified were independent of associations with variants discoverable by a GWAS or single-point exome-wide (ExWAS) analysis, we repeated the aggregate analysis while conditioning on the sentinel SNPs from the single-variant genome-wide or exome-wide analysis. First, we performed GWAS of glycan traits using the same individuals as in the analysis of the exome-sequencing data, but using as genotypes SNP dosages imputed from the HRC imputation panel, as described above. For each glycan trait we defined the sentinel SNPs as the variants having the lowest significant p-

value ($p < 5x10^{-8}$) in a 1Mb window, and MAF > 1%. Then we also performed the exome-wide association analysis (ExWAS), following exactly the same protocol, but with exome sequencing data used for genotypes. We then re-run variant aggregate analysis as previously described, but with adjusting the glycan traits for the genotype of the sentinel SNPs from the GWAS/ExWAS significant loci, in addition to the other covariates listed above. The statistical significance level was determined in the same way as outlined in the main analysis above.

Replication of glycome rare associations in different cohorts and associations with health-related traits

To investigate whether glycome rare-variants associations were cohort specific, each significant gene-glycan trait pair from the cohort-level discovery analysis was tested for associations in the remaining cohorts. The p-value threshold for replication was set to $3.23x10^{-4}$ for transferrin (0.05/31/5) and $5.95x10^{-4}$ for IgG (0.05/21/4) glycans, correcting for the number of independent glycan traits (i.e. 31 for transferrin and 21 for IgG) and the number of discovered glycome-gene pairs (i.e. 5 for transferrin and 4 for IgG in gene-based aggregation analysis).

To investigate whether the glycome associated rare-variants may also affect health-related phenotypes, we tested for association each glycome-associated gene and 116 health-related traits The significance threshold was set to $5.43x10^{-5}$, correcting for the number of health-related traits (116), and the number of discovered glycome-gene pairs and number of glycome-associated genes (8).

3.3 Conclusion

As described in the above pre-print submission, I used gene-based aggregation tests to identify rare pLoF and missense variants associated with the transferrin (N = 1907) and IgG (N = 4912) N-glycome, in both known (transferrin - *FUT8*, *FUT6*, *MSR1* and *FOXI1*, as reported in Chapter 2; IgG - *ST6GAL1*, *MGAT3* and *FUT8*) and not previously reported genes (IgG - *RFXAP*). The synergy of genetic isolates and aggregating effects of multiple rare variants enabled identification of

genetic associations with protein glycome that would otherwise hardly be detectable in the general population, especially given our sample size (low thousands). For example, *ST6GAL1* rs750567016, affecting IgG glycosylation, is a 300-fold up-drifted variant in the ORCADES cohort and much larger sample sizes would be needed to detect this association with a biologically relevant gene in the general population. Further, multiple rare variants in *FUT8*, *MGAT3* and *RFXAP*, were associated with the transferrin/IgG glycome, independently of single-variant GWAS and ExWAS associations. Interestingly, several glycan traits are influenced by common and rare variants, suggesting that, at the current sample sizes, a mixed discovery strategy aimed at identifying the contribution of both common and rare variants, with both small and large effect sizes, should be adopted to elucidate the genetic architecture of protein glycome. I did not identify any robust links between glycome-associated rare variants and health related traits in our cohorts. However, consultation of repositories for aggregated rare-variants associations suggests a pleiotropic effect of the *MSR1*, stop gained rs41341748 variant on transferrin glycome and blood levels of IGF-1, which have been linked to risk of diabetes[240,241].. Nevertheless, the underlying mechanism behind a potential connection between the transferrin glycome, IGF1 and diabetes is currently speculative and requires to be further explored. Although gene-based aggregation tests, combined with intermediate phenotypes, such as protein glycome, and genetically isolated populations have demonstrated the ability to detect rare variant associations even in relatively small sample sizes (low thousands), it would still be beneficial to perform similar analyses in larger cohorts to understand the contribution of rare variants in multifactorial, complex diseases via protein glycome. I discuss further the topic of statistical power and sample size in Chapter 5. Finally, variant annotation in this study was performed using several bioinformatics tools, such as SnpEff, SIFT and PolyPhen, which allow for rapid classification and prioritisation of candidate variants, reducing the need for expensive and labour-intensive functional assays. Nevertheless, these *in silico* methods have inherent caveats and limitations, which I discuss in further details in Chapter 5.

While Chapter 2 and 3 discussed the genetic architecture of protein glycome, genetic variation contributing to another understudied omic trait, the lipidome of bile acids, is the focus of the next chapter.

Chapter 4: Genetic architecture of bile acid lipidome

4.1 Introduction

After focusing on genetic regulation of protein glycomics, in this chapter I investigate the genetic architecture of the bile acid lipidome, another understudied omic trait. Primary bile acids are synthesised from cholesterol in the liver and, after food ingestion, are secreted into the small intestine, where they emulsify lipid-soluble nutrients, promoting their absorption[142]. Once secreted in the gastrointestinal tract, primary bile acids are heavily modified by the gut microbiota to generate secondary bile acids. Then, the majority of bile acids are returned to the liver[143]. Therefore, the levels of bile acid in blood serum reflect the amount that has escaped extraction from the portal blood and, instead of being transported back to the liver, entered the systemic circulation. Bile acids are known to be influenced by environmental factors such as sex, with female sex and oestrogen playing roles in regulating bile acids production and composition[261]. High oestrogen levels in pregnancy can lead to increased serum bile acids, potentially by reducing expression of bile acid receptors and transport proteins[262]. Age-related hormone changes also contribute to differing bile acids production in women. This dimorphism is evident in both mice and humans, influencing aspects such as the rate of bile acid synthesis and the composition of the bile acid pool. Female mice show a larger total bile acid pool than males, excrete fewer bile acids in faeces, and have lower cholesterol catabolism via bile acid production compared to males[263].

Similarly to glycans, bile acids have been involved in both key physiological processes, such as lipid and glucose homeostasis, vitamin absorption and immunity[149,264], and diseases, such as hepatobiliary diseases, inflammatory bowel disease and cancer[152,265,266]. Nevertheless, research focusing on bile acid metabolites in a large sample from a general human population, as opposed to a disease cohort, is currently lacking[267,268]. A number of LC–MS/MS methods have been developed allowing analysis of free and conjugated BAs without

derivatization, however showing disadvantages with time consuming extraction procedures, long analysis times or lack of baseline separation of isobaric species. Bile acids molecules analysed in the current study have been measured by a LC–MS/MS method for simultaneous determination of free and conjugated BAs in plasma and serum with a runtime of 6.5 min[269].

In this chapter, I explore the genetic contributions to plasma bile acid lipidome, reporting associations with both common and low-frequency/rare variants and also sex-specific association signals. Further, I investigate whether the bile acid-associated variants have an effect on complex traits or diseases or vice-versa, that is whether the complex traits and diseases influence bile acid variability.

4.2 Manuscript pre-print

What follows is a manuscript submitted to a preprint server. A copy of the manuscript is included below, with permission from the co-authors. Supplementary Figures and Tables can be found at: https://www.medrxiv.org/content/10.1101/2022.12.16.22283452v1.

For this work, I pre-processed and performed imputation of bile acid data for all cohorts and conducted single-cohort GWAS for the ORCADES and CROATIA-Vis cohorts, both for the full sample and sex-specific analysis. Single-cohort GWAS of bile acids for the NSPHS, MICROS and ERF cohorts was performed by Åsa Johansson, Dariush Ghasemi-Semeskandeh and Shahzad Ahmad, respectively. I conducted meta-analysis of both full sample and sex-stratified sample. I also carried out all down-stream analyses. I performed gene-based aggregation analysis of bile acids using clean exome sequence data provided by Regeneron. Finally, I wrote the first draft of the manuscript, with the support of Gerhard Liebisch and Carsten Gnewuch regarding methods on bile acids quantification. The full list of author contributions can be found in the “Author contributions” section of this article.

Genome-wide association study reveals loci with sex-specific effects on plasma bile acids

Arianna Landini[1,2], Dariush Ghasemi-Semeskandeh[3,4], Åsa Johansson[5], Shahzad Ahmad[6], Gerhard Liebisch[7], Carsten Gnewuch[7], Regeneron Genetics Center[8], Gannie Tzoneva[8], Alan R. Shuldiner[8], Andrew A. Hicks[3], Peter Pramstaller[3], Cristian Pattaro[3], Harry Campbell[2], Ozren Polašek[9,10], Nicola Pirastu[11], Caroline Hayward[1], Mohsen Ghanbari[6], Ulf Gyllensten[5], Christian Fuchsberger[3], James F. Wilson*[1,2] & Lucija Klarić*[1]

1 MRC Human Genetics Unit, Institute for Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom
2 Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
3 Institute for Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck, Bolzano, Italy
4 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
5 Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
6 Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
7 Institute of Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Regensburg, Germany
8 Regeneron Genetics Center, Tarrytown, NY, USA
9 Department of Public Health, School of Medicine, University of Split, Split, Croatia
10 Algebra University College, Zagreb, Croatia
11 Genomics Research Centre, Human Technopole, Milan, Italy
* Authors contributed equally.
Correspondence to: J.F.W ( ) or L.K. ( )

Introduction

Bile acids (BAs) are synthesised from cholesterol in the liver and subsequently stored in the gallbladder. After ingestion of food, BAs are secreted into the small intestine, where they contribute to the digestion of lipid-soluble nutrients[142]. Approximately 95% of BAs are then re-absorbed by the intestinal epithelium and transported back to the liver via the portal vein - a process termed “enterohepatic circulation”[143]. Primary bile acids in humans consists of cholic acid (CA), chenodeoxycholic acid (CDCA), and their taurine- or glycine-bound derivatives (TCA and TCDCA, GCA and GCDCA). Once secreted in the lower gastrointestinal tract, primary BAs are heavily modified by the gut microbiota to produce a broad range of secondary BAs, with deoxycholic acid (DCA), a CA derivative, and lithocholic acid (LCA), a CDCA derivative, being the most prevalent[143]. Bile acids also act as hormone-like signalling molecules, serving as ligands to nuclear (hormone) receptors. Through activation of these diverse signalling pathways, BAs control not only their own transport and metabolism, but also lipid and glucose metabolism, and innate and adaptive immunity[149]. Bile acids are thus involved in regulating several physiological systems, such as fat digestion, cholesterol metabolism, vitamin absorption, and liver function[264]. In addition, given their role in coordinating bile homeostasis, biliary physiology and gastrointestinal functions, impaired signalling of BAs is associated with development of hepatobiliary diseases, such as cholestatic liver disorders, cholesterol gallstone disease and other gallbladder-related conditions[147], and of inflammatory bowel disease[152]. Further, bile acids have been implicated in carcinogenesis - specifically oesophageal, gastric, hepatocellular, pancreatic, colorectal, breast, prostate and ovarian cancer - both as pro-carcinogenic agents and tumour suppressors[156]. Thanks to their role as signalling molecules, BAs have been considered as possible targets for the treatment of metabolic syndrome and various metabolic diseases[270]. Further, BAs are able to facilitate and promote drug permeation through biological membranes, making them of general interest for drug formulation and delivery[271].

While many studies have focused on the genetic determinants of blood metabolites[77,157–160,162], research focusing specifically on bile acids in a large sample from the general population is currently lacking. Here we investigate the genetic architecture of primary and secondary BAs, reporting associations with both common and low-frequency/rare variants. First, we performed a genome-wide association meta-analysis (GWAMA) of plasma blood levels of 18 BA traits (N=4923). For a subset of this sample (female N=1088, male N=820), we perform sex-stratified GWAMA, to describe sex-specific genetic contributions to BA variability. We then explore whether complex traits or diseases have a role in influencing BA variability by using Mendelian Randomisation. We finally employ multiple gene-based aggregation tests to investigate rare (MAF < 5%) predicted loss of function (pLoF) and missense variants from whole exome sequencing affecting the 18 BA traits in a subset of our cohorts (N=1006).

Results

Loci associated with plasma levels of bile acids

To investigate the genetic control of bile acids, we performed a GWAS meta-analysis on five cohorts of European descent (N = 4923), studying the associations of blood plasma levels of 18 primary and secondary bile acid traits with HRC-imputed genotypes/whole exome sequence data. Based on the number of below limit-of-detection (LOD) measurements, BAs were analysed either as quantitative or binary traits (Supplementary Table 1). In addition, two analysis approaches were carried out in parallel for quantitative traits: in one case, <LOD values were considered as missing, in the other case, they were imputed (Methods). An additive linear model was assumed for each bile acid trait, followed by fixed-effect inverse-variance meta-analysis. Overall, we identified 2 loci that passed the significance threshold (p-value $< 3.57 \times 10^{-9}$, Bonferroni adjusted for the number of independent bile acid traits) (Figure1), near the

SLCO1B1 and *PRKG1* genes. The most strongly associated locus ($p=1.14 \times 10^{-16}$), on chromosome 12 near *SLCO1B1*, showed consistent directionality across 4 of the 5 populations (Table 1), with the effect allele T of the sentinel SNP, rs4149056, associated with decreased serum levels of GDCA (quantitative). In the same locus, we found GLCA and the imputed GDCA trait to be significantly associated with the rs73079476 variant (Supplementary Table 2), in high linkage disequilibrium with the sentinel SNP, rs4149056 ($r^2 = 0.97$). On the other hand, rs146800892, the sentinel SNP on chromosome 10 near *PRKG1*, has a minor allele frequency (MAF) lower than 1% in any cohort but CROATIA-Vis and might thus represent a cohort-specific association with GCA (Supplementary Table 2).

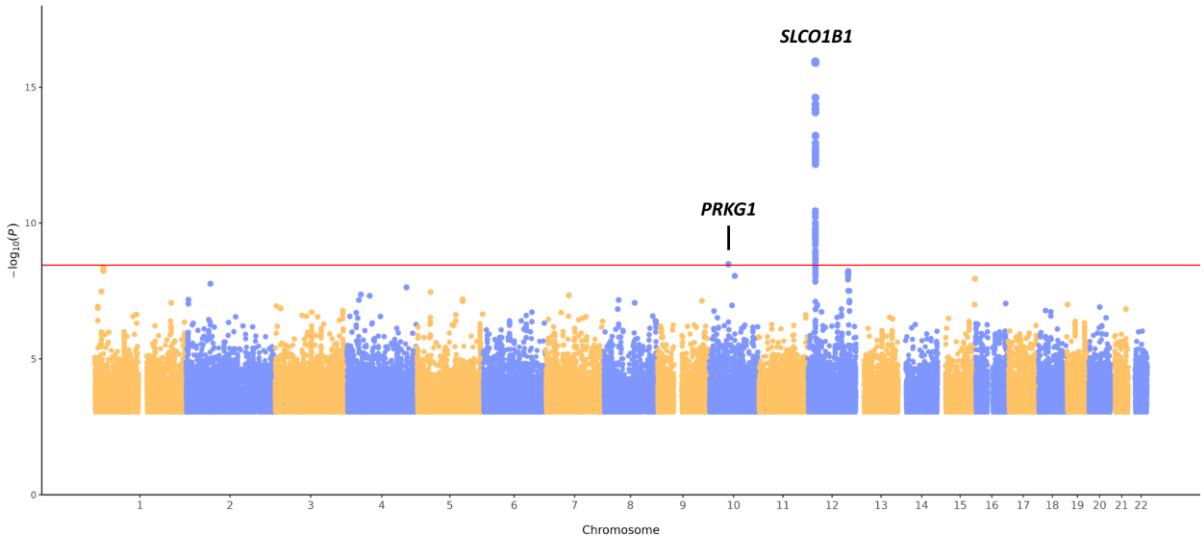


Figure 1. Summary Manhattan plot pooling together meta-analysis results obtained across 18 bile acid traits. The pooling was performed by selecting the lowest *p* value (y-axis) from the 18 bile acids for every genomic position (x-axis). The Bonferroni-corrected genome-wide significance threshold (horizontal red line) corresponds to 3.57×10^{-9}. For simplicity, SNPs with *p* value $> 1 \times 10^{-3}$ are not plotted. *P* values are derived from the two-sided Wald test with one degree of freedom.

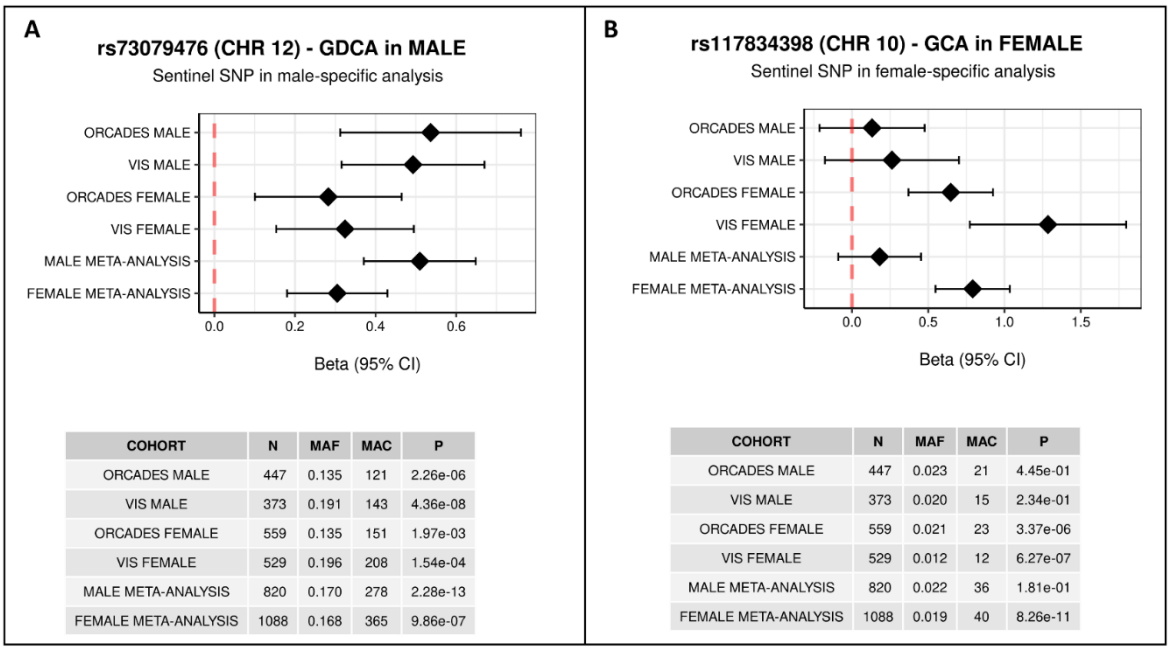
Table 1: Loci genome-wide significantly associated with at least one of the 18 bile acid traits in all samples and sex-stratified GWAMA

All samples										
Locus	**Gene**	**SNP**	**EA**	**OA**	**EAF**	**Beta**	**P**	**SE**	**N**	**Lead BA**
12:20994540-21463812	*SLCO1B1*	rs4149056	T	C	0.839	-0.25	1.10×10^{-16}	0.03	4547	GDCA
10:53832549-53832549	*PRKG1*	rs146800892	T	C	0.988	-0.96	3.30×10^{-09}	0.16	900	GCA
Sex-stratified										
Locus	**Gene**	**SNP**	**EA**	**OA**	**Beta M**	**Beta F**	**P M**	**P F**	**N (M/F)**	**Lead BA**
12:20994540-21463812	*SLCO1B1*	rs73079476	A	C	-0.51	-0.31	2.30×10^{-13}	9.90×10^{-07}	820/1088	GDCA
10:53832549-53832549	*PRKG1*	rs117834398	T	G	-0.18	-0.79	1.80×10^{-01}	8.30×10^{-11}	820/1088	GCA

Each locus is represented by the SNP with the strongest association in the region, according to the p-value rejecting the null hypothesis of no association with at least one of 18 bile acid traits. In all samples analysis, an association was considered significant if the p-value was lower than 3.57×10^{-9}, the genome-wide significance threshold Bonferroni-corrected for the number of independent bile acid traits. In sex-stratified analysis, an association was considered significant if the p-value was lower than 5×10^{-9}, the genome-wide significance threshold Bonferroni-corrected for the number of independent bile acid quantitative traits. The two SNPs in the *SLCO1B1* locus are in high LD (LD $r^2 = 0.97$), while the two SNPs in the *PRKG1* locus represent two distinct signals (LD $r^2 < 0.001$). Locus - coded as 'chromosome: locus start–locus end' (GRCh37 human genome build); Gene - suggested candidate gene; SNP - variant with the strongest association in the locus; EA - SNP allele for which the effect estimate is reported; OA - other allele; EAF - frequency of the effect allele; Beta - effect estimate for the SNP and bile acid with the strongest association in the locus; SE - standard error of the effect estimate, P - p-value of the effect estimate (two-sided Wald test with one degree of freedom); N - sample size; Lead BA - bile acid with the strongest association to the reported SNP; M - male specific analysis; F - female specific analysis.

Sex-specific associations of bile acid plasma levels

To investigate whether the genetic component influencing bile acid variation may differ between men and women, we performed sex-specific GWAS meta-analysis of the 14 quantitative (imputed) bile acid traits for ORCADES and CROATIA-Vis cohorts (female N=1088, male N=820) and discovered two sex-specific associations. The association of GDCA with rs73079476 from the *SLCO1B1* locus was significant in male-only GWAS (beta = -0.51, p-value = 2.28×10^{-13}) (Table 1, Figure 2A). The signal for the same locus in female-only GWAS, while consistent in terms of directionality, has a smaller effect size than in male-only analysis (beta = -0.31) and does not reach the significance threshold (p-value = 9.86×10^{-7}), despite the slightly higher sample size (Figure 2). This suggests that the genetic effect of *SLCO1B1* locus on the plasma levels of GDCA is larger in men than women. We also identified a sex-specific association of GCA at the *PRKG1* locus. In contrast to *SLCO1B1*, the sentinel SNP in *PRKG1*, rs117834398, has a larger effect in females than in males (female beta = -0.79, male beta = -0.18), and passed the significant threshold only in the female-specific analysis (female p-value = 8.26×10^{-11}, male p-value = 1.81×10^{-1}) (Table 1, Figure 2B). Interestingly, the sentinel SNPs at the *PRKG1* locus for the full meta-analysis and for the female-specific analysis are in linkage equilibrium ($r^2 < 0.01$) and represent two independent associations in that locus. Overall, none of the significant association identified in one sex was replicated in the other, suggesting that the genetic contribution to plasma BA levels is likely to be different in males and females. We have identified 13 additional associations (p-value $< 5 \times 10^{-9}$, Bonferroni adjusted for the number of independent quantitative bile acid traits) that might have sex-specific effects (Supplementary Table 3, Supplementary Figure 1). However, given the low allele frequencies and allele counts in the two analysed cohorts, further analyses are required to replicate these associations.



COHORT	N	MAF	MAC	P
ORCADES MALE	447	0.135	121	2.26e-06
VIS MALE	373	0.191	143	4.36e-08
ORCADES FEMALE	559	0.135	151	1.97e-03
VIS FEMALE	529	0.196	208	1.54e-04
MALE META-ANALYSIS	820	0.170	278	2.28e-13
FEMALE META-ANALYSIS	1088	0.168	365	9.86e-07

COHORT	N	MAF	MAC	P
ORCADES MALE	447	0.023	21	4.45e-01
VIS MALE	373	0.020	15	2.34e-01
ORCADES FEMALE	559	0.021	23	3.37e-06
VIS FEMALE	529	0.012	12	6.27e-07
MALE META-ANALYSIS	820	0.022	36	1.81e-01
FEMALE META-ANALYSIS	1088	0.019	40	8.26e-11

Figure 2. Sex-specific associations. The effect of rs73079476 on chromosome 12 on GDCA bile acid is almost as twice strong in males compared to the effect in females (Panel A). The effect of rs117834398 on GCA bile acid is stronger in females than in males (Panel B). N – sample size, MAF – minor allele frequency, MAC – minor allele count, CI – confidence interval.

Link with complex traits and diseases

Next, we assessed whether variants associated with BA levels have been previously associated with any other biochemical traits and diseases. Using Phenoscanner[183,272] we found that rs4149056, sentinel SNP in *SLCO1B1* locus, and its proxies (r^2 > 0.8), were also associated with concentration of bilirubin, non-bile acid metabolites, mean corpuscular haemoglobin, sex hormone binding globulin and estrone conjugates, and various responses to drugs (i.e., statin-induced myopathy, LDL-cholesterol response to simvastatin and methotrexate clearance in acute lymphoblastic leukaemia) (Supplementary Table 4). To obtain

deeper insight into the causal relationship between BAs and diseases, we conducted bi-directional Mendelian Randomisation (MR) analysis. Using the sentinel SNPs associated with GLCA, GDCA and GCA (Table 1) as instrumental variables we tested whether genetically increased levels of BA influence levels or risk for 548 biochemical traits and diseases available in the IEU Open GWAS database[209] (Supplementary Table 5). Levels of GLCA and GDCA were significantly (p-value < $0.05/(548x3) = 3.04\times10^{-5}$) associated with different biochemical measurements, such as levels of sex hormone-binding globulin, testosterone, triglycerides, vitamin D, alanine transaminase and galectin-3; with blood traits, such as mean corpuscular haemoglobin and mean corpuscular volume; and with diseases and their risk factors, such as daytime dozing and stroke (Supplementary Table 6). These MR tests were performed using the Wald ratio test utilising only a single instrument, thus the results of causal relationship between BAs and traits/diseases should be interpreted with caution. Yet our results suggest a possible overlap in genetic regulation, involving the *SLCO1B1* locus. Next, to assess whether complex traits and disease could have an effect on bile acid levels, we performed reverse MR using 548 traits/diseases as exposure and bile acids as outcomes. We found no significant associations, suggesting that none of the tested diseases or complex traits have an effect on BA levels (Supplementary Table 7).

Exome-wide rare variant analysis of bile acids

To assess the contribution of low frequency and rare variants to the bile acid genetic architecture, we performed exome-wide gene-based tests across 18 bile acid traits in the ORCADES cohort (N = 1006) by testing the aggregated effect of rare (MAF <5%) predicted loss-of-function (pLoF) and non-synonymous missense variants. We identified significant association (p-value $<1.79 \times 10^{-7}$) of rare variants from 3 genes with 2 bile acid traits (quantitative CA and binary THDCA). For these associations, a significant p-value was reported by at least 2 of the 4 aggregation tests used. Rare variants significantly associated with

quantitative bile acid trait CA are located in the *OR1G1* gene, while those associated with binary bile acid trait THDCA are located in *SART1* and *SORCS2* genes (Table 2, Supplementary Table 8). We further identified significant association of rare variants from *EPS8L1* gene with quantitative bile acid trait DCA and from *EEF2K* with binary bile acid trait THDCA (Supplementary Table 8). However, a significant p-value was reported by only one of the 4 aggregation tests used. Due to the lack of replication across aggregations tests, we considered these associations as not robust.

Table 2. Gene-based aggregation analysis results for bile acid traits in ORCADES cohort

BA	Trait type	Gene	MAF	Functional consequence	N variants	Aggregation test	P	AC
CA	Quantitative	*OR1G1*	<0.01	Missense variants	2	SKAT-O	1.67×10^{-8}	17
THDCA	Binary	*SORCS2*	<0.05	pLoF and missense variants	10	SMMAT-E	1.44×10^{-8}	174
THDCA	Binary	*SART1*	<0.01	pLoF and missense variants	4	SKAT	1.19×10^{-7}	25

BA- bile acid trait tested for rare variants association; Trait type – whether BA was analysed as a quantitative or binary trait; Gene - gene for which variants were aggregated; MAF - upper bound for minor allele frequency of tested variants; Functional consequence - predicted functional consequence for aggregated variants; N variants - number of variants in the mask; Aggregation test - rare-variants aggregation test reporting the lowest p-value out of 4 aggregation tests; P - p-value for the aggregation test; AC - cumulative allele count of all the variants in a mask. Bonferroni-corrected discovery p-value threshold was set to 1.79×10^{-7} (0.05/20,000 estimate of number of genes in the human genome/14 number of independent bile acids).

Discussion

Bile acids (BAs) are synthesised from cholesterol in the liver and then secreted into the small intestine to emulsify and promote absorption of lipid-soluble nutrients. BAs also act as hormone-like signalling molecules and have been linked to regulation of lipid and glucose metabolism, immunity, vitamin absorption, hepatobiliary diseases, inflammatory bowel disease and cancer. Despite the crucial role of BAs on whole-body physiology, their genetic architecture has not been extensively investigated in a large sample from the general population. In this study, we performed both pooled and sex-stratified genome-wide association meta-analysis of plasma levels of 18 bile acid compounds, including both primary and secondary forms, in 4923 European individuals.

We identified two secondary bile acids (GDCA and GLCA) significantly associated with a locus encompassing the *SLCO1B1* gene. The encoded protein, OATP1B1 (organic anion transporting polypeptide 1B1), is a well-known human hepatocyte transporter mediating the uptake of various endogenous compounds such as bile salts, bilirubin glucuronides, thyroid hormones and steroid hormone metabolites, and also clinically frequently used drugs like statins, HIV protease inhibitors, and the anti-cancer agents irinotecan or methotrexate[273–277]. The sentinel SNP of the *SLCO1B1* locus, rs4149056, is a missense variant (p.Val174Ala) which has been linked by previous GWA studies to blood concentration of several metabolites, including vitamin D[278], triglycerides[20] and bilirubin[279], a compound resulting from the breakdown of haem catabolism and excreted as a major component of bile. This same variant has also been associated with levels of sex hormone-binding globulin and testosterone[280]. The knock-out of the gene in mice results in abnormal liver physiology and abnormal xenobiotic pharmacokinetic phenotypes (Open Targets[281]). A rare variant from the *PRKG1* locus was significantly associated with levels of glycocholic acid (GCA). *PRKG1* encodes a Protein Kinase CGMP-Dependent 1, a protein involved in signal transduction and a key mediator of the nitric oxide/cGMP. The sentinel variant in the region, rs146800892, only passes the MAF threshold (MAF

> 0.01) in the CROATIA-Vis cohort, which is therefore the only cohort contributing to this association. Due to its demographic history and geographic position, CROATIA-Vis is a genetic isolate[282] so it is possible that this variant has increased in frequency compared to a general population[219].The mechanism of how the variation within this gene could relate to bile acid levels is unclear and would need to be further investigated.

In the sex-stratified GWAS meta-analysis, we observed sex-specific associations for the two identified loci. Levels of glycodeoxycholic acid (GDCA) are more strongly associated with the variant in *SLCO1B1* in men than in women, while female levels of GCA are more strongly affected by the variant in *PRKG1* than male levels. Later, our Mendelian randomization analysis did not provide evidence that testosterone, oestradiol, sex hormone-binding globulin or other sex-related traits have causal effects on plasma BA levels. While this could be due to a lack of statistical power of our BA meta-analysis, we currently have no evidence to suggest an effect of sex-related hormones on BA levels mediated by genetics. We also detected associations with variants from the same gene, *PRKG1*, in the main, non-stratified analysis. However, the two associations (sex-specific and pooled) appear to be independent (LD r^2 <0.001). While the association from the pooled analysis might be either false positive or population-specific, the independent association from the sex-stratified analysis replicates well between two analysed cohorts (CROATIA-Vis and ORCADES).

After assaying common variants through GWAS, we performed exome-wide gene-based association tests in a subset of our samples (N = 1006), to investigate the genetic contribution of rare and low frequency (MAF <5%) coding variants (pLoF and missense) to bile acid levels. Overall, we identified associations with rare variants from 3 genes, *OR1G1*, *SART1* and *SORCS2*. *OR1G1* is an olfactory receptor gene, whose coded protein receptor interacts with odorant molecules in the nose to initiate a neuronal response triggering the perception of smell[283,284]. In addition to the nasal level, the olfactory receptor coded by *OR1G1* is expressed also by enterochromaffin cells, specialised

enteroendocrine cells of the gastrointestinal tract. Braun *et al*[285]*.* determined that certain olfactory cues from spices and odorants, such as thymol, present in the luminal environment of the gut may stimulate serotonin release via olfactory receptors present in enterochromaffin cells. Between 90% and 95% of total body serotonin is in fact synthesised by enterochromaffin cells[286]: serotonin controls gut motility and secretion and is implicated in pathologic conditions such as vomiting, diarrhoea, and irritable bowel syndrome[285]. In mice, gut serotonin was shown to stimulate bile acid synthesis and secretion by the liver and gallbladder. Thus, release of serotonin in response to odorant cues increases bile acid turnover[287]. The hypoxia-associated factor (HAF), encoded by *SART1* gene and also known as SART1(800), is involved in proliferation and hypoxia-related signalling. The protein encoded by *SORCS2* is a receptor for the precursor of nerve growth factor, up-regulation of which has been reported for several liver pathologies, such as hepatotoxin- induced fibrosis[288], ischemia-reperfusion injury[289], oxidative injury[290], cholestatic injury[291] and hepatocellular carcinoma[288,292,293]. However, due to unavailability of exome sequencing data in other cohorts these associations were not replicated.

Recently, Chen *et al.*[268] have performed an association analysis on plasma and faecal levels of bile acids in 297 obese individuals. Their study revealed 27 associated loci, including genes involved in transport of GDP-fucose and zinc/manganese and zinc-finger-protein-related genes, mostly associated with bile acid levels in stool. In our study we analysed blood plasma in a much larger sample from a general population and discovered only two associated loci. Neither of genes identified in our study were reported in Chen *et al*, suggesting that genetic regulation of bile acids between stool and blood plasma or between obese and general populations might differ significantly.

We acknowledge several limitations in the present study. We found only a small percentage of BA variability to be affected by genetics, suggesting that a larger sample size is required to further describe BA genetic architecture. BAs are known to be largely influenced by environmental factors, such as sex and gut

microbiota. Female sex and oestrogens are considered relevant regulators of BA production and composition[261,263]. In pregnant women, high levels of circulating oestrogen are associated with development of cholestasis, characterised by increased plasma bile acids, likely via oestrogen reducing the expression of BA receptor and transport proteins[262]. Similarly, age-related differences in hormone levels influence the differential production of BAs in women[294]. The relevant impact of sex on plasma BA levels was confirmed by the sex-stratified analysis, where the two significantly associated loci showed to be sex-specific. Similarly, species-composition of gut microbiota has a great impact on BAs levels, especially for secondary BAs that are a direct result of microbiome activity. A recent study describing the effect of gut microbiota on the human plasma metabolome reported that both primary BA cholic acid (CA) and secondary BA deoxycholic acid (DCA) show a high percentage of variance explained by the microbiota (R^2 = 30% and 36%, respectively), indicating a strong impact on BAs of the variation in microbiota composition[295]. It is important to interpret our findings in the context of the tissue in which BA levels were measured, blood plasma. Bile acids are synthesised in the liver and secreted into the intestine, to be then reabsorbed into portal circulation and returned to the liver: plasma BA levels thus reflect the amount of BAs escaping extraction from the portal blood. Therefore, levels of BAs in plasma are likely to be influenced by genes other than those encoding the particular anabolic and catabolic enzymes, including those involved in hepatic function and dysfunction. In line with this, the major genetic contributor to blood BA levels in our study are variants from the *SLCO1B1* gene, encoding the hepatocyte transporter OATP1B1 and important for flux of bile salts, bilirubin glucuronides and various hormone metabolites, rather than genes encoding key enzymes of primary BA synthesis, such as *CYP7A1* and *CYP7B1*[296]*.* Similarly, some of the genes with rare variants associations have been linked to liver diseases, such as liver cancer[297], and intrahepatic cholestasis of pregnancy[298].

In conclusion, we explored the genetic architecture of plasma bile acid levels, including both common and rare variants. By performing GWAS meta-analysis (N = 4923), we identified 2 significantly associated loci, mapping to the *SLCO1B1* and *PRKG1* genes. In the sex-specific GWAS meta-analysis we observed that variants in these genes have different impact on bile acid levels in men and women. To assess relationships between genetically increased levels of bile acids and risk for diseases we performed Mendelian randomisation, but did not find any bile acids affecting disease risk, nor the reverse, which however might be affected by the lack of statistical power. Using the gene-based aggregated tests and whole exome sequencing, we further identified rare pLoF and missense variants in 3 genes associated with BAs, *OR1G1*, *SART1* and *SORCS2,* some of which are known to be involved in liver disease. Additional studies with larger sample sizes and of more diverse ancestry will be necessary to validate our findings, further unravel the genetic architecture of bile acid levels, and to understand their relationship with human diseases and complex traits.

Materials and methods

Ethics

All studies were approved by local research ethics committees and all participants have given written informed consent. The ORCADES study was approved by the NHS Orkney Research Ethics Committee and the North of Scotland REC. The CROATIA-Vis study was approved by the ethics committee of the medical faculty in Zagreb and the Multi-Centre Research Ethics Committee for Scotland. The Northern Swedish Population Health Study (NSPHS) was approved by the local ethics committee at the University of Uppsala (Regionala Etikprövningsnämnden, Uppsala). The MICROS study was approved by the ethical committee of the Autonomous Province of Bolzano, Italy. The ERF study was approved by the Erasmus institutional medical-ethics committee in Rotterdam, The Netherlands.

Phenotypic data

Bile acids quantification

Bile acid (BA) analysis was performed from plasma or serum (MICROS cohort) samples by liquid chromatography-tandem mass spectrometry (LC-MS/MS) as previously described[269]. The HPLC equipment consisted of a 1200 series binary pump (G1312B), a 1200 series isocratic pump (G1310A) and a degasser (G1379B) (Agilent, Waldbronn, Germany) connected to an HTC Pal autosampler (CTC Analytics, Zwingen, CH). A hybrid triple quadrupole linear ion trap mass spectrometer API 4000 Q-Trap equipped with a Turbo V source ion spray operating in negative ESI mode was used for detection (Applied Biosystems, Darmstadt, Germany). High purity nitrogen was produced by a nitrogen generator NGM 22-LC/MS (cmc Instruments, Eschborn, Germany). Gradient chromatographic separation of BAs was performed on a 50 mm × 2.1 mm (i.d.) Macherey-Nagel NUCLEODUR C18 Gravity HPLC column, packed with 1.8 µm particles equipped with a 0.5 µm pre-filter (Upchurch Scientific, Oak Harbor, WA,

USA). The injection volume was 5 µL and the column oven temperature was set to 50 °C. Mobile phase A was methanol/water (1/1, v/v), mobile phase B was 100% methanol, both containing 0.1% ammonium hydroxide (25%) and 10 mmol/L ammonium acetate (pH 9). A gradient elution was performed with 100% A for 0.5 min, a linear increase to 50% A until 4.5 min, followed by 0% A from 4.6 until 5.5 min and re-equilibration from 5.6 to 6.5 min with 100% A. The flow rate was set to 500 µL/min. To minimize contamination of the mass spectrometer, the column flow was directed only from 1.0 to 5.0 min into the mass spectrometer using a diverter valve. Otherwise, methanol with a flowrate of 250 µL/min was delivered into the mass spectrometer. The turbo ion spray source was operated in the negative ion mode using the following settings: Ion spray voltage = −4500 V, ion source heater temperature = 450 °C, source gas 1 = 40 psi, source gas 2 = 35 psi and curtain gas setting = 20 psi. Analytes were monitored in the multiple reaction monitoring (MRM). Quadrupoles Q1 and Q3 were working at unit resolution. Calibration was achieved by the addition of BAs to EDTA-plasma/serum. A combined BA standard solution containing the indicated amounts (0.5 - 70.5 µmol/L) was placed in a 1.5 ml tube and excess solvent was evaporated under reduced pressure before adding EDTA-plasma/serum. Calibration curves were calculated by linear regression without weighting. Data analysis was performed with Analyst Software 1.4.2. (Applied Biosystems, Darmstadt, Germany). The data were exported to Excel spreadsheets and further processed by self-programmed Excel macros which sort the results, calculate the analyte/internal standard peak area ratios, generate calibration lines and calculate sample concentrations. For the calculation we selected the internal standard with analogous fragmentation and closest retention time to the respective BA species.

Pre-processing of bile acid traits

Prior to genetic analysis, bile acid traits were grouped into three groups based on the percentage of samples with below the limit of detection (<LOD) measurements: high <LOD group (> ~30% of all samples below LOD) and low

<LOD group (< ~7% of all samples below LOD) (Supplementary Table 1 and Supplementary Table 10 for further details). Accordingly, different phenotypic pre-processing and different analysis strategies were applied to the groups. Bile acids within a high <LOD were considered as binary traits: individuals were categorised based on whether their bile acid levels were effectively measured (category 1) or were below the LOD (category 0). Bile acid traits belonging to this group were THDCA, TUDCA, TCA and GHDCA. All other bile acids were considered as quantitative traits and were $\log_{10}$-transformed. However, to increase the sample size, in addition to a complete-case analysis (considering as missing all samples with <LOD), we also imputed <LOD measurements. For each bile acid, imputation of <LOD measurements was performed by fitting a truncated normal distribution, with mean and standard deviation of the effectively measured raw data, truncated (as an upper bound) to the lowest measured value for the given bile acid. To do so, we used the "rtnorm" function from the MCMCglmm R package[299]. After imputation, measurements were $\log_{10}$-transformed.

Genome-wide association analysis

Genome-wide association studies (GWAS) were performed in 5 cohorts of European descent, CROATIA-Vis (N=971), ORCADES (Orkney Complex Disease Study) (N=1019), NSPHS (Northern Sweden Population Health Study) (N=718), MICROS (Micro-Isolates in South Tyrol) (N=1336) and ERF (Erasmus Rucphen Family Study) (N=879), for a combined sample size of 4923. Specific sample size for each bile acid molecule, in both meta-analysis and single cohort GWAS, can be found in Supplementary Table 10. Bile acid traits were adjusted for age, sex, batch, population structure/cryptic relatedness by including population principal components or applying linear mixed models and using a kinship matrix estimated from genotyped data. Within each cohort, residuals of covariate and population structure/relatedness correction were tested for association with Haplotype Reference Consortium (HRC)[28] imputed SNP dosages or SNP genotypes from whole genome sequencing, applying an additive

genetic model of association. Details of cohorts, individual-level pre-imputation QC, GWAS software and parameters specific for each cohort can be seen in Supplementary Table 11 Single-cohort summary statistics were filtered for minor allele frequency (MAF) > 0.01. The genomic control inflation factor (λ_{GC}) was calculated for each bile acid trait. Cohort-level λ_{GC} overall ranged from 0.9 to 1.1 for quantitative bile acid traits, both imputed and not, suggesting little residual influence of population stratification and family structure (Supplementary Table 12). In a few cases, ERF cohort reported somewhat deflated λ_{GC} (GCDCA at 0.884 and GLCA at 0.899). On the other hand, there was considerable inflation for binary bile acid in the case of NSPHS (Supplementary Table 12), with values of λ_{GC} above 1.1, suggesting that population structure/cryptic relatedness was not fully controlled for these traits in the NSPHS cohort.

Meta-analysis

Prior to meta-analysis, cohort-level GWAS were quality controlled using the EasyQC software package, following the protocol described in Winkler *et al*.[300] Cohort-level results were corrected for the genomic control inflation factor, then pooled and analysed with METAL v2011-03-25 software[196], applying the fixed-effect inverse-variance method. The mean genomic control inflation factor after the meta-analysis was 0.991 (range 0.938 – 1.009), suggesting that the confounding effects of the family structure were correctly accounted for (Supplementary Table 12). The standard genome-wide significance threshold was Bonferroni corrected for the number of independent bile acid traits, calculated as 14 ($5x10^{-8}/14 = 3.57x10^{-9}$). The number of independent bile acid traits was estimated as the sum of the number of binary traits (4) and the number of principal components that jointly explained 99% of the total variance of log_{10}-transformed quantitative traits in each cohort (10) (Supplementary Table 13).

Sex-stratified GWAS meta-analysis

To identify possible differences in the genetic contribution to bile acid variability between men and women, we performed sex-specific GWAS of the 14 quantitative bile acid traits for ORCADES and CROATIA-Vis cohorts. Given that for the sex-stratified GWAS we implicitly halve our sample size, we performed these analyses only on the imputed bile acid traits. The same analysis steps and procedures already described for the full meta-analysis were applied. Bile acid traits were adjusted for age, sex and batch as fixed effects, and relatedness (estimated as the kinship matrix calculated from genotyped data) as a random effect in a linear mixed model, calculated using the 'polygenic' function from the GenABEL R package[194]. Residuals of covariate and relatedness correction were tested for association with HRC-imputed[28] SNP dosages using the RegScan v0.5 software[195], applying an additive genetic model of association. Prior to meta-analysis, SNPs having a difference in allele frequency between the two cohorts higher than ±0.3 or a minor allele count (MAC) lower or equal to 6 were filtered out. Cohort-level GWAS were corrected for genomic control inflation factor and then meta-analysed (N =820 for male and N =1088 for female individuals) using METAL v2011-03-25 software[196], applying the fixed-effect inverse-variance method. The mean λ_{GC} was 0.993 (range 0.978–1.011) for male-specific meta-analysis and 0.996 (range 0.984–1.003) for female-specific meta-analysis. The Bonferroni-corrected significance threshold applied is 5×10^{-9}.

Phenoscanner and Mendelian Randomisation

To assess link between bile acids and diseases we explored the overlap of SNPs associated with BAs with complex human traits by using PhenoScanner v1.1 database[183,272], taking into account significant genetic association ($p < 5 \times 10^{-9}$) at the same or strongly (LD $r2 > 0.8$) linked SNPs in populations of European ancestry. We then performed bi-directional Mendelian Randomisation (MR) to investigate the effect of 548 complex traits and diseases available in the IEU Open GWAS database[209] (manually curated list of studies from identifiers ebi-a,

ieu-a, ieu-b and ukb-a; the complete list reported in the Supplementary Table 5) on BA levels, and vice-versa. The set of genome-wide significant, LD clumped SNPs used as instruments for complex traits/diseases was extracted from the selected studies by using the "extract_instruments" function from the TwoSampleMR 0.5.6 R package[208]. Similarly, sentinel SNPs from BAs meta-analysis (Supplementary Table 2) were selected as instruments. MR tests were performed by using fixed effects inverse variance-weighted (IVW) in case of multiple instruments or Wald Ratio method in case of a single instrument, as implemented in the TwoSampleMR 0.5.6 R package[208]. Multiple testing correction was controlled for using either the Bonferroni correction or false discovery rate (FDR).

Whole-exome sequencing data

Exome sequencing

The "Goldilocks" exome sequence data for ORCADES cohort was prepared at the Regeneron Genetics Center, following the protocol detailed in Van Hout *et al.*[217] for the UK Biobank whole-exome sequencing project. In summary, sequencing was performed using S2 flow cells on the Illumina NovaSeq 6000 platform with multiplexed samples. DNAnexus platform[250] was used for processing raw sequencing data. The files were converted to FASTQ format and aligned using the BWA-mem[251] to GRCh38 genome reference. The Picard tool[252] was used for identifying and flagging duplicated reads, followed by calling the genotypes for each individual sample using the WeCall variant caller[253]. During quality control, 33 samples genetically identified as duplicates, 3 samples showing disagreement between genetically determined and reported sex, 4 samples with high rates of heterozygosity or contamination, 2 samples having low sequence coverage (less than 80% of targeted bases achieving 20X coverage) and 1 being discordant with genotyping chip were excluded. Finally, the "Goldilocks" dataset was generated by (i) filtering out genotypes with read

depth lower than 7 reads, (ii) keeping variants having at least one heterozygous variant genotype with allele balance ratio greater than or equal to 15% (AB ≥ 0.15) or at least one homozygous variant genotype, and (iii) filtering out variants with more than 10% of missingness and HWE $p<10^{-6}$. Overall, a total of 2,090 ORCADES (820 male and 1,270 female) participants passed all exome sequence and genotype quality control thresholds. A pVCF file containing all samples passing quality control was then created using the GLnexus joint genotyping tool[254].

Variant annotation

Exome sequencing variants were annotated as described in Van Hout, *et al.*[217] Briefly, they were annotated with the most severe consequence across all protein-coding transcripts using SnpEff[255]. Gene regions were defined based on Ensembl release 85[301]. Predicted loss-of function (pLoF) variants were defined as variants annotated as start lost, stop gained/lost, splice donor/acceptor and frameshift. The deleteriousness of missense variants was based on dbNSFP 3.2[302,303] and assessed using the following algorithms: (1) SIFT[304]: “D” (Damaging), (2) Polyphen2_HDIV: “D” (Damaging) or “P” (Possibly damaging), (3) Polyphen2_HVAR[305]: “D” (Damaging) or “P” (Possibly damaging), (4) LRT[256]: “D” (Deleterious) and (5) MutationTaster[257]: “A” (Disease causing automatic) or “D” (Disease causing). If not predicted as deleterious by any of the algorithms the missense variants were considered “likely benign”, “possibly deleterious” if predicted as deleterious by at least one of the algorithms and “likely deleterious” if predicted as deleterious by all five algorithms.

Exome-wide gene-based aggregation analysis of rare variants

Generation of gene masks

For each gene, the variants were grouped into four categories (masks), based on severity of their functional consequence. The first mask (mask 1) included only pLoF variants. Masks 3 and 4 included both pLoF and variants predicted to be deleterious, by 5/5 algorithms (mask 3) or by at least one algorithm (mask 4). The most permissive mask (mask 2) included pLoF and all missense variants. These masks were then further split by the frequencies of the minor allele (MAF ≤ 5%, e.g. mask1_maf5; and MAF ≤ 1%, e.g. mask1_maf1), resulting in up to 8 burden tests for each gene (Supplementary Table 9).

ORCADES gene-based aggregation analysis

We performed variant Set Mixed Model Association Tests (SMMAT)[48] on the 18 bile acid traits from ORCADES cohort, quantified and pre-processed as previously described, fitting a GLMM adjusting for age, sex, batch, and familial or cryptic relatedness by kinship matrix. The kinship matrix was estimated from the genotyped data using the 'ibs' function from GenABEL R package[194]. The SMMAT framework includes 4 variant aggregate tests: burden test, sequence kernel association test (SKAT), SKAT-O and SMMAT-E, a hybrid test combining the burden test and SKAT. The 4 variant aggregate tests were performed on 8 different pools of genetic variants, called "masks", each one including a different set of variants based on both MAF and predicted consequence of variants (e.g., loss of function and missense) (Supplementary Table 9), as described above. Discovery significance threshold was Bonferroni corrected for the rough estimate of the number of genes in the human genome, 20,000, and the number of independent bile acid traits, 14, calculated as previously described ($0.05/20000/14 = 1.79 \times 10^{-7}$). A gene association was considered significant if it passed the above reported Bonferroni corrected significance threshold in at least two of the 4 performed variant aggregate tests and if the cumulative allele count of the variants included in the gene was equal or higher than 10.

4.3 Conclusion

As described in the above pre-print submission, I explored the genetic architecture of plasma bile acid levels, including both rare and common genetic variants. Using GWAS meta-analysis (N = 4923), I identified 2 significantly associated loci, near the *SLCO1B1* and *PRKG1* genes. I observed that variants in these genes have different impacts on bile acid levels in men and women. Performing gene-based aggregated tests on whole exome sequencing data (N = 1006), I identified rare and low frequency (MAF<5%) predicted loss-of-function (pLoF) and missense variants associated with bile acids, mapping to *OR1G1*, *SART1* and *SORCS2*. Based on the literature, I suggest possible biological mechanisms how these genes could influence bile acid levels. Finally, I tested for relationships between genetically increased levels of bile acids and complex traits/diseases risk and vice-versa, but found no significant associations. Since the sentinel SNP of the *SLCO1B1* locus has been identified by previous studies as associated with multiple traits (e.g. concentration of bilirubin, non-bile acid metabolites, mean corpuscular haemoglobin, sex hormone binding globulin, estrone conjugates, responses to various drugs), the current study may have insufficient statistical power to detect the effect of complex traits or diseases on bile acid levels. I discuss further the topic of statistical power and sample size in the next Chapter.

Chapter 5: Discussion

A deeper understanding of the genetic factors that contribute to complex traits and diseases is of great relevance for overall improvement of human health. In this thesis I explored the genetic architecture of protein glycomics and bile acid lipidomics, two types of omics that have been implicated in various complex diseases but have not been extensively researched. I successfully identified both common and rare association signals, which, for the most part, are located in genes with a clear biological link to the phenotypes of interests. Genetically isolated populations studied offered higher statistical power for discovery of rare and low frequency association signals, thanks to the increased allele frequency of certain variants compared to the general population. The statistical power was however not enough to assess the impact of glycans- and bile acids-associated variants on health-related traits and diseases, given their effect sizes, for which increased sample size is likely required.

5.1 Genetic regulation of transferrin and IgG glycome

In Chapter 2 I investigated genes and genetic variants influencing the transferrin glycome, and then compared those with the ones contributing to IgG glycosylation. Using the genome-wide association meta-analysis (N = 1890) for 35 transferrin glycan traits, I identified ten loci significantly associated with transferrin glycosylation, two of which (near *FOXI1* and *MSR1* genes) have not previously been associated with the glycosylation of any protein. The other eight loci (*MGAT5*, *TF*, *NXPE1*/*NXPE4*, *ST3GAL4*, *B3GAT1*, *HNF1A*, *FUT8* and *FUT6*) have been previously linked with the glycosylation of total plasma proteins and/or IgG[131–138] or to the prevalence of carbohydrate-deficient transferrin (CDT)[188], a measure that provides partial information on the sialylation status of transferrin. Several of the above-mentioned loci contain genes encoding glycotransferases, the enzymes catalysing the transfer of sugar molecules from donor molecules to acceptor proteins, resulting in the formation of glycosidic bonds.

I described rs6785596 as a “cis-glyQTL”, a genomic locus that explains variation in glycosylation levels and is local to the gene (*TF*) encoding the protein being glycosylated (transferrin). In this study, glycan traits were quantified as the percentage of the total transferrin N-glycome, in order to detect changes in glycosylation (relative abundance of individual glycan species in relation to the whole transferrin glycome), rather than the absolute amounts of specific glycans, which would be affected by changes in protein levels. Nevertheless, we assessed that variance of transferrin glycan traits associated at *TF* locus is partially (for TfGP3 and TfGP8 traits) or completely (for TfGP9 trait) explained by rs8177240, the strongest association with transferrin protein levels reported in GWAS catalog (p-value = $8x10^{-610}$)[178], which I used as a proxy for transferrin protein abundance (see Chapter 2 Supplementary Results, Supplementary Table 6 and 9 for further details). Interestingly, rs8177240 is a splicing QTL for the *TF* gene in liver (p-value = $5.9x10^{-25}$, GTEx v8[306]). Alternative splicing of the *TF* gene can generate isoforms of the transferrin protein, which may differ in the presence or location of glycosite motifs[307,308]. Composition and abundance of glycans on the transferrin molecule may thus vary between protein isoforms by the effect of rs8177240, a variant in control transferrin splicing, in addition to the glyQTL rs6785596.

Transferrin has two N-linked disialylated biantennary oligosaccharide chains, followed by other minor isoforms which vary depending on the number of oligosaccharide chains[309]. Hyposialylation or desialylation of transferrin produces altered isoform patterns, which are used as diagnostic marker for alcohol abuse[310,311] and glycosylation defects like congenital disorders of glycosylation[312]. Transferrin isoforms thus differ in their glycosylation patterns, meaning that the composition and abundance of glycans on the transferrin molecule can vary between isoforms. Unfortunately, HILIC-UHPLC, the glycosylation measuring approach used in this study, separates all glycan structures from the carrying proteins, losing the information about which glycan structures were attached to which transferrin molecule or isoform. The analysis of intact glycoprotein would enable in this case the characterisation of transferrin isoforms and their glycosylation patters; however, this technique currently faces

limitations in sensitivity and glycoform resolution and is thus not routinely used to characterise large cohorts[108].

I also performed genome-wide association meta-analysis (N = 2020) for 24 IgG N-glycan traits and compared the results to transferrin glycan-associated loci. Of the 7 genomic regions associated with transferrin or IgG glycome and containing glycosyltransferase coding genes, 3 are unique for transferrin glycosylation (*MGAT5, ST3GAL4, B3GAT1*), 2 are IgG glycan-specific (*ST6GAL1*, *MGAT3*), while other 2 are associated with the glycosylation of both IgG and transferrin (*FUT8* and *FUT6*). Therefore, while some glycotransferase enzymes seem to be protein-specific, other instead are able to glycosylate both transferrin and IgG. To investigate how the same glycosyltransferase enzymes are genetically regulated in different proteins, I focused on the shared associations at *FUT8* and *FUT6* genes. Colocalisation analysis suggests that association patterns of transferrin and IgG glycan traits at *FUT8* and *FUT6* loci are driven by independent causal variants, at both genomic regions. Therefore, while the same glycosyltransferase enzymes are involved in glycosylation of both transferrin and IgG, the process is independently regulated by protein-specific causal variants.

Next, I propose a biological mechanism by which independent genetic variants in the *FUT8* region could have a protein-specific effect on the transferrin and IgG glycomes. Since these protein-specific variants are located in the regulatory region of the genes, and are not in strong LD with coding variants from the enzymes' active site, I suggest that they could affect the expression of other enzymes, such as transcription factors, in different tissues. The majority of the IgG found in blood plasma is produced by bone marrow plasma cells, the fully differentiated form of B-cells[186]. On the other hand, transferrin protein found in blood plasma is mostly produced by liver hepatocytes[185]. The GWAS meta-analyses performed in this study showed that the glycomes of transferrin and IgG are associated with variants from genetic regions coding for two different transcription factors, HNF1A and IKZF1. The gene encoding transcription factor HNF1A is mainly expressed in the liver hepatocytes, while the gene coding for

IKZF1 is mainly expressed in plasma cells. Similarly, *TF*, the gene coding for transferrin protein, is most highly expressed in the liver hepatocytes, while *IGHG1*, gene encoding the constant region of immunoglobulin heavy chains, is most highly expressed in plasma cells. *IKZF1* has been functionally validated as a key regulator of IgG glycosylation by Klarić and colleagues[135]: knock-down of *IKZF1* resulted in change in expression of the *FUT8* gene and further on in change in levels of fucosylation of IgG. On the other hand, using the Regulatory Sequence Analysis Tool (RSAT)[182], I identified a potential binding site for the HNF1A transcription factor, spanning the sentinel SNP of the transferrin-glycome association in the *FUT8* locus. While HNF1A's role in transferrin glycosylation has yet to be experimentally determined, this transcription factor was shown to regulate the expression of the *FUT8* and *FUT6* genes in the HepG2 hepatocyte cell line, with *HNF1A* knockdown resulting in downregulation of *FUT6* and upregulation of *FUT8*[132]. Overall, I suggest that the two different causal variants may affect the binding of different transcription factors, HNF1A and IKZF1, in different tissues, liver hepatocytes and B-cells, and therefore have independent effects on the glycosylation of transferrin and IgG proteins. However, the effect of specific SNPs on the binding of the two transcription factors and their downstream impact on the expression of fucosyltransferase enzymes in a tissue-specific fashion still needs to be functionally validated, using for example Chromatin immunoprecipitation (ChIP) assay or CRISPR-Cas9 genome editing. The ChIP assay allows for the identification and quantification of protein-DNA interactions in vivo. It involves the cross-linking of DNA and proteins in cells, followed by the isolation of the protein-DNA complex, the amplification and sequencing of the DNA fragment containing the binding site[313]. This process can be used to determine whether variants in the *FUT8* gene may affect the putative binding of the HNF1A transcription factor. Similarly, it was employed by Klarić and collegues[135] to confirm the binding of IKZF1 transcription factor at a *FUT8* binding site in an IgG-secreting human B cell. Another way to test the effect of *FUT8* genetic variant(s) on the binding of HNF1A transcription binding is by causing the disruption or alteration of the transcription binding site using CRISPR-Cas9 genome editing, and then comparing the transcriptional activity of the edited

cells to cells with the wild-type sequence[314].In addition to HNF1A, *FUT8* variants associated with transferrin glycosylation might also be affecting the binding of the FOXI1 transcription factor. According to RSAT, I also identified a potential binding site for the FOX1 transcription factor in the *FUT8* region. Unlike HNF1A, FOXI1's involvement in regulating gene expression of fucosylation genes is to date undocumented and requires functional validation.

Finally, using a framework incorporating Mendelian randomisation and colocalisation analysis, I assessed the effect of transferrin glycosylation on biochemical measurements and diseases, and the reverse. I found no evidence of complex traits/diseases influencing transferrin glycome, while I identified an effect of transferrin glycans on levels of C-reactive protein, LDL and total cholesterol at the *HNF1A* locus. However, these MR results rely on a single instrumental variable (or two maximum) and were driven by associations in a single locus, so should be interpreted with caution.

5.2 Rare and low frequency variants contributing to variation of protein glycome

Building on the work carried out in Chapter 2, where I described similarities and differences in the genes and variants contributing to the glycome of transferrin and IgG proteins, in chapter 3 I further investigated the genetic architecture of protein glycome by focusing on rare and low frequency variants affecting the transferrin and IgG glycosylation. By performing exome-wide gene-based tests across 51 transferrin traits (CROATIA-Korcula N = 948, VIKING N = 959) and 94 IgG glycan traits (ORCADES N = 1960, CROATIA-Korcula N = 1866, VIKING N = 1086), testing low frequency and rare (MAF <5%) predicted loss-of-function (pLoF) and missense variants, I identified 16 significant associations for the transferrin glycome (p-value < 8.06×10^{-8}) and 32 for the IgG glycome (p-value < 1.19×10^{-7}). Meta-analysis of IgG glycans for the ORCADES and VIKING cohorts added *FUT6* to the list of genes whose rare variants are significantly associated

with the IgG glycome. Similarly to my previous work (Chapter 2), most of the identified genes are associated with protein-specific glycans (transferrin - *TIRAP, MSR1* and *FOXI1*; IgG - *MGAT3, ST6GAL1* and *RFXAP*), while *FUT8* and *FUT6* are associated with glycosylation of both proteins. Almost all associated genes encode key enzymes of protein glycosylation (*MGAT3, ST6GAL1, FUT6, FUT8*) or have been previously associated with transferrin and IgG glycan traits in GWAS analysis (*MSR1, FOXI1*)[135,215], as seen in Chapter 2, with the exception of *TIRAP* and *RFXAP*, which have no previously known link to the glycome. The novel gene *TIRAP* is located in close proximity to *ST3GAL4*, a glycosyltransferase-coding gene which has been associated with the transferrin glycosylation[215]. The *TIRAP* rare variant rs8177399 is a splicing QTL for *ST3GAL4* in whole blood and may thus influence the transferrin glycome by controlling the splicing of the gene encoding ST3GAL4 sialyltransferase, an enzyme catalysing the transfer of sialic acid to a glycan structure. This proposed mechanism for *TIRAP's* influence on transferrin glycome is further corroborated by the fact the glycan associated with *TIRAP* rare variants is a derived trait expressing the percentage of trisialylated structures on the total glycome (Chapter 3 - Supplementary Table 15). *RFXAP* encodes the regulatory factor X-associated protein, part of the RFX DNA-binding complex, that binds to certain major histocompatibility (MHC) class II gene promoters and activates their transcription. MHC-II molecules are transmembrane proteins found on the surface of professional antigen-presenting cells, such as B cells[235]. B cells produce IgG antibodies and play a crucial role in the regulation and development of the immune response. The *HLA* super-locus, a gene-rich region counting at least 132 genes encoding MHC molecules, has been previously associated with IgG glycosylation[137]. Nevertheless, the precise role of *RFXAP* in IgG glycosylation still needs to be investigated, identifying for example signaling pathways or molecular networks associated with this gene to predict potential interactions and downstream effects, or performing a gene expression analysis, to profile gene expression changes upon overexpression of *RFXAP*.

Coupling gene-based strategies, aggregating the contribution of multiple rare variants, with genetically isolated populations, having otherwise rare variants at

higher frequency due to genetic drift, I identified rare variant associations with the protein glycome independent from single-variant GWAS or ExWAS association signals. After conditioning the aggregated rare variant association tests on sentinel SNPs from GWAS on imputed genotypes, 24 of the 32 IgG glycome-gene aggregate pairs remained significant. Conversely, 14 out of the 16 transferrin glycome-gene aggregate pairs failed to reach genome-wide significance after adjusting for GWAS sentinel SNPs, suggesting that a considerable part of the rare variant signal for transferrin glycosylation was dependent on variants identifiable by GWAS. More specifically, gene-based associations with rare variants from *FUT8*, *ST6GAL1* and *MGAT3* remained significant following the conditioning on the GWAS sentinel variants and no significant associations were found for the *RFXAP* locus in the GWAS analysis. On the other hand, gene-based associations at *FUT6*, *MSR1*, *FOXI1* and *TIRAP* genes were explained by sentinel GWAS variants. It is worth noting that the GWAS sentinels of *FUT6*, *MSR1* and *FOXI1* are low frequency (0.02<MAF<0.05) variants. Next, to test whether the rare variant glycome associations were driven by a single variant detectable by a single-point ExWAS, or rather by a group of multiple rare variants, I conditioned the aggregated rare variant association tests on sentinel ExWAS associations. Compared to the GWAS, ExWAS analysis can uncover single-variant associations with variants that are too rare to be imputed accurately and therefore could not be detected using a GWAS of imputed data. With this analysis I discovered that aggregated association of variants with IgG glycosylation in the *ST6GAL1* gene was driven by a single variant that is sufficiently rare to not be available in the imputed data. On the other hand, associations at *FUT8* and *MGAT3* genes remained significant after conditioning on the sentinel ExWAS variants. In the case of these two genes, the sentinel ExWAS variants were common, suggesting that both common and aggregates of multiple rare variants in these two genes independently contribute to protein glycosylation.

Finally, I assessed the effect of glycome-associated rare variants on diseases by performing gene-based association tests of 116 quantitative health-related traits

available in our cohorts. While I did not find any significant associations with health-related traits, by querying public repositories of gene-based tests[229,230] and Open Targets Genetics portal[315,316], I discovered a potential pleiotropic effect of the stop gained rs41341748 variant from *MSR1* on transferrin glycosylation, galectin-3 and insulin-like growth factor 1 levels. In turn, both galectin-3-binding protein measurement[243–247] and insulin-like growth factor 1 levels[81,241,242] have been associated with type 2 diabetes. Nevertheless, the underlying mechanism behind a potential connection between transferrin glycome, IGF1, galectin-3 and diabetes is currently speculative and requires to be further explored, likely with a larger sample size. By searching on Phenoscanner database[183], I did not find any overlap between loci associated with transferrin glycosylation and those associated with diabetes, insulin resistance or HbA1c levels (Supplementary Table 11a of Chapter 2). Further, including HbA1c levels as an additional covariate in the transferrin glycome GWAS did not influenced the results of VIKING cohort, since the association effect sizes and p-values obtained for the HbA1c-adjusted GWAS are very similar to those obtained in the original GWAS. Accordingly, I could not find any evidence suggesting that prediabetes, insulin resistance or HbA1c levels might have an effect on transferrin levels mediated by common genetic variation. Although the use of isolate cohort samples allowed for the identification of genetic associations with intermediate phenotypes like glycomics, statistical power was not enough to detect associations with complex traits and diseases, especially for variants of low frequency. While this analysis, at the current sample size and given the effect size of associated variants, lacked statistical power to assess the impact of glycome-associated rare variants on health-related traits, the combined use of intermediate phenotypes, rare-variants aggregating tests and isolated populations is still a strategy worth exploring to achieve this goal. Drastically larger sample sizes from the general population may in fact still lack power for discovery of rare variants contributing to complex diseases. A recent study[36] performed exome-wide gene-based analyses in UK Biobank to evaluate the contributions of ultra-rare (MAF<0.1%) damaging variants to several health-related traits and diseases. Despite the 100-fold higher sample size compared to our study, the authors noted that rare variant discovery

power was still limited. For example, only three associations for type 2 diabetes and one for atrial fibrillation were significant, despite both phenotypes counting over 12,000 cases. A previous exome sequencing study of type 2 diabetes argued that rare variant gene-level signals are likely distributed across numerous genes, with the vast majority of them having extremely small effects on the disease[317]. Accordingly, authors estimated that 75,000–185,000 sequenced cases, or 600,000-1,275,000 samples from population-based biobanks, would be necessary to identify known diabetes drug targets at 80% statistical power. An alternative to such large sample sizes is the use of genetic isolates, which are advantageously characterised by different allele frequency and disease prevalence than the general population. For example, in this study I detected two instances of isolate-specific glycome associations that are driven by variants increased in frequency compared to the general population. The rs750567016 variant in *ST6GAL1*, affecting IgG glycosylation, is over 300 times more common in ORCADES (MAF=3.3×10^{-3}) than in UK Biobank (MAF=1.0×10^{-5}) or gnomAD (MAF=9.0×10^{-6}), and is absent from the CROATIA-Korcula and VIKING cohorts. The rs115399307 variant in *FOXI1*, associated with transferrin glycosylation, is seven times more common in VIKING (MAF=2.1×10^{-2}) than in the CROATIA-Korcula cohort (MAF=2.7×10^{-3}), UK Biobank (MAF=8.5×10^{-3}) and gnomAD (MAF=7.1×10^{-3}). These findings suggest that, for some specific variants, genetic isolates may retain higher statistical power to identify rare variant associations than large population-based biobanks, despite the limited sample size. The increased statistical power offered by health-related, quantitative traits compared to binary disease endpoint traits can be noted also in a recent exome sequencing study exploring the impact of rare protein-altering variants on health in 454,787 UK Biobank participants[39]. Eighty gene-based associations were significantly identified by testing 3702 binary traits, while 484 gene-based associations were observed by testing only 292 quantitative traits. Accordingly, carefully selected intermediate quantitative phenotypes represent a viable option to increase statistical power for detecting rare variant associations with complex diseases even in UK biobank, one of the largest WES cohorts currently available worldwide. Viking Genes, a family-based cohort including individuals from the

ORCADES and VIKING cohorts here studied, has now reached over 9000 participants from the Scottish Isles. We may therefore have the opportunity in the future to test, in a genetic isolate of a larger size, the efficacy of intermediate phenotypes and rare variant aggregating tests in identifying rare variants affecting impacting health and complex diseases.

While WES rare variant studies are a powerful tool for identifying genetic variant associations, they may be prone to false positives and errors due to several factors inherent to the WES technology and analysis process. To obtain whole exome sequences, genomic DNA extracted from participant samples is fragmented into smaller segments, to which adapters are attached to facilitate the subsequent steps. The exonic regions of interest are selectively enriched using capture probes and amplificated by polymerase chain reaction (PCR)[318]. The WES design used in Chapter 3 and Chapter 4 studies is set to target approximately 39 Mb of the human genome and an additional 100 bp flanking region upstream and downstream of each capture target are also included[217]. The amplified DNA libraries are then loaded onto a high-throughput sequencing platform, such as the Illumina platform. Sequencing is followed by variant calling, aimed at identifying differences between an individual's sequenced DNA and a reference genome. It is important to note that certain rare variants, especially singletons, raise concerns about potential sequencing errors. To mitigate this, variants were called using joint calling with multiple cohorts (ORCADES, VIKING and CROATIA-Korcula), thereby increasing confidence in their accuracy. To limit the possibility of false positives in gene-based aggregation tests of Chapters 3 and 4, I considered as not reliable, and thus discarded from the presented significant results, genes whose tested variants have a cumulative allele count lower than 10.

To further increase statistical power and to facilitate biological interpretation, the current studies have been limited to the analysis of variants impacting the gene product, namely variants which were annotated *in silico* as pLoF or missense. While detailed molecular studies are considered the gold standard for confirming

variant function, several bioinformatics tools, as SnpEff[255], SIFT[319], PolyPhen[320], which were used in this study, and VEP[180], employed in the previous chapter, have been developed for rapidly classifying and prioritizing candidate variants, reducing the need for expensive and labour-intensive functional assays. However, these methods have inherent limitations due to differences in assumptions, reference/training datasets and alignment algorithms, which may lead to incorrect conclusions.

It has been observed that pLoF variants, given their rarity, often result to be sequencing errors rather than true genetic variants, thus not actually resulting in a loss of protein function[321]. A common example of this phenomenon are nonsense variants in the final exon of a gene, causing a premature termination codon. The impact of most of these end truncations, due to their position in the aminoacidic chain, is however not as deleterious as complete LoF, where the protein is not produced altogether. In other cases, LoF variants are unexpectedly observed in apparently healthy individuals. One proposed explanation for this paradox involves alternative splicing of mRNA, which allows exons of a gene to be expressed at varying levels across different cell types[322]. pLoF variants in weakly expressed regions have been observed to have similar effect sizes to those of synonymous variants, whereas pLoF variants in highly expressed exons are most strongly enriched among cases. The great majority of annotation tools do not systematically incorporate information about exon expression into the interpretation of variants and, in case of multiple protein-coding transcripts, variants are usually labelled with the most severe consequence[217]. Finally, variant impact scores, as SIFT, PolyPhen etc, are generated using algorithms considering multiple variant features to predict their impact or deleteriousness. These classifiers require training sets that identify the precise set and combination of input features associated with "deleterious" alleles. The choice of data used as the training set is thus crucial, as it determines the type of deleterious variants that can be accurately predicted. Accordingly, each algorithm, based on the feature set used, has its own strengths, weaknesses, and caveats[323,324].

While the study presented in Chapters 3 and 4 of this thesis serves as a first exploratory analysis of rare, protein-coding variants affecting the levels of transferrin/IgG glycans and bile acids, additional statistical tools can be applied in the future to confirm the annotation of associated variants and gain deeper insight in their functional role. Common types of LoF annotation errors can be identified for example by the Loss-Of-Function Transcript Effect Estimator, or LOFTEE, which applies a conservative filtering strategy to distinguishes high-confidence pLoF variants from annotation artefacts[325]. For genes where transcript expression differs between exons, the "proportion expressed across transcripts", or pext, metric can be used to differentiate between weakly and highly expressed exons, and thus flag variants which may be less likely to be pathogenic[322]. Finally, experimental methods, such as measuring the effect of the variant on protein function in vitro or in vivo, or population-based methods, such as comparing the frequency of the variant in cases versus controls to determine its association with disease, can be applied to confirm the impact of glycan-associated variants on complex traits or disease[326].

5.3 Genetic architecture of bile acid lipidome

After focusing on the genetic architecture of the protein glycome, in Chapter 4 I performed similar analyses for a different "ome", the bile acid lipidome. By performing GWAS meta-analysis of blood plasma levels for 18 primary and secondary bile acids in five cohorts of European descent (CROATIA-Vis N = 971, ORCADES N = 1019, NSPHS N = 718, MICROS N = 1336, ERF N = 879), I identified 2 significantly associated loci, near the *SLCO1B1* and *PRKG1* genes. While direction of the effect for the *SLCO1B1* locus is consistent across 4 of the 5 cohorts tested, the sentinel SNP of *PRKG1* locus passed the MAF threshold (MAF > 1%) only in the CROATIA-Vis cohort, which is therefore the only cohort contributing to this association.

Since the rate of bile acids synthesis and their pool composition are known to be sexually dimorphic[261], I performed sex-specific GWAS meta-analysis of bile acid traits in two of the available cohorts (female N = 1088, male N = 820), observing sex-specific associations for the 2 loci identified in the pooled analysis containing all samples. The bile acid association signal at the *SLCO1B1* locus, significant in the male-only analysis, has a smaller effect size and does not reach the significance threshold in the female-only analysis, despite the slightly larger sample size. On the contrary, the association signal at the *PRKG1* locus has a larger effect in females than in males and passes the significance threshold only in the female-specific analysis. Interestingly, the *PRKG1* sentinel SNPs from the pooled-analysis and for the female-specific analysis are in linkage equilibrium and thus represent two independent associations at the locus. Nevertheless, Mendelian randomisation analysis did not provide evidence that testosterone, oestradiol, sex hormone-binding globulin or other sex-related traits have causal effects on levels of plasma bile acids. While this could be due to a lack of statistical power of the bile acids meta-analysis, there is currently no evidence to suggest an effect of sex-related hormones on bile acid levels mediated by genetics. In addition to *SLCO1B1* and *PRKG1*, I identified another 13 sex-specific association signals, suggesting that the genetic effect on the plasma levels of bile acids at these loci is different in men and women. These, however, need to be further validated in other cohorts, also testing a genotype x sex interaction model (G x S), which is a statistical model helpful to investigate whether the effect of genetic variation on a trait is influenced by the sex of the individual. The interaction term that combines genotype and sex can capture whether the effect of the genotype on the trait differs between males and females, providing thus insights into the complex interplay between genetics and biology in the context of sexual dimorphism[327].

I next performed exome-wide gene-based association tests in one of the cohorts with available exome sequencing data (N = 1006), and identified associations of bile acids with rare pLOF and missense variants in the *OR1G1*, *SART1* and *SORCS2* genes. The literature suggests that olfactory cues, recognised by

olfactory receptors such as the one encoded by *OR1G1*, can stimulate the release of serotonin by specialised enteroendocrine gastrointestinal cells[285], which has been shown to increase bile acid synthesis and secretion in mice[287], offering a potential link between the *OR1G1* and bile acid levels. The hypoxia-associated factor (HAF), encoded by *SART1*, has been shown to promote the transcription of HIF-2α[283], a hypoxia-inducible factor (HIF) transcription factor involved in proliferation and hypoxia-related signalling. Increased expression of HIF-2α has been reported in several liver diseases, such as non-alcoholic steatohepatitis[328] and hepatocellular carcinoma[329], characterised by impaired levels of plasma bile acids[330,331]. *SORCS2* encodes a receptor for the precursor of nerve growth factor, up-regulation of which has been reported for several liver pathologies[288–293]. As observed for the protein glycome and described in Chapter 3, several rare variants found associated with bile acid lipidome have increased frequency in the tested isolate cohort (ORCADES) compared to the general population. For example, *OR1G1* rs777878604 is ~75-fold more common in ORCADES (MAF=6.23×10^{-3}) than in gnomAD (MAF=1.47×10^{-5}), and *SORCS2* rs777878604 is over 400-fold increased in frequency in ORCADES (MAF=4.5×10^{-3}) compared to gnomAD (MAF=5.88×10^{-5}). Once again, the increase in allele frequency of certain genetic variants in isolated populations can increase power for discovery of rare variant associations.

Finally, I used Mendelian randomisation to assess relationships between genetically increased levels of bile acids and biochemical measurements/risk for diseases, but did not find any bile acids affecting these traits, nor the reverse. All of these results, however, might be affected by the lack of statistical power.

5.4 Similarities and differences in genetic regulation of protein glycome and bile acid lipidome

By performing similar genetic analyses on two different omics datasets I am able to compare the genetic architecture of the two intermediate phenotypes. First, despite the bile acid lipidome analyses having double the sample size compared to the glycosylation-based analyses, using GWAS I discovered only 2 genes significantly associated with bile acid levels, compared to 10 and 11 genes associated with transferrin and IgG glycosylation. In contrast to the GWAS analyses, for the gene-based aggregation of rare-variant analyses the sample size of glycosylation analysis was at least double that of the bile acid analyses. However, I have detected a similar number of associations across all phenotypes - aggregated rare variants from 5 genes were associated with transferrin and 4 with IgG glycosylation levels, and 3 genes with bile acid levels.

For both omics, the majority of associated loci contain genes with a clear biological function in the context of the studied phenotype. However, there are some differences between two phenotypes. In particular, the majority of glycome-associated loci contain genes encoding glycotransferases, key enzymes in protein glycosylation, namely, the acetylglucosaminyltransferases MGAT5 and MGAT3, glucuronyltransferases B3GAT1, sialyltransferases ST3GAL4 and ST6GAL1, and the fucosyltransferases FUT6 and FUT8. Another example of a glycome-associated gene having a clear biological link to the phenotype of interest is the *TF* gene, associated with transferrin glycosylation and encoding transferrin protein. The observation for IgG glycosylation was similar, with variants in the gene coding for the heavy chain of immunoglobulin G were also associated with glycosylation of this protein[135,137]. In addition to glycosyltransferases and transferrin protein coding genes, another “class” of genes found associated with the protein glycome are transcription factor-coding genes, namely *HNF1A* and *FOXI1* for transferrin, *IKZF1* and *RUNX3* for the IgG glycome. In Chapter 2, I showed that the binding of both FOXI1 and HNF1A transcription factors might be affected by transferrin glycome-associated variants

in the *FUT8* locus. Lauc *et al.*[132] have shown that *HNF1A* knockdown results in downregulation of *FUT6* and upregulation of *FUT8* in HepG2 hepatocyte cell line. While it might be expected that a change in levels of fucosyltransferase enzymes FUT6 and FUT8 would impact levels of antennary and core fucosylation, this link, especially in the context of transferrin glycosylation, has yet to be experimentally proven. Similarly, a possible involvement of FOXI1 in the regulation of the transferrin fucosylation is to date unknown and would require functional validation. In the GWAS of IgG glycosylation, Klarić *et al.*[135] suggest that IgG glycome-associated SNPs from the *MGAT3* locus disrupt the binding site for RUNX3 transcription factor and are pleiotropic with *MGAT3* expression in lymphoblastoid cells. Further, Klarić *et al.*[135] provided experimental evidence that IKZF1 transcriptionally regulates *FUT8*, showing, first, that IKZF1 binds to regulatory regions of *FUT8* and, second, that *IKZF1* knockdown results in increased *FUT8* expression and increased IgG fucosylation.

Overall, genetic analysis of protein glycome identified association at genes encoding key enzymes of glycan synthesis and transcription factors (potentially) regulating them. On the contrary, no association signals in genes encoding key enzymes of bile acid synthesis, such as *CYP7A1* and *CYP7B148*, were identified in Chapter 4 analyses. The strongest GWAS association with bile acid lipidome is in *SLCO1B1*, whose encoded protein is a well-known hepatocyte transporter mediating the uptake of various endogenous compounds, such as bile salts, bilirubin glucuronides, thyroid hormones and steroid hormone metabolites, and drugs, like statins, HIV protease inhibitors, and the anti-cancer agents irinotecan or methotrexate. Accordingly, the knock-out of the *SLCO1B1* gene in mice results in abnormal liver physiology, abnormal xenobiotic pharmacokinetic phenotypes and altered blood chemistry (e.g. abnormal level of circulating serum albumin, decreased circulating serum amylase, increased serum alanine transaminase, bilirubin and cholesterol)[332]. Furthermore, the *SLCO1B1* sentinel SNP rs4149056, for which I observed a sex-specific effect on plasma bile acid levels, has been reported to impact the pharmacokinetics of pravastatin differently in men and women[333]. *PRKG1*, another gene whose variants I found associated

with the bile acid lipidome, encodes the protein kinase cGKI, which Franko *et al.*[334] reported to regulate the activation of hepatic stellate cells, liver-specific mesenchymal cells that play vital roles in liver physiology and fibrogenesis. Activation of hepatic stellate cells is characteristic of hepatic fibrogenesis, a process occurring during chronic liver injury and that is primarily involved in the progression of chronic liver diseases, irrespective of their specific aetiology[335]. Nevertheless, the mechanism of how the variation within the *PRKG1* gene relates to bile acid levels requires further investigation. Finally, also the products of *SART1* and *SORCS2*, whose rare variants I found associated with bile acid lipidome, have been linked indirectly to several liver pathologies, e.g. non-alcoholic steatohepatitis[328], hepatocellular carcinoma[329], cirrhosis[291] and hepatocellular carcinoma[292].

Overall, genetic analysis of the bile acid lipidome identified association signals at genes involved in hepatic function and dysfunction, rather than genes encoding key anabolic or catabolic enzymes of the pathway, as observed instead for protein glycome. It is important to interpret these findings in the context of the tissue in which bile acid levels were measured, blood plasma. Bile acids are synthesised in the liver and secreted into the intestine, to be then reabsorbed into the bloodstream and returned to the liver. Bile acid levels in plasma thus reflect the amount of bile acid escaping extraction from the bloodstream and returning to the liver. In healthy subjects, the fasting total serum bile acid concentration is relatively low (2-10 μmol/L), thanks to the efficiency of the liver in removing bile acids from the portal-hepatic circulation. Increase in serum total bile acids could be mostly related to a dysfunction in the hepatocellular uptake of bile acid, impairment of transport of bile acids in hepatocytes, or dysfunction of bile acid efflux[336]. The exploration of BAs' genetic regulation could thus potentially benefit from examining tissues beyond blood plasma. Alternative tissues, such as the liver, offering insights into BAs synthesis and metabolism, along with the gallbladder and small intestine, where BAs significantly contribute to digestion and nutrient absorption, might yield more pertinent information. Regrettably,

obtaining samples on a large scale from these human tissues poses practical challenges, which is not the case for blood samples.

5.5 Future work

Since the first GWAS of the protein glycome[132], several methodological improvements have been applied in order to increase statistical power for discovery of genetic associations. First, the use of glycome profiling technologies specifically tailored for large scale population studies (e.g. ultra-performance liquid chromatography), characterised by high sensitivity, resolution and speed, and also able to provide branch-specific information of glycan structures[337]. Then, by employing new imputation panels, like 1000 Genomes[338] and Haplotype Refence Consortium (HRC)[28], or exome sequence data to increase the resolution and power of genetic mapping. Finally, in the most recent GWAS of IgG glycosylation, Frkatović-Hodžić *et al*.[339] increased the available sample size by developing a protocol for harmonisation of glycan data generated using different analytical platforms, ultra-performance liquid chromatography (UPLC) and liquid chromatography–mass spectrometry (LC-MS). IgG is composed of four subclasses (IgG1, IgG2, IgG3, and IgG4), which have distinct structural and functional characteristics[340]; LC-MS provides IgG subclass-specific glycan information, while UPLC measures the total IgG glycosylation. Glycan information obtained from these analytical platforms has not been previously combined, increasing the sample size to over 13,000 individuals in the discovery phase and more than 7,000 participants in the replication phase, the highest GWAS sample size to date for IgG glycosylation (in particular, the focus of this study is galactosylation, a trait measuring the presence of zero, one, or two galactose moieties in IgG glycan structures). Zaytseva *et al.*[341] recently performed large-scale investigation of possible causal relationships between IgG glycosylation traits and risk of 12 autoimmune, inflammatory, neurodegenerative, cardiovascular and cancer diseases, whose aberrant glycosylation profile is well

characterised. The authors were unable to detect any significant effects of IgG glycan traits on the risk of the diseases, due to the generally lower power of the IgG glycome GWAS (N=9000) compared with diseases GWAS (mean N=144,760, min N=8477, max N=462,013). The selected instrumental variables on average explained 5.3% of variance in IgG glycosylation traits, ranging from a minimum value of 0.67% to a maximum value of 20.06% variance explained. The authors estimated that a median required sample size of 25,500 for the IgG glycosylation GWAS would be needed to detect significant MR associations, if any exist. For example, the required sample size for the IgG glycome ranges from 9500 for effect on hypertension to 101,500 for effect on Parkinson's disease. They thus suggest to re-assay the effect of IgG glycome on diseases when an IgG glycome GWAS of $N \sim 25{,}500/2 = 12{,}750$ will be available, assuming that, in the future, available GWAS for diseases will count twice the sample size as that used in the current study[342]. While no similar calculations have been performed for glycome of non-IgG proteins, these data suggest that a reasonable increase in sample size will improve our chances of identifying the effect of protein glycome on complex diseases. As discussed in section 5.5, the protein glycome potentially represents a great example of intermediate phenotype to facilitate the discovery of biological mechanisms underlying complex diseases. Glycans are sufficiently close to their genetic substrate to allow the identification of biologically relevant genes even in cohorts of limited sample size, and changes in their pattern have been observed in several pathological states.

Similarly, identification of the genetic contribution to the bile acid lipidome and its effect on complex diseases has been limited by the available sample size. Bile acids are known to be largely influenced by environmental factors, such as sex, diet and gut microbiota, which can confound the genetic signal. Female sex and oestrogens are considered relevant regulators of bile acid production and composition, as seen for example in pregnancy[262] and menopause[294]. Also, species-composition of gut microbiota has a great impact on the bile acid lipidome, especially for secondary bile acids that are a direct result of microbial activity[295]. In a recent study on genetic and dietary determinants of bile acids, Li

et *al.*[267] used a genetically diverse population of ~360 mice to gain insight into the determinants of bile acid homeostasis, reducing the impact of confounding environmental factors typical of human studies. The authors found most bile acids to have a high heritability (h^2>0.5), indicating a strong genetic influence. However heritability of bile acids significantly dropped when ignoring dietary environmental factors. Since exhaustively controlling for the confounding effect of diet is a difficult task in large population-based cohorts, an increased sample size is necessary to further describe the genetic architecture of bile acid lipidome and its effect on complex diseases.

Research into the genetic determinants of the protein glycome has largely been limited to samples of European ancestry, as, to the best of my knowledge, all currently available population-based GWAS of the protein glycome have been performed in European cohorts. A study aimed at describing the inter-ethnic differences in serum glycome among US origin, South Indian, Japanese, and Ethiopian populations, found the Ethiopian ethnic group to exhibit a peculiar glycome pattern, with exclusive glycoforms, and greatly increased levels of both specific glycan structures and total serum glycome[343]. Interestingly, some of the glycoforms observed to be exclusive or elevated in healthy Ethiopian participants have previously been proposed as serum biomarkers of hepatocellular carcinoma in Japanese individuals[344] or as prognostic biomarkers in renal cell carcinoma[345], while some have been significantly associated with castration-resistant status in prostate cancer[346]. These findings highlight the importance of considering ethnic differences in variations of serum glycome, as ignoring these differences may lead to inaccurate and misleading conclusions when using glycans as disease biomarkers. A similar study[347], comparing IgG glycosylation patterns in individuals from 14 different countries and 25 different ethnic groups, found that indices describing a country's development level, expected lifespan and numerous health related indicators were positively correlated with levels of galactosylated IgG glycan structures. An MR study identified an effect of levels of IgG glycan traits on susceptibility of rheumatoid arthritis in a Chinese cohort[348], while a previous study in a large, well-powered cohort of European-descendent

concluded that loci associated with IgG glycosylation do not affect the risk of rheumatoid arthritis[349]. Authors suggest that this discrepancy in findings among the two studies might be due to ancestry-specific genetic effects. Therefore, more efforts need to be invested to expand the current analyses to more diverse cohorts.

Significant differences among different ethnic groups have been noted also for concentration of individual bile acid traits, with subjects of Asian descent displaying significantly higher concentrations of serum CDCA, TCDCA, GCDCA and GCA than subjects of other ethnic backgrounds[331]. Similarly, faecal concentrations of total bile acid have been reported to be significantly higher in Asian vegetarians than European-heritage vegetarians[350]. Accordingly, these different bile acid concentrations may be due to genetic differences in bile acid synthesis/metabolism or possibly different dietary patterns, noting also that individuals of Asian ancestry are more susceptible to certain liver diseases[351].

Protein glycome and the bile acid lipidome are influenced both by genetics and environmental factors. It is therefore important to establish which of the two is the main cause of observed differences in the glycan and bile acid profiles of different populations. Understanding to which extent genetics of glycome and bile acid lipidome possibly contribute to disease risk in an ethnic-specific manner is important both from the perspective of unravelling the underlying mechanisms and pathologies and for the discovery of new drugs and better treatments. Additionally, rather than searching for differences among populations, individuals from multiple ancestries can be analysed together in a trans-ethnic GWAS, which has increased power to detect true associations with the trait of interest, as shared genetic effects across ethnicities tend to be directionally consistent[352].

As discussed in Chapter 2 and 3, genes and variants regulating transferrin and IgG glycosylation are, for the most part, protein-specific. The current knowledge of genetic architecture of glycome is however limited to the IgG and transferrin proteins or to total plasma protein, quantifying the glycome of all proteins in

plasma, but without information on which glycan was bound to which protein. While genetic studies of the glycome of proteins other than IgG and transferrin would allow for identification of protein-specific glycosylation pathways, these are currently hampered by technical challenges in isolating the glycoproteins of interest in large cohorts. As N-glycome profiling technologies develop, the study of individual N-glycosylation profiles across a wider range of proteins will hopefully become possible in the close future. For example, fast, efficient, and robust high-throughput procedures have been developed for isolation of haptoglobin[353] and alpha-1-acid glycoprotein[354], acute phase proteins whose altered glycosylation has been described in different types of diseases. Additionally, high-throughput technologies for N-glycome profiling in human tissues other than blood are likely to emerge, allowing to assess how glycosylation is genetically regulated in tissue of primary interest for the protein studied (e.g. liver for transferrin). For example, site-specific glycosylation analysis of immunoglobulin A and G antibodies was reported for the first time in human saliva[355]. Finally, compared with studies of other complex traits, the sample sizes employed in genetic analyses of human N-glycomics remain relatively modest, causing limitations on the utilization of quantitative genetics methodologies. Anticipated future progress includes increased GWAS statistical power thanks to larger sample sizes, allowing more precise estimations of genetic impact and enhancing the reliability of quantitative genetics analyses such as MR, colocalisation and genetic correlation. These advancements are expected to enhance our understanding of N-glycosylation regulation in human proteins and shed light on the role of glycosylation in the development of glycome-associated diseases. This, in turn, could lead to the creation of new methods for predicting, preventing, diagnosing, and managing such diseases[356].

5.6 Conclusion

The main goal of this thesis was to expand the current knowledge about the genetic architecture of two under-studied omic traits, protein glycomics and bile acid lipidomics, and to explore their contribution to complex diseases. I first described the genetic regulation of the transferrin glycome and then compared it to that of IgG glycosylation, noting both protein-specific and shared associations. I found these shared associations to be likely regulated by independent causal variants, suggesting that glycosylation of transferrin and IgG is genetically regulated by both shared and protein-specific mechanisms. Next, I investigated rare pLOF and missense variants associated with the glycome of transferrin and IgG, increasing statistical power for discovery by using multiple gene-based aggregation tests in isolated populations. Overall, the protein glycome appeared to be mainly associated with genes encoding key enzymes of the glycosylation process. Finally, I applied a similar approach to bile acid lipidomics, exploring the genetic contribution of both common and rare variants and identifying also sex-specific association signals. Associated genes reflect hepatic function and dysfunction, rather than core enzymes for the synthesis of bile acids. Unfortunately, both the protein glycome and the bile acid lipidome analysis did not have enough statistical power to identify an effect on complex disease. Additional studies with larger sample sizes and of more diverse ancestries will be necessary to validate findings, to further unravel the genetic architecture of protein glycome and bile acid lipidome, and to understand their relationship with human diseases and complex traits. This in turn could contribute to understanding biological mechanisms underlying complex diseases, developing informed disease screening tests, improving disease diagnosis and prognosis, and finally designing innovative and more customised treatment strategies to enhance human health.

References

1. Mackay, T. F. C. The genetic architecture of quantitative traits. *Annual Review of Genetics* **35**, 303–339 (2001).
2. Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res.* **6**, 399–408 (2003).
3. Jelenkovic, A. *et al.* Genetic and environmental influences on adult human height across birth cohorts from 1886 to 1994. *Elife* **5**, 14 (2016).
4. Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet. 2022 543* **54**, 263–273 (2022).
5. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nat. 2022 6107933* **610**, 704–712 (2022).
6. Scriver, C. R. The PAH gene, phenylketonuria, and a paradigm shift. *Hum. Mutat.* **28**, 831–845 (2007).
7. Hillert, A. *et al.* The Genetic Landscape and Epidemiology of Phenylketonuria. *Am. J. Hum. Genet.* **107**, 234–250 (2020).
8. MacKay, T. F. Q & A: Genetic analysis of quantitative traits. *Journal of Biology* **8**, 1–5 (2009).
9. Timpson, N. J., Greenwood, C. M. T. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110–124 (2018).
10. Forgetta, V. *et al.* Rare Genetic Variants of Large Effect Influence Risk of Type 1 Diabetes. *Diabetes* **69**, 784–795 (2020).
11. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
12. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
13. Deaton, A. M. *et al.* Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *Sci. Reports 2021 111* **11**, 1–16 (2021).
14. Polfus, L. M. *et al.* Genetic discovery and risk characterization in type 2 diabetes across diverse populations. *Hum. Genet. Genomics Adv.* **2**, 100029 (2021).
15. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet. 2020 527* **52**, 680–691 (2020).
16. Cai, L. *et al.* Genome-wide association analysis of type 2 diabetes in the EPIC-InterAct study. *Sci. Data 2020 71* **7**, 1–6 (2020).
17. Mahajan, A. *et al.* Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet. 2022 545* **54**, 560–572 (2022).
18. Jiang, X. *et al.* Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. *Nat. Commun. 2018 91* **9**, 1–12 (2018).
19. Manousaki, D. *et al.* Genome-wide Association Study for Vitamin D Levels Reveals 69 Independent Loci. *Am. J. Hum. Genet.* **106**, 327–337 (2020).
20. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
21. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nat. 2021 6007890* **600**, 675–679 (2021).
22. Pilia, G. *et al.* Heritability of Cardiovascular and Personality Traits in 6,148

Sardinians. *PLOS Genet.* **2**, e132 (2006).
23. Shea, M. K. *et al.* Genetic and non-genetic correlates of vitamins K and D. *Eur. J. Clin. Nutr. 2009 634* **63**, 458–464 (2007).
24. Zheng, H. F. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526**, 112–117 (2015).
25. Bradfield, J. P. *et al.* A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.* **7**, (2011).
26. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet. 2012 449* **44**, 981–990 (2012).
27. Manousaki, D. *et al.* Low-Frequency Synonymous Coding Variation in CYP2R1 Has Large Effects on Vitamin D Levels and Risk of Multiple Sclerosis. *Am. J. Hum. Genet.* **101**, 227–238 (2017).
28. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016).
29. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nat. 2015 5267571* **526**, 82–90 (2015).
30. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Communications Biology* **2**, 1–11 (2019).
31. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5–23 (2014).
32. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 1–14 (2018).
33. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits : From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
34. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biology* **18**, 1–17 (2017).
35. Gibson, G. Rare and common variants: Twenty arguments. *Nature Reviews Genetics* **13**, 135–145 (2012).
36. Jurgens, S. J. *et al.* Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nat. Genet.* **54**, 240–250 (2022).
37. Sarnowski, C. *et al.* Impact of Rare and Common Genetic Variants on Diabetes Diagnosis by Hemoglobin A1c in Multi-Ancestry Cohorts: The Trans-Omics for Precision Medicine Program. *Am. J. Hum. Genet.* **105**, 706–718 (2019).
38. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186 (2017).
39. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, (2021).
40. Schwarze, K., Buchanan, J., Taylor, J. C. & Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine* **20**, 1122–1130 (2018).
41. Li, B. & Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
42. Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34**, 188–193 (2010).
43. Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7**, e1001322 (2011).
44. Basu, S. & Pan, W. Comparison of statistical tests for disease association with

rare variants. *Genet. Epidemiol.* **35**, 606–619 (2011).
45. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
46. 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).
47. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am. J. Hum. Genet.* **104**, 410–421 (2019).
48. Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *Am. J. Hum. Genet.* **104**, 260–274 (2019).
49. Scherag, A., Müller, H. H., Dempfle, A., Hebebrand, J. & Schäfer, H. Data adaptive interim modification of sample sizes for candidate-gene association studies. in *Human Heredity* **56**, 56–62 (Hum Hered, 2003).
50. Gordon, D., Levenstien, M. A., Finch, S. J. & Ott, J. Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies. *Pac. Symp. Biocomput.* 490–501 (2003). doi:10.1142/9789812776303_0046
51. Pfeiffer, R. M. & Gail, M. H. Sample size calculations for population- and family-based case-control association studies on marker genotypes. *Genet. Epidemiol.* **25**, 136–148 (2003).
52. Charlesworth, B. Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**, 195–205 (2009).
53. Otto, S. P. & Whitlock, M. C. The probability of fixation in populations of changing size. *Genetics* **146**, 723–733 (1997).
54. Andrews, C. Natural Selection, Genetic Drift, and Gene Flow Do Not Act in Isolation in Natural Populations | Learn Science at Scitable. *Nat. Educ.* **3**, (2017).
55. Halachev, M. *et al.* Increased ultra-rare variant load in an isolated Scottish population impacts exonic and regulatory regions. *PLOS Genet.* **15**, e1008480 (2019).
56. Service, S. K., Ophoff, R. A. & Freimer, N. B. The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum. Mol. Genet.* **10**, 545–551 (2001).
57. Devlin, B., Roeder, K., Otto, C., Tiobech, S. & Byerley, W. Genome-wide distribution of linkage disequilibrium in the population of Palau and its implications for gene flow in Remote Oceania. *Hum. Genet.* **108**, 521–528 (2001).
58. Colonna, V. *et al.* Small effective population size and genetic homogeneity in the Val Borbera isolate. *Eur. J. Hum. Genet. 2013 211* **21**, 89–94 (2012).
59. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
60. Charrow, J. Ashkenazi Jewish genetic disorders. *Fam. Cancer* **3**, 201–206 (2004).
61. Norio, R. Finnish Disease Heritage I: Characteristics, causes, background. *Human Genetics* **112**, 441–456 (2003).
62. Knowler, W. C., Bennett, P. H., Hamman, R. F. & Miller, M. Diabetes incidence and prevalence in pima indians: A 19-fold greater incidence than in rochester, minnesota. *Am. J. Epidemiol.* **108**, 497–505 (1978).

63. Baier, L. J. & Hanson, R. L. Genetic Studies of the Etiology of Type 2 Diabetes in Pima Indians: Hunting for Pieces to a Complicated Puzzle. *Diabetes* **53**, 1181–1186 (2004).
64. Dabelea, D. *et al.* Increasing prevalence of type II diabetes in American Indian children. in *Diabetologia* **41**, 904–910 (Diabetologia, 1998).
65. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. **13**, (2014).
66. Abbafati, C. *et al.* Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **396**, 1204–1222 (2020).
67. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
68. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
69. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).
70. Mahajan, A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes article. *Nat. Genet.* **50**, 559–571 (2018).
71. Sing, C. F., Haviland, M. B. & Reilly, S. L. Genetic architecture of common multifactorial diseases. *CIBA Found. Symp.* 211–232 (1996). doi:10.1002/9780470514887.ch12
72. Schork, N. J. Genetics of complex disease: Approaches, problems, and solutions. in *American Journal of Respiratory and Critical Care Medicine* **156**, (Am J Respir Crit Care Med, 1997).
73. Deschepper, C. F., Boutin-Ganache, I., Zahabi, A. & Jiang, Z. In search of cardiovascular candidate genes: Interactions between phenotypes and genotypes. in *Hypertension* **39**, 332–336 (Hypertension, 2002).
74. Carlsson, E., Groop, L. & Ridderstråle, M. Role of the FOXC2 -512C> T polymorphism in type 2 diabetes: Possible association with the dysmetabolic syndrome. *Int. J. Obes.* **29**, 268–274 (2005).
75. Andrulionytè, L., Zacharova, J., Chiasson, J. L. & Laakso, M. Common polymorphisms of the PPAR-gamma2 (Pro12Ala) and PGC-1alpha (Gly482Ser) genes are associated with the conversion from impaired glucose tolerance to type 2 diabetes in the STOP-NIDDM trial. *Diabetologia* **47**, 2176–2184 (2004).
76. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).
77. Surendran, P. *et al.* Rare and common genetic determinants of metabolic individuality and their effects on human health. *Nat Med* **28**, 2321–2332 (2022).
78. Gudjonsson, A. *et al.* A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nat. Commun.* **13**, 1–13 (2022).
79. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science (80-.).* **361**, 769–773 (2018).
80. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).
81. Pietzner, M. *et al.* Genetic architecture of host proteins interacting with SARS-CoV-2. *bioRxiv Prepr. Serv. Biol.* (2020). doi:10.1101/2020.07.01.182709
82. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).

83. Gilly, A. *et al.* Whole-genome sequencing analysis of the cardiometabolic proteome. *Nat. Commun.* **11**, 1–9 (2020).
84. Frayling, T. M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science (80-.).* **316**, 889–894 (2007).
85. Dina, C. *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat. Genet.* **39**, 724–726 (2007).
86. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
87. La Cognata, V., Morello, G. & Cavallaro, S. Omics data and their integrative analysis to support stratified medicine in neurodegenerative diseases. *International Journal of Molecular Sciences* **22**, 4820 (2021).
88. Ota, M. & Fujio, K. Multi-omics approach to precision medicine for immune-mediated diseases. *Inflamm. Regen.* **41**, 1–6 (2021).
89. Manchia, M. *et al.* The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases. *PLoS One* **8**, e76295 (2013).
90. Buyske, S., Yang, G., Matise, T. C. & Gordon, D. When a case is not a case: Effects of phenotype misclassification on power and sample size requirements for the transmission disequilibrium test with affected child trios. *Hum. Hered.* **67**, 287–292 (2009).
91. Edwards, B. J., Haynes, C., Levenstien, M. A., Finch, S. J. & Gordon, D. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genet.* **6**, 1–12 (2005).
92. Thornton-Wells, T. A., Moore, J. H. & Haines, J. L. Dissecting trait heterogeneity: A comparison of the clustering methods applied to genotypic data. *BMC Bioinformatics* **7**, 1–18 (2006).
93. Meyer-Lindenberg, A. & Weinberger, D. R. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience* **7**, 818–827 (2006).
94. Qian, W., Schweizer, T., Munoz, D. & Fischer, C. E. Misdiagnosis of Alzheimer's Disease: Inconsistencies Between Clinical Diagnosis and Neuropathological Confirmation. *Alzheimer's Dement.* **12**, P293–P293 (2016).
95. Singh, T. & Rajput, M. Misdiagnosis of Bipolar Disorder. *Psychiatry (Edgmont)* **3**, 57 (2005).
96. Solomon, A. J. *et al.* The contemporary spectrum of multiple sclerosis misdiagnosis. *Neurology* **87**, 1393–1399 (2016).
97. Yamada, R., Okada, D., Wang, J., Basak, T. & Koyama, S. Interpretation of omics data analyses. *Journal of Human Genetics* **66**, 93–102 (2021).
98. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nat. Rev. Genet.* **19**, 299–310 (2018).
99. Xu, Q., Yang, C. & Pei, Y. F. Editorial: Genetic Pleiotropy in Complex Traits and Diseases. *Front. Genet.* **13**, 859 (2022).
100. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).
101. Thanassoulis, G. & O'Donnell, C. J. Mendelian Randomization: Nature's Randomized Trial in the Post-Genome Era? *JAMA* **301**, 2386 (2009).
102. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of Causal Effects Using Instrumental Variables. *J. Am. Stat. Assoc.* **91**, 444–455 (1996).
103. Labrecque, J. & Swanson, S. A. Understanding the Assumptions Underlying

Instrumental Variable Analyses: a Brief Review of Falsification Strategies and Related Tools. *Curr. Epidemiol. Reports* **5**, 214–220 (2018).
104. Wallace, C. Statistical testing of shared genetic control for potentially related traits. *Genet. Epidemiol.* **37**, 802–813 (2013).
105. Zuber, V. *et al.* Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *Am. J. Hum. Genet.* **109**, 767 (2022).
106. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).
107. Varki, A. & Lowe, J. B. *Biological Roles of Glycans*. *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, 2009).
108. De Haan, N. *et al.* Developments and perspectives in high-throughput protein glycomics: enabling the analysis of thousands of samples. *Glycobiology* **32**, 651–663 (2022).
109. Trbojević-Akmačić, I., Vilaj, M. & Lauc, G. High-throughput analysis of immunoglobulin G glycosylation. *http://dx.doi.org/10.1080/14789450.2016.1174584* **13**, 523–534 (2016).
110. Trbojević-Akmačić, I. *et al.* High-Throughput Glycomic Methods. *Chem. Rev.* **122**, 15865–15913 (2022).
111. Agakova, A., Vučković, F., Klarić, L., Lauc, G. & Agakov, F. Automated integration of a UPLC glycomic profile. *Methods Mol. Biol.* **1503**, 217–233 (2017).
112. Rudd, P., Karlsson, N. G., Khoo, K.-H. & Packer, N. H. *Glycomics and Glycoproteomics*. *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, 2015). doi:10.1101/GLYCOBIOLOGY.3E.051
113. Lauc, G., Rudan, I., Campbell, H. & Rudd, P. M. Complex genetic regulation of protein glycosylation. *Mol. Biosyst.* **6**, 329–335 (2010).
114. Lauc, G., Vojta, A. & Zoldoš, V. Epigenetic regulation of glycosylation is the quantum mechanics of biology. *Biochimica et Biophysica Acta - General Subjects* **1840**, 65–70 (2014).
115. Wahl, A. *et al.* IgG glycosylation and DNA methylation are interconnected with smoking. *Biochim. Biophys. Acta - Gen. Subj.* **1862**, 637–648 (2018).
116. Freeze, H. H. Genetic defects in the human glycome. *Nat. Rev. Genet.* **7**, 537–551 (2006).
117. Clerc, F. *et al.* Human plasma protein N-glycosylation. *Glycoconj. J.* **33**, 309–343 (2016).
118. Kristic, J. *et al.* Glycans are a novel biomarker of chronological and biological ages. *Journals Gerontol. - Ser. A Biol. Sci. Med. Sci.* **69**, 779–789 (2014).
119. Russell, A. C. *et al.* The N-glycosylation of immunoglobulin G as a novel biomarker of Parkinson's disease. *Glycobiology* **27**, 501–510 (2017).
120. Trbojević-Akmačić, I. *et al.* Plasma N-glycome composition associates with chronic low back pain. *Biochim. Biophys. Acta - Gen. Subj.* **1862**, 2124–2133 (2018).
121. Gudelj, I. *et al.* Low galactosylation of IgG associates with higher risk for future diagnosis of rheumatoid arthritis during 10 years of follow-up. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1864**, 2034–2039 (2018).
122. Trbojević-Akmačić, I. *et al.* Inflammatory bowel disease associates with proinflammatory potential of the immunoglobulin G glycome. *Inflamm. Bowel Dis.* **21**, 1237–1247 (2015).
123. Rudman, N., Gornik, O. & Lauc, G. Altered N-glycosylation profiles as potential

biomarkers and drug targets in diabetes. *FEBS Letters* **593**, 1598–1615 (2019).
124. Munkley, J. & Elliott, D. J. Hallmarks of glycosylation in cancer. *Oncotarget* **7**, 35478–89 (2016).
125. Taniguchi, N. & Kizuka, Y. Glycans and Cancer: Role of N-Glycans in Cancer Biomarker, Progression and Metastasis, and Therapeutics. *Adv. Cancer Res.* **126**, 11–51 (2015).
126. Vajaria, B. N. & Patel, P. S. Glycosylation: a hallmark of cancer? *Glycoconjugate Journal* **34**, 147–156 (2017).
127. Rodríguez, E., Schetters, S. T. T. & Van Kooyk, Y. The tumour glyco-code as a novel immune checkpoint for immunotherapy. *Nat. Rev. Immunol.* **18**, 204–211 (2018).
128. Adamczyk, B., Tharmalingam, T. & Rudd, P. M. Glycans as cancer biomarkers. *Biochim. Biophys. Acta - Gen. Subj.* **1820**, 1347–1353 (2012).
129. Peng, W. *et al.* Clinical application of quantitative glycomics. *Expert Rev. Proteomics* **15**, 1007–1031 (2018).
130. Thanabalasingham, G. *et al.* Mutations in HNF1A result in marked alterations of plasma glycan profile. *Diabetes* **62**, 1329–1337 (2013).
131. Huffman, J. E. *et al.* Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults. *Hum. Mol. Genet.* **20**, 5000–5011 (2011).
132. Lauc, G. *et al.* Genomics meets glycomics-the first GWAS study of human N-Glycome identifies HNF1α as a master regulator of plasma protein fucosylation. *PLoS Genet.* **6**, e1001256–e1001256 (2010).
133. Sharapov, S. Z. *et al.* Defining the genetic control of human blood plasma N-glycome using genome-wide association study. *Hum. Mol. Genet.* **28**, 2062–2077 (2019).
134. Sharapov, S. Z. *et al.* Replication of 15 loci involved in human plasma protein N-glycosylation in 4802 samples from four cohorts. *Glycobiology* **31**, 82–88 (2021).
135. Klarić, L. *et al.* Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. *Sci. Adv.* **6**, eaax0301 (2020).
136. Lauc, G. *et al.* Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers. *PLoS Genet.* **9**, e1003225 (2013).
137. Shen, X. *et al.* Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation. *Nat. Commun.* **8**, 447 (2017).
138. Wahl, A. *et al.* Genome-Wide Association Study on Immunoglobulin G Glycosylation Patterns. *Front Immunol* **9**, 277 (2018).
139. Li, T. *et al.* Modulating IgG effector function by Fc glycan engineering. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 3485–3490 (2017).
140. Trbojević-Akmačić, I. *et al.* Comparative analysis of transferrin and IgG N-glycosylation in two human populations. *Commun. Biol.* **6**, 312 (2023).
141. Huffman, J. E. *et al.* Comparative performance of four methods for high-throughput glycosylation analysis of immunoglobulin G in genetic and epidemiological research. *Mol. Cell. Proteomics* **13**, 1598–1610 (2014).
142. Lorbek, G., Lewinska, M. & Rozman, D. Cytochrome P450s in the synthesis of cholesterol and bile acids – from mouse models to human diseases. *FEBS J.* **279**, 1516–1533 (2012).
143. Chiang, J. Y. L. Bile Acid Metabolism and Signaling. *Compr. Physiol.* **3**, 1191–1212 (2013).
144. Liebisch, G. The role of LC-MS in lipidomics. *LC-GC Eur.* **30**, 240–242 (2017).

145. Astarita, G., Kendall, A. C., Dennis, E. A. & Nicolaou, A. Targeted lipidomic strategies for oxygenated metabolites of polyunsaturated fatty acids. *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* **1851**, 456–468 (2015).
146. Sandra, K. & Sandra, P. Lipidomics from an analytical perspective. *Current Opinion in Chemical Biology* **17**, 847–853 (2013).
147. Perino, A., Demagny, H., Velazquez-Villegas, L. & Schoonjans, K. Molecular physiology of bile acid signaling in health, disease, and aging. *Physiol. Rev.* **101**, 683–731 (2021).
148. Berger, E. & Haller, D. Structure-function analysis of the tertiary bile acid TUDCA for the resolution of endoplasmic reticulum stress in intestinal epithelial cells. *Biochem. Biophys. Res. Commun.* **409**, 610–615 (2011).
149. Thomas, C., Pellicciari, R., Pruzanski, M., Auwerx, J. & Schoonjans, K. Targeting bile-acid signalling for metabolic diseases. *Nature Reviews Drug Discovery* **7**, 678–693 (2008).
150. Kawamata, Y. *et al.* A G protein-coupled receptor responsive to bile acids. *J. Biol. Chem.* **278**, 9435–9440 (2003).
151. Ding, L., Yang, L., Wang, Z. & Huang, W. Bile acid nuclear receptor FXR and digestive system diseases. *Acta Pharmaceutica Sinica B* **5**, 135–144 (2015).
152. Fiorucci, S. *et al.* Bile Acid Signaling in Inflammatory Bowel Diseases. *Dig Dis Sci* **66**, 674–693 (2021).
153. Prawitt, J., Caron, S. & Staels, B. Bile acid metabolism and the pathogenesis of type 2 diabetes. *Current Diabetes Reports* **11**, 160–166 (2011).
154. Li, R., Andreu-Sánchez, S., Kuipers, F. & Fu, J. Gut microbiome and bile acids in obesity-related diseases. *Best Practice and Research: Clinical Endocrinology and Metabolism* **35**, 101493 (2021).
155. Gottlieb, A. & Canbay, A. Why bile acids are so important in non-alcoholic fatty liver disease (NAFLD) progression. *Cells* **8**, (2019).
156. Režen, T. *et al.* The role of bile acids in carcinogenesis. *Cell. Mol. Life Sci.* **79**, 243 (2022).
157. Bomba, L. *et al.* Whole-exome sequencing identifies rare genetic variants associated with human plasma metabolites. *Am. J. Hum. Genet.* **109**, 1038–1054 (2022).
158. Demirkan, A. *et al.* Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet* **11**, e1004835 (2015).
159. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7**, 11122 (2016).
160. Lotta, L. A. *et al.* A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat. Genet.* **53**, 54–64 (2021).
161. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54–60 (2011).
162. Shin, S.-Y. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
163. Rhee, E. P. *et al.* A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.* **18**, 130–143 (2013).
164. Shen, X. *et al.* Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation. *Nat Commun* **8**, 447 (2017).
165. Deribe, Y. L., Pawson, T. & Dikic, I. Post-translational modifications in signal integration. *Nat. Struct. Mol. Biol.* **17**, 666–672 (2010).
166. Choudhary, C. *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science (80-.).* **325**, 834–840 (2009).
167. Santos, A. L. & Lindner, A. B. Protein Posttranslational Modifications: Roles in

Aging and Age-Related Disease. *Oxidative Medicine and Cellular Longevity* **2017**, (2017).
168. Berdasco, M. & Esteller, M. Aberrant Epigenetic Landscape in Cancer: How Cellular Identity Goes Awry. *Developmental Cell* **19**, 698–711 (2010).
169. Chatterjee, B. & Thakur, S. S. Investigation of post-translational modifications in type 2 diabetes. *Clin. Proteomics* **15**, 32 (2018).
170. Shafi, S. *et al.* Deciphering the Role of Aberrant Protein Post-Translational Modification in the Pathology of Neurodegeneration. *CNS Neurol. Disord. - Drug Targets* **20**, 54–67 (2020).
171. Mimura, Y. *et al.* Glycosylation engineering of therapeutic IgG antibodies: challenges for the safety, functionality and efficacy. *Protein and Cell* **9**, 47–62 (2018).
172. McClain, D. A. *et al.* Adipose tissue transferrin and insulin resistance. *J. Clin. Endocrinol. Metab.* **103**, 4197–4208 (2018).
173. Karlsson, I., Ndreu, L., Quaranta, A., Thorsen, G. & Thorsén, G. Glycosylation patterns of selected proteins in individual serum and cerebrospinal fluid samples. *J Pharm Biomed Anal* **145**, 431–439 (2017).
174. Spik, G. *et al.* Studies on glycoconjugates. LXIV. Complete structure of two carbohydrate units of human serotransferrin. *FEBS Lett.* **50**, 296–299 (1975).
175. Zaytseva, O. O. *et al.* Heritability of Human Plasma N-Glycome. *J. Proteome Res.* **19**, 85–91 (2019).
176. Frkatovic, A., Zaytseva, O. O. & Klaric, L. Genetic Regulation of Immunoglobulin G Glycosylation. in 259–287 (Springer, Cham, 2021). doi:10.1007/978-3-030-76912-3_8
177. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1-3 (2012).
178. Benyamin, B. *et al.* Novel loci affecting iron homeostasis and their effects in individuals at risk for hemochromatosis. *Nat. Commun.* **5**, 4926 (2014).
179. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
180. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
181. Mollicone, R. *et al.* Molecular basis for plasma α(1,3)-fucosyltransferase gene deficiency (FUT6). *J. Biol. Chem.* **269**, 12662–12671 (1994).
182. Turatsinze, J.-V. V., Thomas-Chollier, M., Defrance, M. & van Helden, J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.* **3**, 1578–1588 (2008).
183. Staley, J. R. *et al.* PhenoScanner: A database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
184. Martensson, O., Harlin, A., Brandt, R., Seppa, K. & Sillanaukee, P. Transferrin Isoform Distribution: Gender and Alcohol Consumption. *Alcohol. Clin. Exp. Res.* **21**, 1710–1715 (1997).
185. Ogun, A. S. & Adeyinka, A. *Biochemistry, Transferrin*. *StatPearls* (2021).
186. Allen, H. C. & Sharma, P. Histology, Plasma Cells. in *StatPearls* (2021).
187. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1–10 (2018).
188. Kutalik, Z. *et al.* Genome-wide association study identifies two loci strongly affecting transferrin glycosylation. *Hum. Mol. Genet.* **20**, 3710–7 (2011).
189. Zemunik, T. *et al.* Genome-wide association study of biochemical traits in Korcula Island, Croatia. *Croat. Med. J.* **50**, 23–33 (2009).

190. Kerr, S. M. *et al.* An actionable KCNH2 Long QT Syndrome variant detected by sequence and haplotype analysis in a population research cohort. *Sci. Rep.* **9**, (2019).
191. Trbojević-Akmačić, I. *et al.* Chromatographic monoliths for high-throughput immunoaffinity isolation of transferrin from human plasma. *Croat. Chem. Acta* **89**, 203–211 (2016).
192. Trbojević Akmačić, I. *et al.* High-throughput glycomics: Optimization of sample preparation. *Biochem.* **80**, 934–942 (2015).
193. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
194. Karssen, L. C., van Duijn, C. M. & Aulchenko, Y. S. The GenABEL Project for statistical genomics. *F1000Research* **5**, 914 (2016).
195. Haller, T. *et al.* RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Brief. Bioinform.* **16**, 39–44 (2015).
196. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
197. Pruim, R. J. *et al.* LocusZoom: Regional visualization of genome-wide association scan results. in *Bioinformatics* **27**, 2336–2337 (2011).
198. Zeileis, A. & Hothorn, T. Diagnostic Checking in Regression Relationships. *R News* **2**, 7–10 (2002).
199. Ziyatdinov, A. *et al.* lme4qtl: Linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics* **19**, (2018).
200. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
201. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
202. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
203. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
204. Vlieghe, D. *et al.* A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34**, D95–D97 (2006).
205. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
206. Momozawa, Y. *et al.* IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, 2427 (2018).
207. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
208. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, e34408 (2018).
209. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv* 2020.08.10.244293 (2020). doi:10.1101/2020.08.10.244293
210. Goldstein, D. B. Common Genetic Variation and Human Traits. *https://doi.org/10.1056/NEJMp0806284* **360**, 1696–1698 (2009).
211. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
212. Cirulli, E. T. *et al.* Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science (80-.).* **347**, 1436–1441 (2015).

213. Petrovski, S. *et al.* An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **196**, 82–93 (2017).
214. Allen, A. S. *et al.* Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. *Lancet Neurol.* **16**, 135–143 (2017).
215. Landini, A. *et al.* Genetic regulation of post-translational modification of two distinct proteins. *Nat. Commun. 2022 131* **13**, 1–13 (2022).
216. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-.).* **337**, 1190–1195 (2012).
217. Van Hout, C. V *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
218. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* **11**, 773–785 (2010).
219. Zuk, O. *et al.* Searching for missing heritability : Designing rare variant association studies. (2014). doi:10.1073/pnas.1322563111
220. Vanhooren, V. *et al.* Serum N-glycan profile shift during human ageing. *Exp Gerontol* **45**, 738–743 (2010).
221. Vilaj, M., Gudelj, I., Trbojević-Akmačić, I., Lauc, G. & Pezer, M. IgG Glycans as a Biomarker of Biological Age. in *Biomarkers of Human Aging* 81–99 doi:10.1007/978-3-030-24970-0_7
222. Ząbczyńska, M., Link-Lenczowski, P. & Pocheć, E. Glycosylation in Autoimmune Diseases. in *The Role of Glycosylation in Health and Disease* (eds. Lauc, G. & Trbojević-Akmačić, I.) 205–218 (Springer International Publishing, 2021). doi:10.1007/978-3-030-70115-4_10
223. Gudelj, I. & Lauc, G. Protein N-Glycosylation in Cardiovascular Diseases and Related Risk Factors. *Curr. Cardiovasc. Risk Rep.* **12**, (2018).
224. Rebelo, A. L., Chevalier, M. T., Russo, L. & Pandit, A. Role and therapeutic implications of protein glycosylation in neuroinflammation. *Trends Mol Med* **28**, 270–289 (2022).
225. Costa, A. F., Campos, D., Reis, C. A. & Gomes, C. Targeting Glycosylation: A New Road for Cancer Drug Discovery. *Trends Cancer* **6**, 757–766 (2020).
226. Bondt, A. *et al.* Immunoglobulin G (IgG) Fab glycosylation analysis using a new mass spectrometric high-throughput profiling method reveals pregnancy-associated changes. *Mol Cell Proteomics* **13**, 3029–3039 (2014).
227. Wuhrer, M. *et al.* Glycosylation profiling of immunoglobulin G (IgG) subclasses from human serum. *Proteomics* **7**, 4070–4081 (2007).
228. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
229. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168 (2022).
230. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
231. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
232. Fitzgerald, K. A. *et al.* Mal (MyD88-adapter-like) is required for Toll-like receptor-4 signal transduction. *Nature* **413**, 78–83 (2001).
233. Horng, T., Barton, G. M. & Medzhitov, R. TIRAP: an adapter molecule in the Toll signaling pathway. *Nat Immunol* **2**, 835–841 (2001).

234. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
235. Jones, E. Y., Fugger, L., Strominger, J. L. & Siebold, C. MHC class II proteins and disease: a structural perspective. *Nat Rev Immunol* **6**, 271–282 (2006).
236. Juszczak, A. *et al.* Plasma fucosylated glycans and C-reactive protein as biomarkers of HNF1A-MODY in young adult–onset nonautoimmune diabetes. *Diabetes Care* **42**, 17–26 (2019).
237. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
238. Verhelst, X. *et al.* Protein Glycosylation as a Diagnostic and Prognostic Marker of Chronic Inflammatory Gastrointestinal and Liver Diseases. *Gastroenterology* **158**, 95–110 (2020).
239. Wittenbecher, C. *et al.* Plasma N-Glycans as Emerging Biomarkers of Cardiometabolic Risk: A Prospective Investigation in the EPIC-Potsdam Cohort Study. *Diabetes Care* **43**, 661–668 (2020).
240. Meyer, N. M. T. *et al.* Low IGF1 and high IGFBP1 predict diabetes onset in prediabetic patients. *Eur J Endocrinol* **187**, 555–565 (2022).
241. Segev, Y. *et al.* Systemic and renal growth hormone-IGF1 axis involvement in a mouse model of type 2 diabetes. *Diabetologia* **50**, 1327–1334 (2007).
242. Gardner, E. J. *et al.* Damaging missense variants in IGF1R implicate a role for IGF-1 resistance in the etiology of type 2 diabetes. *Cell Genomics* doi:10.1016/j.xgen.2022.100208
243. Atalar, M. N. *et al.* Assessment of serum galectin-3, methylated arginine and Hs-CRP levels in type 2 diabetes and prediabetes. *Life Sci* **231**, 116577 (2019).
244. Lin, D. *et al.* Galectin-3/adiponectin as a new biological indicator for assessing the risk of type 2 diabetes: a cross-sectional study in a community population. *Aging (Albany NY)* **13**, 15433–15443 (2021).
245. Ohkura, T. *et al.* Low serum galectin-3 concentrations are associated with insulin resistance in patients with type 2 diabetes mellitus. *Diabetol Metab Syndr* **6**, 106 (2014).
246. Vora, A., de Lemos, J. A., Ayers, C., Grodin, J. L. & Lingvay, I. Association of Galectin-3 With Diabetes Mellitus in the Dallas Heart Study. *J Clin Endocrinol Metab* **104**, 4449–4458 (2019).
247. Weigert, J. *et al.* Serum galectin-3 is elevated in obesity and negatively correlates with glycosylated hemoglobin in type 2 diabetes. *J Clin Endocrinol Metab* **95**, 1404–1411 (2010).
248. Carlsson, M. C., Bengtson, P., Cucak, H. & Leffler, H. Galectin-3 guides intracellular trafficking of some human serotransferrin glycoforms. *J Biol Chem* **288**, 28398–28408 (2013).
249. Cederfur, C. *et al.* Different affinity of galectins for human serum glycoproteins: galectin-3 binds many protease inhibitors and acute phase proteins. *Glycobiology* **18**, 384–394 (2008).
250. Reid, J. G. *et al.* Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics* **15**, 30 (2014).
251. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
252. “Picard Toolkit”. 2019. Broad Institute, GitHub Repository: https://broadinstitute.github.io/picard, Broad Institute.
253. ‘weCall’. 2018. GitHub Repository: https://github.com/Genomicsplc/wecall. Genomics PLC.
254. Lin, M. F. *et al.* GLnexus: joint variant calling for large cohort sequencing.

bioRxiv (2018). doi:10.1101/343970
255. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
256. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553–1561 (2009).
257. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575–576 (2010).
258. Pučić, M. *et al.* High throughput isolation and glycosylation analysis of IgG-variability and heritability of the IgG glycome in three isolated human populations. *Mol. Cell. Proteomics* **10**, M111.010090-M111.010090 (2011).
259. Trbojević Akmačić, I., Ugrina, I. & Lauc, G. *Methods in Enzymology, Volume 586: Chapter Three - Comparative Analysis and Validation of Different Steps in Glycomics Studies*. (2017). doi:https://doi.org/10.1016/bs.mie.2016.09.027
260. Kassambara, A. and Mundt, F. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7. *https://CRAN.R-project.org/package=factoextra* (2020).
261. Phelps, T., Snyder, E., Rodriguez, E., Child, H. & Harvey, P. The influence of biological sex and sex hormones on bile acid synthesis and cholesterol homeostasis. *Biol Sex Differ* **10**, 52 (2019).
262. Abu-Hayyeh, S. *et al.* Intrahepatic cholestasis of pregnancy levels of sulfated progesterone metabolites inhibit farnesoid X receptor resulting in a cholestatic phenotype. *Hepatology* **57**, 716–726 (2013).
263. Li-Hawkins, J. *et al.* Cholic acid mediates negative feedback regulation of bile acid synthesis in mice. *J Clin Invest* **110**, 1191–1200 (2002).
264. de Aguiar Vallim, T. Q., Tarling, E. J. & Edwards, P. A. Pleiotropic roles of bile acids in metabolism. *Cell Metab* **17**, 657–669 (2013).
265. Perino, A., Demagny, H., Velazquez-Villegas, L. & Schoonjans, K. Molecular Physiology of Bile Acid Signaling in Health, Disease, and Aging. *Physiol Rev* **101**, 683–731 (2021).
266. Rezen, T. *et al.* The role of bile acids in carcinogenesis. *Cell Mol Life Sci* **79**, 243 (2022).
267. Li, H. *et al.* Integrative systems analysis identifies genetic and dietary modulators of bile acid homeostasis. *Cell Metab.* **34**, 1594-1610.e4 (2022).
268. Chen, L. *et al.* Genetic and Microbial Associations to Plasma and Fecal Bile Acids in Obesity Relate to Plasma Lipids and Liver Fat Content. *Cell Rep.* **33**, 108212 (2020).
269. Scherer, M., Gnewuch, C., Schmitz, G. & Liebisch, G. Rapid quantification of bile acids and their conjugates in serum by liquid chromatography-tandem mass spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **877**, 3920–3925 (2009).
270. Danic, M. *et al.* Pharmacological Applications of Bile Acids and Their Derivatives in the Treatment of Metabolic Syndrome. *Front Pharmacol* **9**, 1382 (2018).
271. Stojančević, M., Pavlović, N., Goločorbin-Kon, S. & Mikov, M. Application of bile acids in drug formulation and delivery. *Front. Life Sci.* **7**, 112–122
272. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).
273. Hagenbuch, B. & Meier, P. J. Organic anion transporting polypeptides of the OATP/ SLC21 family: phylogenetic classification as OATP/ SLCO superfamily, new nomenclature and molecular/functional properties. *Pflugers Arch* **447**, 653–

665 (2004).
274. Ho, R. H. & Kim, R. B. Transporters and drug therapy: implications for drug disposition and disease. *Clin Pharmacol Ther* **78**, 260–277 (2005).
275. International Transporter, C. *et al.* Membrane transporters in drug development. *Nat Rev Drug Discov* **9**, 215–236 (2010).
276. Niemi, M., Pasanen, M. K. & Neuvonen, P. J. Organic anion transporting polypeptide 1B1: a genetically polymorphic transporter of major importance for hepatic drug uptake. *Pharmacol Rev* **63**, 157–181 (2011).
277. Nies, A. T., Schwab, M. & Keppler, D. Interplay of conjugating enzymes with OATP uptake transporters and ABCC/MRP efflux pumps in the elimination of drugs. *Expert Opin Drug Metab Toxicol* **4**, 545–568 (2008).
278. Revez, J. A. *et al.* Genome-wide association study identifies 143 loci associated with 25 hydroxyvitamin D concentration. *Nat Commun* **11**, 1647 (2020).
279. Johnson, A. D. *et al.* Genome-wide association meta-analysis for total serum bilirubin levels. *Hum Mol Genet* **18**, 2700–2710 (2009).
280. Ruth, K. S. *et al.* Using human genetics to understand the disease impacts of testosterone in men and women. *Nat Med* **26**, 252–258 (2020).
281. Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res* **49**, D1302–D1310 (2021).
282. Vitart, V. *et al.* 3000 years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia. *Eur. J. Hum. Genet.* **14**, 478–487 (2006).
283. Koh, M. Y., Lemos Jr., R., Liu, X. & Powis, G. The hypoxia-associated factor switches cells from HIF-1alpha- to HIF-2alpha-dependent signaling promoting stem cell characteristics, aggressive tumor growth and invasion. *Cancer Res* **71**, 4015–4027 (2011).
284. Semenza, G. L. Hypoxia, clonal selection, and the role of HIF-1 in tumor progression. *Crit Rev Biochem Mol Biol* **35**, 71–103 (2000).
285. Braun, T., Voland, P., Kunz, L., Prinz, C. & Gratzl, M. Enterochromaffin cells of the human gut: sensors for spices and odorants. *Gastroenterology* **132**, 1890–1901 (2007).
286. Erspamer, V. Pharmacology of indole-alkylamines. *Pharmacol Rev* **6**, 425–487 (1954).
287. Watanabe, H. *et al.* Peripheral serotonin enhances lipid metabolism by accelerating bile acid turnover. *Endocrinology* **151**, 4776–4786 (2010).
288. Oakley, F. *et al.* Hepatocytes express nerve growth factor during liver injury: evidence for paracrine regulation of hepatic stellate cell apoptosis. *Am J Pathol* **163**, 1849–1858 (2003).
289. Ohkubo, T. *et al.* Early induction of nerve growth factor-induced genes after liver resection-reperfusion injury. *J Hepatol* **36**, 210–217 (2002).
290. Valdovinos-Flores, C. & Gonsebatt, M. E. Nerve growth factor exhibits an antioxidant and an autocrine activity in mouse liver that is modulated by buthionine sulfoximine, arsenic, and acetaminophen. *Free Radic Res* **47**, 404–412 (2013).
291. Gigliozzi, A. *et al.* Nerve growth factor modulates the proliferative capacity of the intrahepatic biliary epithelium in experimental cholestasis. *Gastroenterology* **127**, 1198–1209 (2004).
292. Rasi, G. *et al.* Nerve growth factor involvement in liver cirrhosis and hepatocellular carcinoma. *World J Gastroenterol* **13**, 4986–4995 (2007).
293. Tokusashi, Y. *et al.* Expression of NGF in hepatocellular carcinoma cells with its receptors in non-tumor cell components. *Int J Cancer* **114**, 39–45 (2005).
294. Frommherz, L. *et al.* Age-Related Changes of Plasma Bile Acid Concentrations

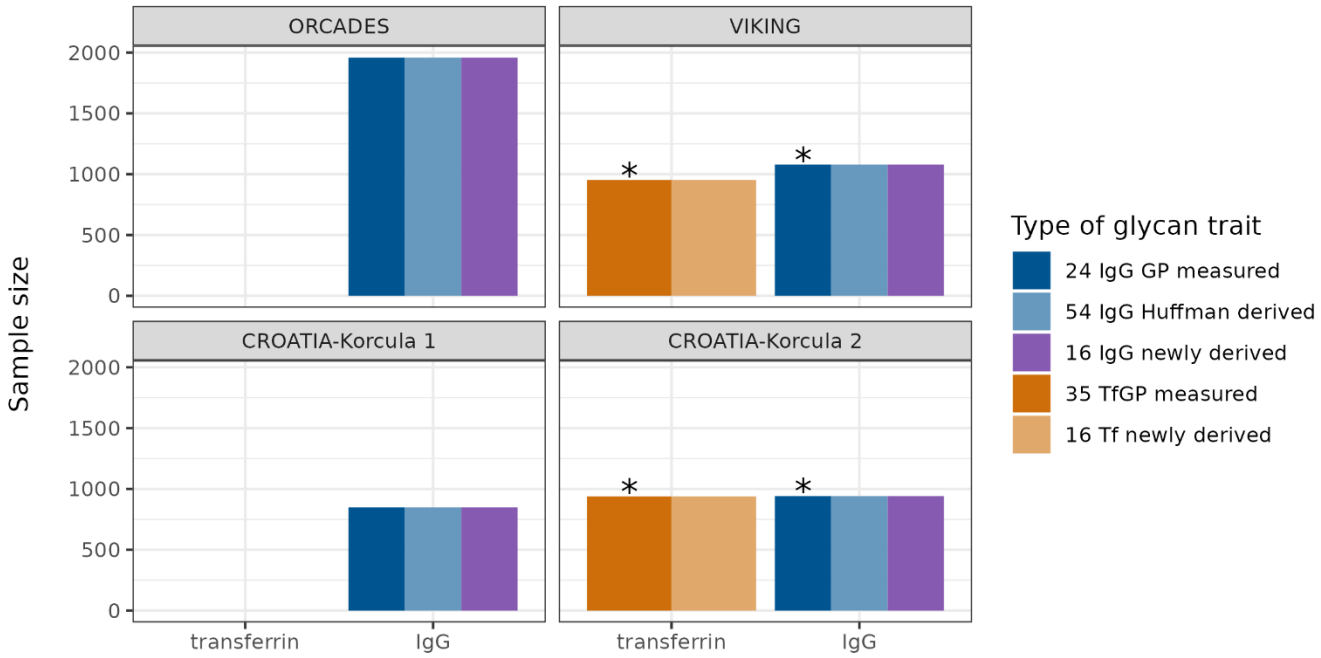
in Healthy Adults--Results from the Cross-Sectional KarMeN Study. *PLoS One* **11**, e0153959 (2016).
295. Dekkers, K. F. *et al.* An online atlas of human plasma metabolite signatures of gut microbiome composition. *Nat. Commun.* **13**, 1–12 (2022).
296. Russell, D. W. The Enzymes, Regulation, and Genetics of Bile Acid Synthesis. *Annu. Rev. Biochem.* **72**, 137–174 (2003).
297. Thomas, C. E. *et al.* Association between Pre-Diagnostic Serum Bile Acids and Hepatocellular Carcinoma: The Singapore Chinese Health Study. *Cancers (Basel)* **13**, (2021).
298. Manzotti, C., Casazza, G., Stimac, T., Nikolova, D. & Gluud, C. Total serum bile acids or serum bile acid profile, or both, for the diagnosis of intrahepatic cholestasis of pregnancy. *Cochrane Database Syst Rev* **7**, CD012546 (2019).
299. Hadfield, J. D. MCMC Methods for Multi-Response Generalized Linear Mixed Models: TheMCMCglmmRPackage. *J. Stat. Softw.* **33**, (2010).
300. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
301. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
302. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235–241 (2016).
303. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894–899 (2011).
304. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat Protoc* **11**, 1–9 (2016).
305. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
306. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (80-.).* **369**, 1318–1330 (2020).
307. Bach, M. A., Roberts, C. T., Smith, E. P. & Leroith, D. Alternative splicing produces messenger rnas encoding insulinlike growth factor-1 prohormones that are differentially glycosylated in vitro. *Mol. Endocrinol.* **4**, 899–904 (1990).
308. Reale, M. A. *et al.* Expression and Alternative Splicing of the Deleted in Colorectal Cancer (DCC) Gene in Normal and Malignant Tissues. *Cancer Res.* **54**, 4493–4501 (1994).
309. Quintana, E. *et al.* Screening for congenital disorders of glycosylation (CDG): Transferrin HPLC versus isoelectric focusing (IEF). *Clin. Biochem.* **42**, 408–415 (2009).
310. Stibler, H. Carbohydrate-deficient transferrin in serum: A new marker of potentially harmful alcohol consumption reviewed. *Clinical Chemistry* **37**, 2029–2037 (1991).
311. Helander, A., Eriksson, G., Stibler, H. & Jeppsson, J. O. Interference of transferrin isoform types with carbohydrate-deficient transferrin quantification in the identification of alcohol abuse. *Clin. Chem.* **47**, 1225–1233 (2001).
312. Lefeber, D. J., Morava, E. & Jaeken, J. How to find and diagnose a CDG due to defective N-glycosylation. *J. Inherit. Metab. Dis.* **34**, 849 (2011).
313. Gade, P. & Kalvakolanu, D. V. Chromatin immunoprecipitation assay as a tool for analyzing transcription factor activity. *Methods Mol. Biol.* **809**, 85–104 (2012).
314. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nature Communications* **9**, 1–13 (2018).
315. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants

and genes at all published human GWAS trait-associated loci. *Nat. Genet. 2021 5311* **53**, 1527–1533 (2021).
316. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res* **49**, D1311–D1320 (2021).
317. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nat. 2019 5707759* **570**, 71–76 (2019).
318. Seaby, E. G., Pengelly, R. J. & Ennis, S. Exome sequencing explained: a practical guide to its clinical application. *Brief. Funct. Genomics* **15**, 374–384 (2016).
319. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
320. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **0 7**, Unit7.20 (2013).
321. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science (80-.).* **335**, 823–828 (2012).
322. Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).
323. Ritchie, G. R. & Flicek, P. Computational approaches to interpreting genomic sequence variation. *Genome Medicine* **6**, 87 (2014).
324. Castellana, S. & Mazza, T. Congruency in the prediction of pathogenic missense mutations: State-of-the-art web-based tools. *Brief. Bioinform.* **14**, 448–459 (2013).
325. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
326. Duzkale, H. *et al.* A systematic approach to assessing the clinical significance of genetic variants. *Clin. Genet.* **84**, 453–463 (2013).
327. Ober, C., Loisel, D. A. & Gilad, Y. Sex-specific genetic architecture of human disease. *Nature Reviews Genetics* **9**, 911–922 (2008).
328. Cai, H., Bai, Z. & Ge, R. L. Hypoxia-inducible factor-2 promotes liver fibrosis in non-alcoholic steatohepatitis liver disease via the NF-κB signalling pathway. *Biochem. Biophys. Res. Commun.* **540**, 67–74 (2021).
329. Foglia, B. *et al.* Hepatocyte-Specific Deletion of HIF2α Prevents NASH-Related Liver Carcinogenesis by Decreasing Cancer Cell Proliferation. *Cell. Mol. Gastroenterol. Hepatol.* **13**, 459–482 (2022).
330. Liu, N. *et al.* Role of bile acids in the diagnosis and progression of liver cirrhosis: A prospective observational study. *Exp. Ther. Med.* **18**, 4058–4066 (2019).
331. Luo, L. *et al.* Assessment of serum bile acid profiles as biomarkers of liver injury and liver disease in humans. *PLoS One* **13**, e0193824 (2018).
332. Lu, H. *et al.* Characterization of organic anion transporting polypeptide 1b2-null mice: essential role in hepatic uptake/toxicity of phalloidin and microcystin-LR. *Toxicol. Sci.* **103**, 35–45 (2008).
333. Niemi, M., Pasanen, M. K. & Neuvonen, P. J. SLCO1B1 polymorphism and sex affect the pharmacokinetics of pravastatin but not fluvastatin. *Clin. Pharmacol. Ther.* **80**, 356–366 (2006).
334. Franko, A. *et al.* cGMP-dependent protein kinase I (cGKI) modulates human hepatic stellate cell activation. *Metabolism.* **88**, 22–30 (2018).
335. Novo, E. *et al.* Liver fibrogenesis: un update on established and emerging basic concepts. *Archives of Biochemistry and Biophysics* **689**, 108445 (2020).
336. Azer, S. A. & Hasanato, R. Use of bile acids as potential markers of liver

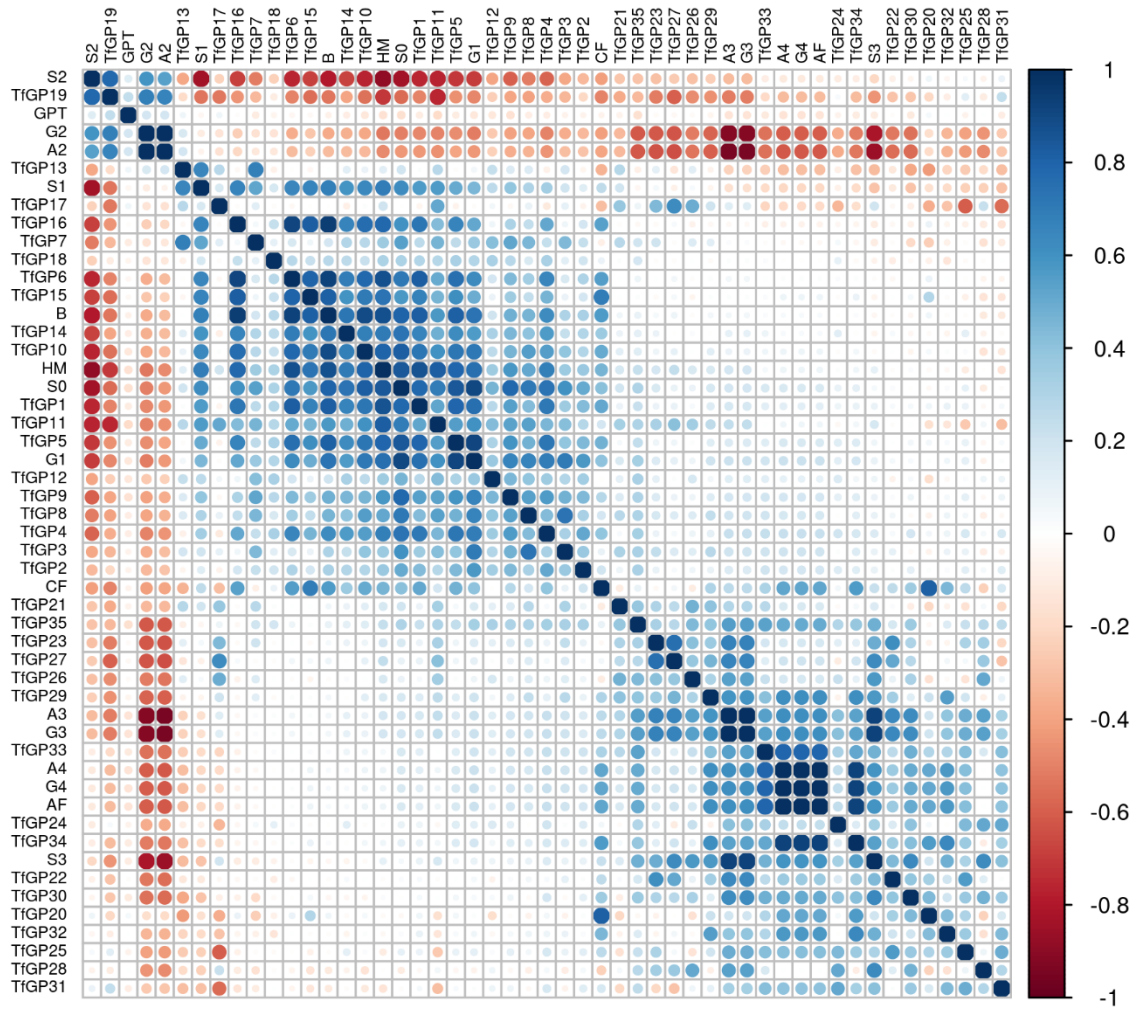
dysfunction in humans: A systematic review. *Medicine (Baltimore).* **100**, e27464 (2021).
337. Knežević, A. *et al.* High throughput plasma N-glycome profiling using multiplexed labelling and UPLC with fluorescence detection. *Analyst* **136**, 4670 (2011).
338. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
339. Frkatović-Hodžić, A. *et al.* Mapping of the gene network that regulates IgG galactosylation. *Under Rev.*
340. Pan, Q. & Hammarström, L. Molecular basis of IgG subclass deficiency. *Immunol. Rev.* **178**, 99–110 (2000).
341. Zaytseva, O. O. *et al.* Investigation of the causal relationships between human IgG N-glycosylation and 12 common diseases associated with changes in the IgG N-glycome. *Hum. Mol. Genet.* **31**, 1545–1559 (2021).
342. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet. 2018 5011* **50**, 1593–1599 (2018).
343. Gebrehiwot, A. G. *et al.* Healthy human serum N-glycan profiling reveals the influence of ethnic variation on the identified cancer-relevant glycan biomarkers. *PLoS One* **13**, (2018).
344. Kamiyama, T. *et al.* Identification of novel serum biomarkers of hepatocellular carcinoma using glycomic analysis. *Hepatology* **57**, 2314–2325 (2013).
345. Hatakeyama, S. *et al.* Serum N-glycan alteration associated with renal cell carcinoma detected by high throughput glycan analysis. *J. Urol.* **191**, 805–813 (2014).
346. Ishibashi, Y. *et al.* Serum tri- and tetra-antennary N-glycan is a potential predictive biomarker for castration-resistant prostate cancer. *Prostate* **74**, 1521–1529 (2014).
347. Štambuk, J. *et al.* Global variability of the human IgG glycome. *Aging (Albany. NY).* **12**, 15222–15259 (2020).
348. Liu, D. *et al.* Genome-Wide Mapping of Plasma IgG N-Glycan Quantitative Trait Loci Identifies a Potentially Causal Association between IgG N-Glycans and Rheumatoid Arthritis. *J. Immunol.* **208**, 2508–2514 (2022).
349. Yarwood, A. *et al.* Loci associated with N-glycosylation of human IgG are not associated with rheumatoid arthritis: A Mendelian randomisation study. *Ann. Rheum. Dis.* **75**, 317–320 (2016).
350. Costarelli, V., Sanders, T. & Reddy, S. Fasting plasma bile acid concentrations in Asian vegetarians, Caucasian vegetarians and Caucasian omnivores. *Nutr. Food Sci.* **36**, 153–158 (2006).
351. Sarin, S. K. *et al.* Liver diseases in the Asia-Pacific region: a Lancet Gastroenterology & Hepatology Commission. *Lancet. Gastroenterol. Hepatol.* **5**, 167 (2020).
352. Carlson, C. S. *et al.* Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLOS Biol.* **11**, e1001661 (2013).
353. Šimunović, J. *et al.* High-throughput immunoaffinity enrichment and N-glycan analysis of human plasma haptoglobin. *Biotechnol. Bioeng.* (2022). doi:10.1002/BIT.28280
354. Keser, T. *et al.* High-Throughput and Site-Specific N-Glycosylation Analysis of Human Alpha-1-Acid Glycoprotein Offers a Great Potential for New Biomarker Discovery. *Mol. Cell. Proteomics* **20**, (2021).
355. Plomp, R. *et al.* Comparative glycomics of immunoglobulin A and G from saliva and plasma reveals biomarker potential. *Front. Immunol.* **9**, 2436 (2018).

356. Timoshchuk, A., Sharapov, S. & Aulchenko, Y. S. Twelve Years of Genome-Wide Association Studies of Human Protein N-Glycosylation. *Engineering* (2023). doi:10.1016/J.ENG.2023.03.013

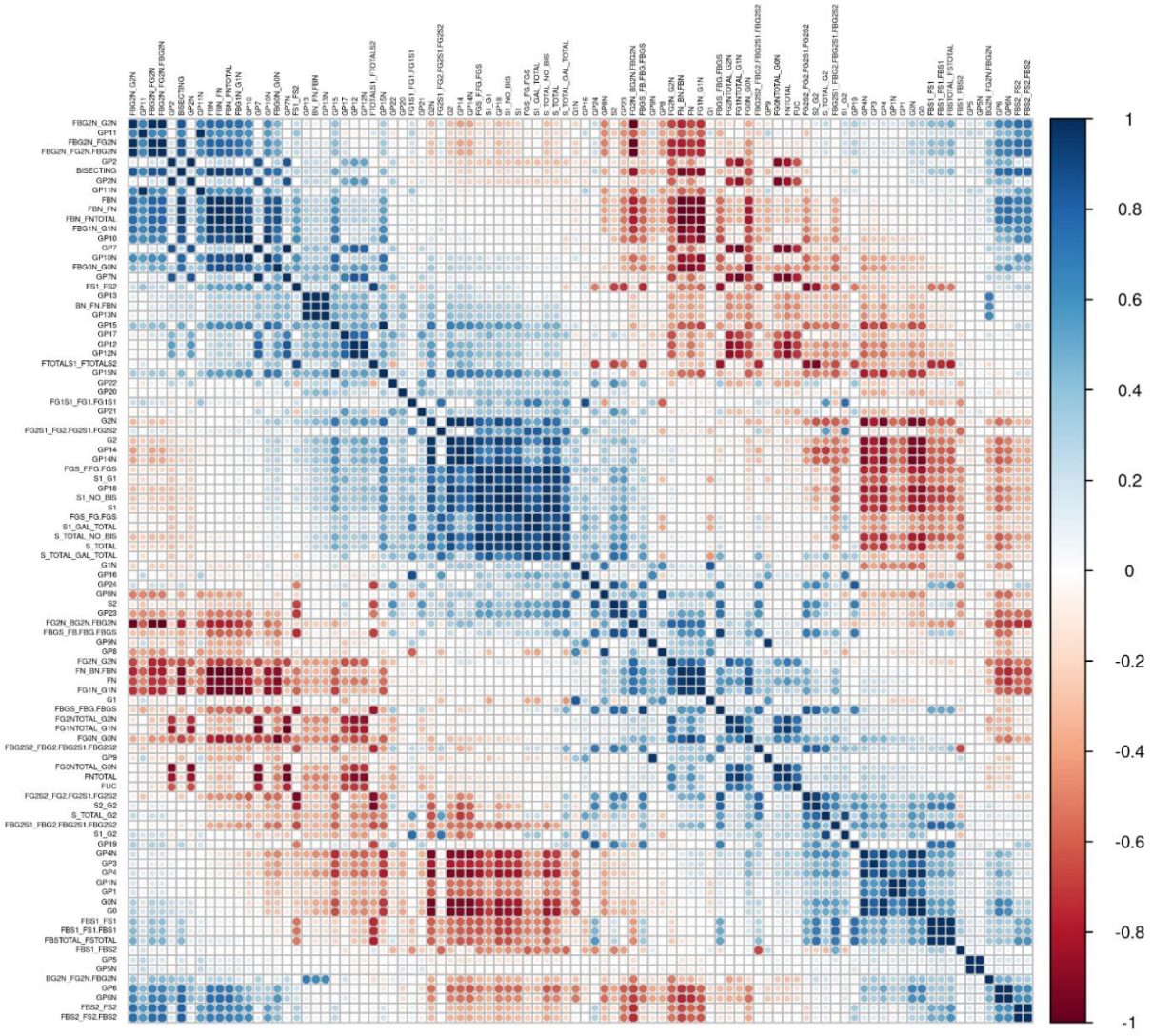
Appendix



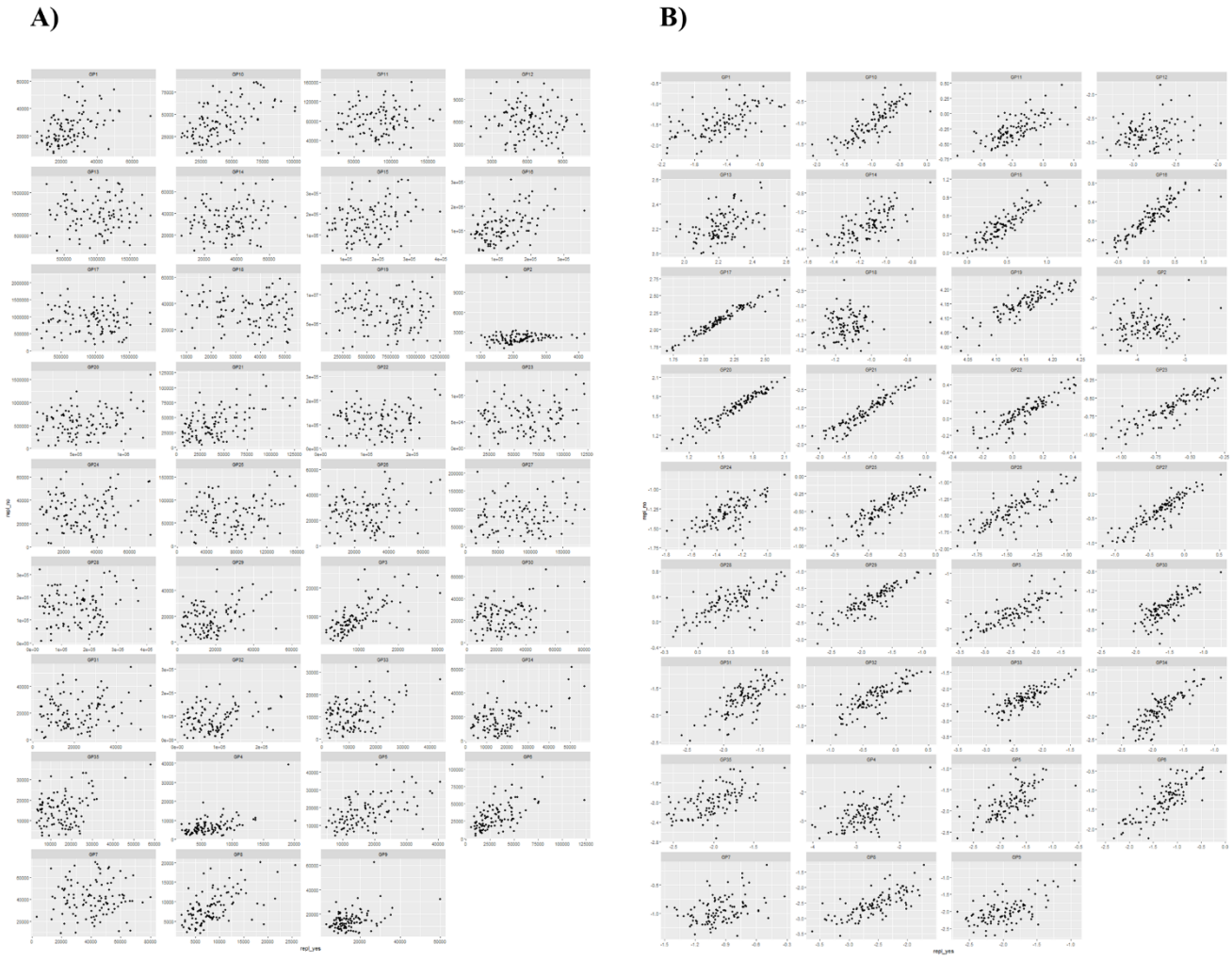
Supplementary Figure 1. Summary of cohorts and sample sizes for all transferrin and IgG glycan traits assayed in Chapters 2 and 3. Transferrin glycan traits overall assayed in this thesis are 35 directly measured traits, whose structural characterization is reported by Trbojević-Akmačić et al.[140] and 16 derived traits, calculated as detailed in the Supplementary Table 15 of Chapter 3, from VIKING (N=952) and batch 2 of CROATIA-Korcula (N=938) cohorts. IgG glycan traits overall assayed in this thesis are 24 directly measured traits, whose structural characterization is detailed in Huffman et al.[141], 54 derived traits as defined by Huffman et al.[141] and 16 newly derived traits, calculated as detailed in the Supplementary Table 16 of Chapter 3, from ORCADES (N=1959), VIKING (N=1079), batch 1 (N=849) and 2 (N=941) of CROATIA-Korcula cohorts. While all of the above described measured and derived glycan traits are included in the analyses of Chapter 3, only the actually measured glycan traits (35 for transferrin and 24 for IgG) of VIKING and batch 2 of CROATIA-Korcula, marked in the bar chart by a "*", are analysed in Chapter 2.



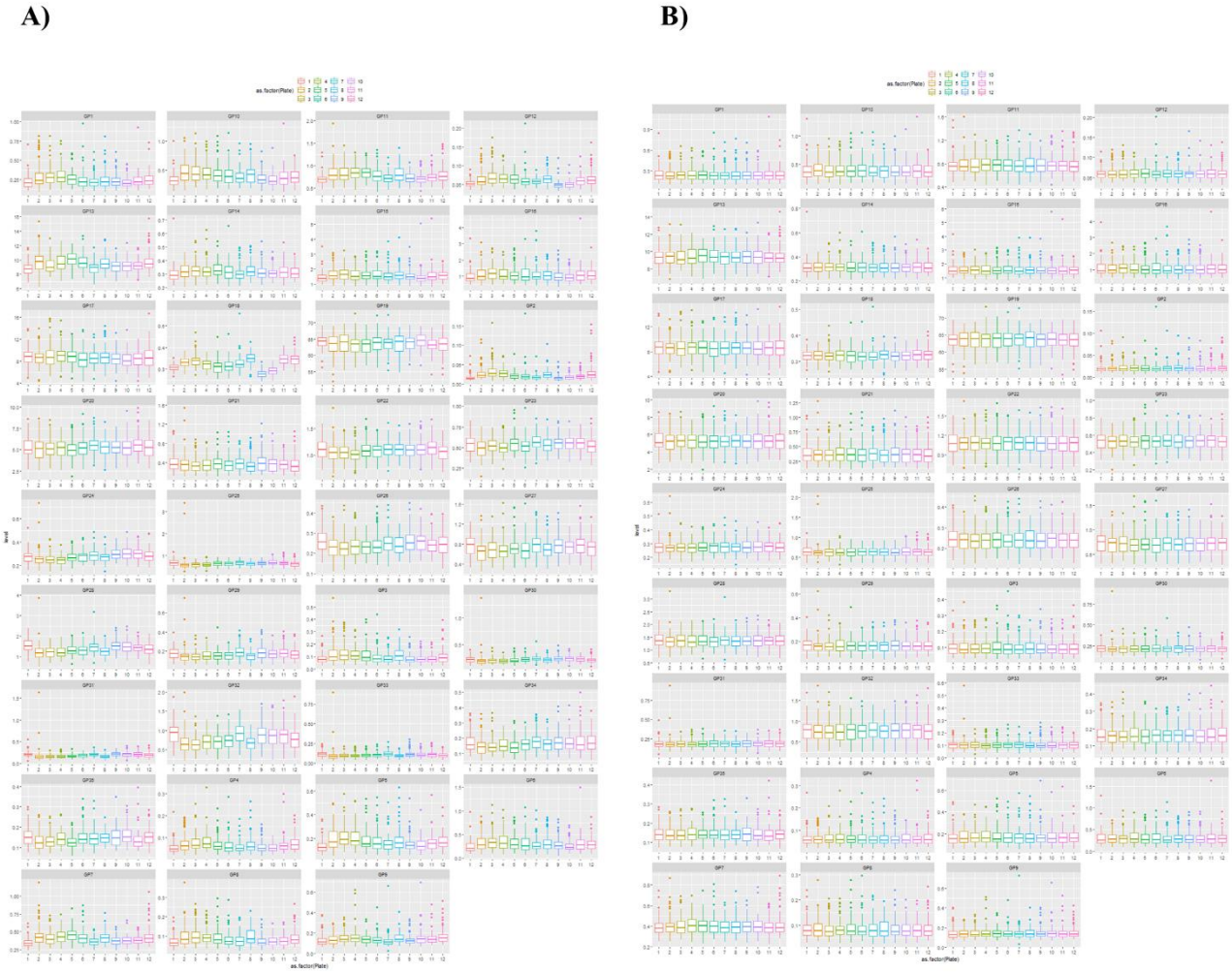
Supplementary Figure 2. Correlation of transferrin glycan measurements in VIKING cohort. Prior to correlation analysis, transferrin glycan measurements were normalised, batch corrected, rank-transformed and adjusted for age, sex and cryptic relatedness. The structural characterization of measured glycan traits, called by the number of their glycan peak (TfGP), are available at Supplementary Table 2 of Trbojević-Akmačić et al.[140] The description and computing formulas of newly derived glycan traits are available at Supplementary Table 15 of the Chapter 3 of this thesis. The blue colour indicates a positive Pearson correlation, while the red colour indicates a negative Pearson correlation. The areas of circles show the absolute value of corresponding correlation coefficients.



Supplementary Figure 3. Correlation of IgG glycan measurements in VIKING cohort. Prior to correlation analysis, IgG glycan measurements were normalised, batch corrected, rank-transformed and adjusted for age, sex and cryptic relatedness. The structural characterization of measured IgG glycan traits, called by the number of their glycan peak (GP), and the description and computing formulas of derived glycan traits are available at Supplementary Table 1 of Huffman et al.[141] The description and computing formulas of newly derived IgG glycan traits are available at Supplementary Table 16 of the Chapter 3 of this thesis. The blue colour indicates a positive Pearson correlation, while the red colour indicates a negative Pearson correlation. The areas of circles show the absolute value of corresponding correlation coefficients.



Supplementary Figure 4: Effectiveness of normalisation and batch correction on correlation of duplicated samples in transferrin glycan traits. Plots show the correlation of duplicated samples in raw data (A) and normalised and batch corrected data (B). All the 35 transferrin glycan traits examined are shown. The correlation between duplicated samples is overall increased across glycan traits after batch correction and normalisation. However, while some traits are characterised by a dramatic increase in correlation (e.g. TfGP17 and TfGP20), some still contain considerable measurement error (e.g. TfGP2). In these latter cases, the pre-processing procedure has been less effective in removing experimental noise.



Supplementary Figure 5: Effectiveness of batch correction on transferrin glycan traits. Plots show the distribution of samples, divided by plates, before (A) and after batch correction (B). All the 35 glycan traits examined are shown. Batch effect has been remarkably reduced thanks to batch correction, even for strongly affected glycan traits (e.g. TfGP18 and TfGP2).

Chapter 2 - Supplementary Methods

Transferrin isolation

Flowthrough during IgG purification was collected for immediate subsequent Tf isolation using previously developed preconditioned CIMac-@Tf 96-well monolithic plate[1]. Unbound proteins during Tf isolation were washed away with 1x PBS (0.25 mol L^{-1} NaCl), pH 7.4. Bound Tf was eluted with 0.7 mL of 0.1 mol L^{-1} formic acid pH 3.0 (pH adjusted with 25 % ammonia solution, Merck) and immediately neutralized with 1 mol L^{-1} ammonium hydrogencarbonate (Sigma-Aldrich) to pH 7.0. Monolithic plate was regenerated and stored at 4 °C until the next isolation. Each elution fraction (300 µL) was dried in a vacuum centrifuge (Thermo Scientific) and stored at -20 °C until subsequent N-glycan release. To ensure quality of transferrin purification from isolation to isolation we randomly selected transferrin eluates (7.5x concentrated) per each plate and analysed them by SDS-PAGE to check for potential contaminants. As can be seen in the Supplementary Figure 11, the purification was successful and no other contaminants were detected. We next analysed transferrin eluate for transferrin purity by performing trypsin digestion and LC-MS analysis of obtained (glyco)peptides. Pooled isolated transferrin sample was reduced with dithiothreitol and alkylated with iodoacetamide prior to trypsin digestion. Tryptic glycopeptides and peptides were separated and analysed by nano liquid chromatography coupled to electrospray ionization quadrupole time of flight mass spectrometry (nanoLC-ESI-qTOF-MS). A search for specific tryptic peptides with a maximum of 2 miscleavages was done in MaxQuant software[2] against *Homo sapiens* proteins sequences (UniProt fasta file) with the methionine oxidation and asparagine carrying *N*-acetylhexosamine as variable modifications and carbamidomethyl on cysteine as the fixed modification. Analysis was performed in triplicates and the average intensity extracted for serotransferrin (UniProt P02787) was 99.36%, which confirms high purity of the transferrin sample.

N-glycan release and fluorescent labelling

Dried Tf eluates were denatured with 30 µl of 13.3 g L^{-1} sodium dodecyl sulfate (SDS, Invitrogen) and by incubation at 65 °C for 10 min. After cooling down to room temperature for 30 min, 10 µl of 4 % (v/v) Igepal CA-630 (Sigma-Aldrich) was added and the mixture was shaken for 15 min on a plate shaker. N-glycans were released after the addition of 10 µL of 5x PBS and 1.2 U of PNGase F (Promega) by incubation at 37 °C for 18 hours. Released N-glycans were labeled with 2-aminobenzamide (2-AB, Sigma-Aldrich). The labeling mixture was freshly prepared by dissolving 2-AB and 2-methylpyridine borane complex (2-PB, Sigma-Aldrich) (final concentrations of 19.2 mg mL^{-1} and 44.8 mg mL^{-1}, respectively) in the mixture of dimethyl sulfoxide (Sigma-Aldrich) and glacial acetic acid (Merck) (7:3). Labeling mixture (25 µL) was added to each sample and the plate was sealed using an adhesive seal. After 10 minutes of shaking, samples were incubated for 2 hours at 65 °C. Excess of reagents from previous steps was removed from the samples using hydrophilic interaction liquid chromatography solid phase extraction (HILIC-SPE). After free N-glycan labeling samples were cooled down to room temperature for 30 min and 700 µL of acetonitrile (previously cooled down to 4 °C) was added to each sample. The cleanup procedure was performed on a hydrophilic 0.2 µm AcroPrep GHP filter plate (Pall) using a vacuum manifold (Pall) at around 25 mm Hg. All wells of a GHP filter plate were prewashed with 200 µL of 70 % (v/v) ethanol in water, 200 µL of ultrapure water, and 200 µL of 96 % (v/v) acetonitrile in water (previously cooled down to 4 °C). Diluted samples were loaded to the GHP filter plate wells, and after short incubation subsequently washed with 5x 200 µL of 96 % (v/v) acetonitrile in water. The last washing step was followed by centrifugation at 164 *g* for 5 minutes. Glycans were eluted from the plate with 2x 90 µL of ultrapure water after 15 min shaking at room temperature and centrifugation at 164 *g* for 5 minutes in each step. Combined eluates of 2-AB labeled Tf N-glycans were stored at -20 °C until ultra-high-performance liquid chromatography (UHPLC) analysis.

Glycan analysis by ultra-high-performance liquid chromatography

Fluorescently labelled and purified Tf N-glycans were analyzed by UHPLC based on hydrophilic interactions (HILIC-UHPLC) and detected using excitation and emission wavelengths of 250 and 428 nm, respectively. Acquity UHPLC instrument (Waters) was under the control of Empower 3 software, build 3471 (Waters). Mobile phases were 100 mmol L^{-1} ammonium formate, pH 4.4 (solvent A) and acetonitrile (solvent B) and samples were maintained at 10 °C before injection. Tf 2-AB labeled N-glycans prepared in 75 % acetonitrile were separated on a Waters BEH Glycan column, 150 × 2.1 mm i.d., 1.7 µm BEH particles at 25 °C in a linear gradient of 30-47 % solvent A at a flow rate of 0.56 mL min^{-1} during a 23 minute analytical run. The HILIC-UHPLC system was calibrated using a dextran ladder (external standard of hydrolysed and 2-AB labelled glucose oligomers) according to which the retention times for the individual chromatographic peaks (representing the 2-AB labeled glycan) were converted to glucose units (GU). Data processing was performed using an automatic processing method with a traditional integration algorithm. Each Tf N-glycans chromatogram integrated into 35 peaks was manually corrected to maintain the same intervals of integration for all the samples. The amount of glycans in each chromatographic peak was expressed as a percentage of the total integrated area (% Area).

Replication of transferrin N-glycans loci

To assess robustness of our findings we used the VIKING cohort as replication cohort and CROATIA-Korcula as discovery cohort. Each significant sentinel SNP-top glycan trait pair from the discovery cohort was tested for associations in the replication cohort, with replication significance threshold set to the p-value ≤ 0.00625 (0.05/8, number of discovery cohort genome-wide significant loci). Where the SNP of interest was not available in the replication cohort, a proxy SNP in high linkage disequilibrium ($r2 \geq 0.8$) was used instead. In addition to

statistical significance, we also assessed if the direction of estimated effect was concordant between discovery and replication study.

Chapter 2 - Supplementary Results

Transferrin N-glycans shared genetic associations with complex traits and diseases

For the shared associations between transferrin glycosylation from the *ST3GAL4* locus and LDL, total cholesterol levels and platelet-related traits; *HNF1A* and coronary artery disease, levels of C-reactive protein and of gamma-glutamyl transferase; *FUT6* and age-related macular degeneration (Supplementary Data 11a); the SMR p-value was not significant (Supplementary Data 11b), so the inference on colocalisation could not be performed.

Colocalisation analysis of transferrin and IgG glycan traits with multiple independent association signals at genomic region

To investigate whether the same variant within the *FUT8* and *FUT6* loci is regulating glycosylation of both proteins, and, at the same time, account for the presence of multiple conditionally distinct association signals within the same locus, we applied the PwCoCo pipeline[3], integrating Approximate Bayes Factor (ABF) colocalisation[4] and conditional analyses (for details see Supplementary Figure 1). Briefly, to address the problem of multiple associations within a locus, this approach tests for colocalisation using not only the trait's unconditioned GWAS association statistics, but also their conditioned ones, assessing if any of the independent associations colocalise[3]. We first tested for evidence of multiple SNPs independently contributing to IgG glycan levels at the *FUT6* and *FUT8* loci. While no secondary associated variants were observed in the *FUT6* locus using GCTA-COJO stepwise analysis, two independent variants are likely to contribute to variation in transferrin and IgG glycan traits, namely transferrin TfGP32 and IgG GP20, in the *FUT8* locus (Supplementary Data 5). In this case, colocalisation analyses were conducted between full unconditioned association statistics, association statistics conditioned by one association signal (i.e. transferrin TfGP20 conditioned on rs72716459 and IgG GP7 conditioned on rs8022094) and those conditioned by the other association signal (i.e. transferrin TfGP20 conditioned on rs2411815 and IgG GP7 conditioned on rs8006608), for a total of

nine pairwise combinations. We obtained robust evidence against the colocalisation hypothesis for all tested traits, except for transferrin TfGP20 conditioned on rs72716459 and IgG GP7 unconditioned association statistics. In this case in fact it was not possible to strongly support either the hypothesis of different causal variants at the locus or trait colocalisation (PP.H3 = 46.82%, PP.H4 = 53.18%) and therefore whether this glycans pair share genetic architecture at this locus remains unclear (Supplementary Data 15, Supplementary Figure 7-9). It is important to note that another transferrin glycosylation associated genetic region, harbouring *NXEP1* and *NXEP4* genes, was also associated with IgG glycosylation in Klaric *et al*.[5] Since the role of these genes is unknown and therefore interpretation of their role in glycosylation is not straightforward, we did not proceed with colocalisation analysis for this region.

Colocalisation of TfGP32 and plasma glycosylation traits containing antennary fucose

Transferrin is one of the most abundant proteins in plasma and it can be expected it contributes to plasma glycosylation glycan peaks. To assess whether TfGP32 is likely to have antennary fucose we performed colocalisation analysis of TfGP32 and two plasma glycosylation traits containing antennary fucose (PGP32 - A4F1G3S[3,3+6,3+6]3 and PGP36 - A4F1G4S[3,3,3,6]4) from Sharapov *et al*[6]. and a plasma glycosylation trait reflecting total antennary fucosylation (A-FUC) from Huffman *et al*.[7] By using the approximate Bayes factor colocalisation analysis implemented in coloc R package[4] using the default priors (10^{-4} for p1 and p2 and 10^{-5} for p12), we obtained robust evidence supporting the hypothesis that plasma-derived antennary fucosylation traits colocalise with the TfGP32 (Supplementary Figure 4).

Effect of protein levels on glycosylation

Glycosylation was analysed by releasing total N-glycans from the isolated protein and each glycan structure was quantified as the percentage of the total IgG N-glycome or total transferrin N-glycome. With this approach, only changes in glycosylation are detected (relative abundance of individual glycan species in

relation to the whole IgG or transferrin glycome) and not the absolute amounts of specific glycans, which would be affected by changes in protein levels. Unfortunately, abundance of transferrin and IgG was not measured in this study and, consequently, the relation between protein abundance and N-glycan measurements could not be directly investigated. In the case of IgG, we checked whether the specific protocols used for glycan analysis have a bias depending on the initial, already isolated, protein amount. The protocol we used was robust in measuring very similar levels of glycan traits for different IgG abundance (Supplementary Figure 12). In addition, Bermingham *et al.*[8] reported that, while the IgG N-glycan profile varied with IgG levels, adjusting for IgG levels in the analyses made no meaningful difference to associations of glycans with markers of glycaemic control.

To assess the potential impact of transferrin protein levels on transferrin glycome associations we used transferrin cis-protein QTL (pQTL) rs8177240 (LD R^2= 0.02 with the glyQTL rs6785596), the strongest association with transferrin protein levels reported in GWAS catalog (p-value = $8x10^{-610}$)[9], as a proxy for TF abundance. Interestingly, the pQTL is not an expression QTL (eQTL), but rather a splicing QTL for transferrin levels in liver ($p=5.9\times10^{-25}$, GTEx v8). We then tested its association with transferrin glycans and assessed whether the glycan associations with variants from the *TF* region are likely to be driven by this variant. We considered four models:

M0: glycan ~ age + sex

M1: glycan ~ age + sex + pQTL (rs8177240)

M2: glycan ~ age + sex + glyQTL (rs6785596)

M3: glycan ~ age + sex + pQTL (rs6785596) + glyQTL (rs8177240)

and performed likelihood ratio test between:

- M0 and M1 to assess associations of glycans and pQTL (rs8177240)
- M1 and M3 to assess whether glyQTL contributes to glycan levels even when the pQTL is included in the model

- M2 and M3 to assess whether pQTL contributes to glycan levels even when the glyQTL is included in the model

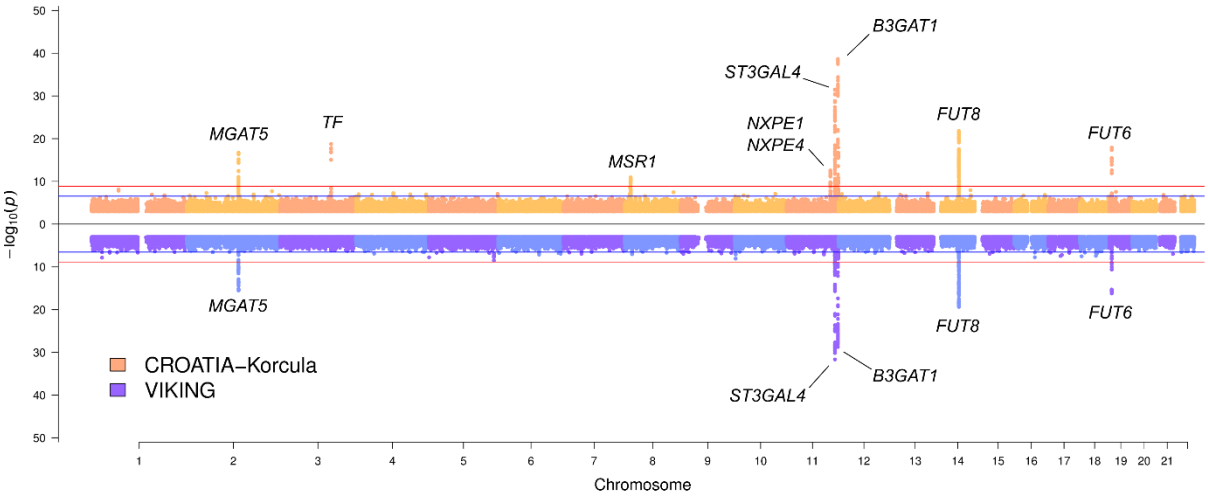
To control for increased levels of relatedness between subjects in our studies, the models were fitted using linear mixed models as implemented in the lme4qtl R package[10], with age, sex, pQTL and glyQTL as fixed effects and kinship matrix as a random effect. The kinship matrix was estimated from the genotyped data using the "ibs" function from GenABEL[11] R package.

Two glycan traits were significantly ($P \leq 0.05/35 = 1.4 \times 10^{-3}$) associated with the pQTL. For one of the two traits, TfGP3, the glyQTL contributes to glycan levels in addition to the pQTL, while for the TfGP9 no additional variation is explained by the glyQTL (Supplementary Data 6). To further corroborate these findings we also repeated the meta-analysis conditioning on the transferrin pQTL rs8177240[9], using the conditional approach implemented in the GCTA-COJO "–cond", and genotypes of 10,000 unrelated individuals of white British ancestry from UK Biobank[12] as independent LD reference panel. The only glycan trait that showed a relevant change in effect size and significance of its association was TfGP9 (Supplementary Data 19), suggesting that its association was dependent on the transferrin protein levels. In case of two glycan traits, TfGP3 and TfGP8, the associations were somewhat less significant, but the effect sizes remained very similar. Accordingly, we consider that transferrin protein levels are likely not affecting associations with 2 out of 3 transferrin glycan traits associated with variants from the *TF* gene. This is in accordance with findings from Kutalik *et al*.[13], who used the same approach to show that associations of a-/disialylated transferrin with the *TF* region were independent of associations with transferrin pQTL.

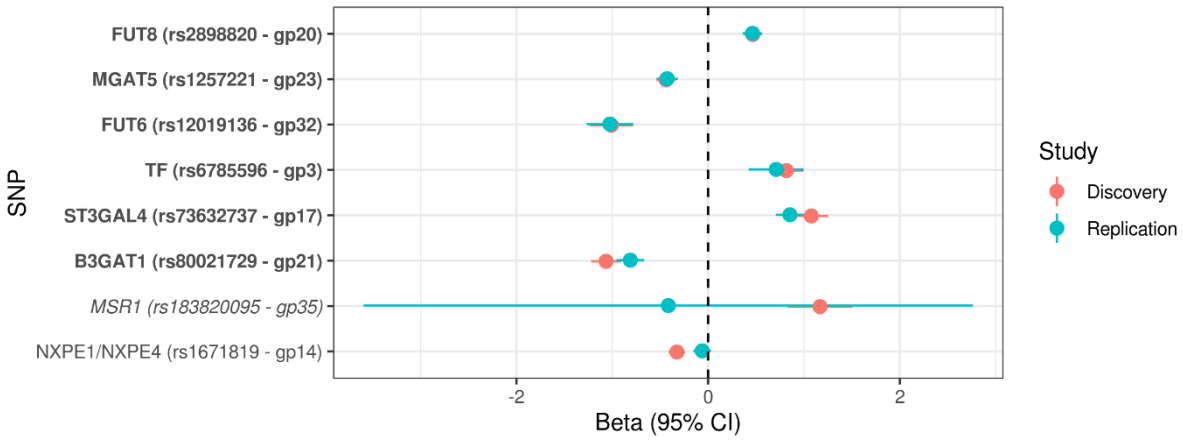
The sentinel glycosylation variant, rs6785596, is a cis eQTL in adipose tissue (Supplementary Data 8a) and it colocalises with TfGP3, but it does not colocalise with expression of *TF* in blood (eQTLGen, Supplementary Figure 13). However, as outlined in the main text, transferrin is predominantly expressed in liver, for which there are no robust transferrin eQTLs (the strongest eQTL in GTEx v8

rs60770862, p=3.3x10^{-6}, LD with glyQTL rs6785596 = 0.0001). The glyQTL variant rs6785596 is also in middling LD (0.57) with a missense variant rs1799899. Overall, further analyses are needed to unravel the complex mechanism behind the associations of transferrin glycans and variants from the *TF* region.

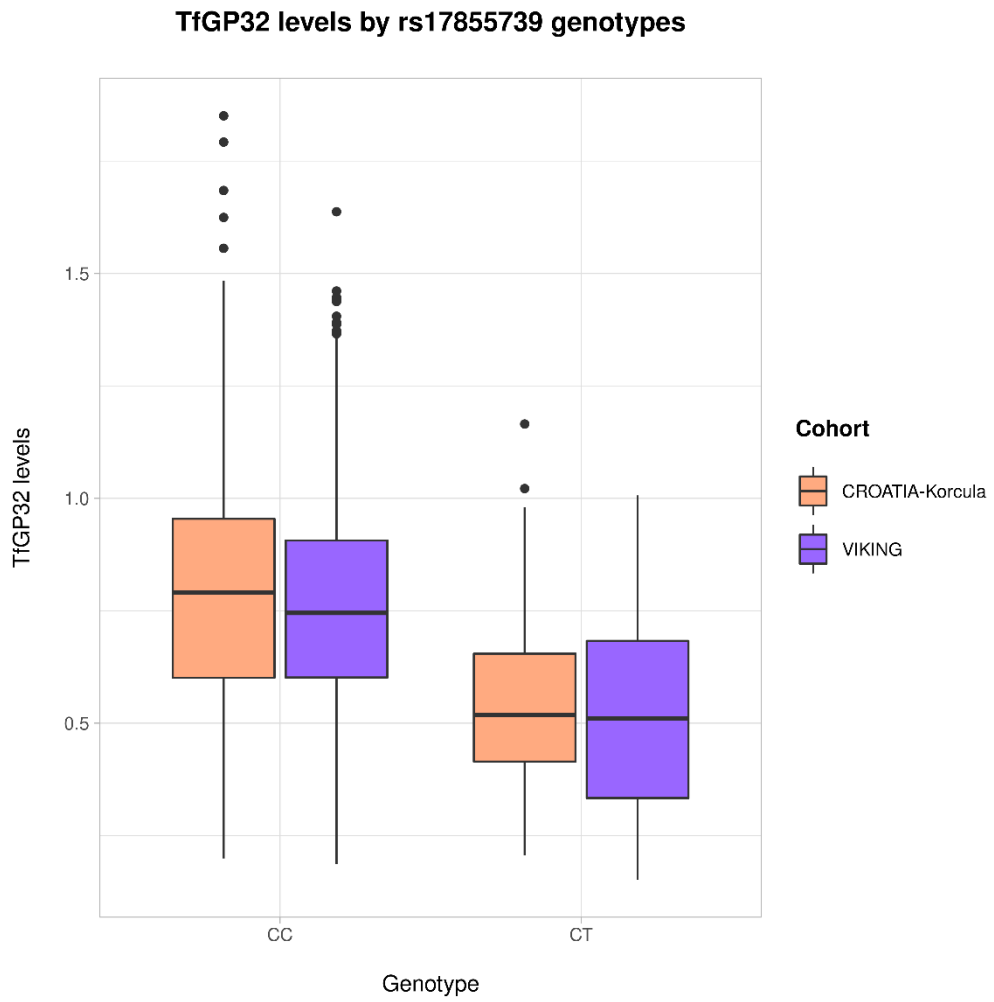
Chapter 2 - Supplementary Figures



Transferrin N-glycome CROATIA-Korcula and VIKING cohorts GWAS summary Miami plot. Miami plot pooling together individual cohort GWAS results obtained across all 35 transferrin glycan traits, at the top in orange for CROATIA-Korcula cohort and at the bottom in blue for VIKING cohort. For each SNP, the lowest p-value across the 35 traits is reported. The y axis shows the strength of the association, and the x axis the genomic position of the SNP. P-values are derived from two-sided Wald test with one degree of freedom. The horizontal red line corresponds to the multiple testing corrected genome-wide significance threshold of 1.43×10^{-9}. The horizontal blue line corresponds to the multiple tests corrected genome-wide suggestive threshold of 2.86×10^{-7}. For simplicity, SNPs with p-value > 1×10^{-3} are not reported. GWAS effect size, standard error and p-value of each sentinel SNP are available in Supplementary Data 1 for CROATIA-Korcula, and in Supplementary Data 2 for VIKING cohort.

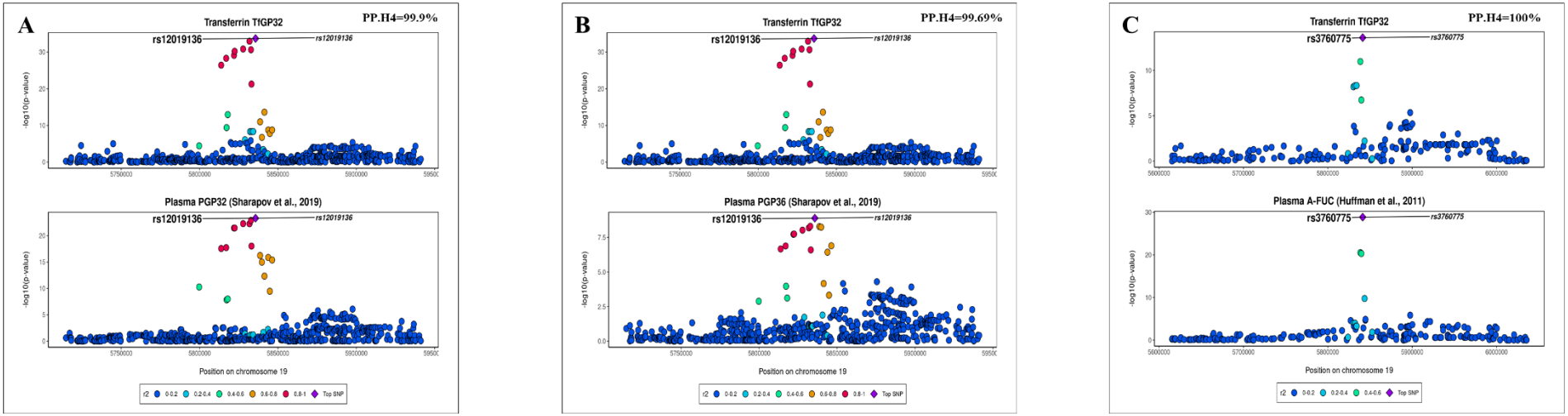


Supplementary Figure 2. Replication of transferrin N-glycome discovery GWAS. Each locus is represented by the strongest SNP-transferrin glycan (reported in brackets) association in the region (y axis). The x axis shows the GWAS effect estimate, with lines representing 95% confidence intervals (CI). For each locus significantly associated with transferrin N-glycome in CROATIA-Korcula cohort, effect size and CI of the sentinel SNP are reported in red when estimated in CROATIA-Korcula cohort (Discovery), and in blue when estimated in VIKING cohort (Replication). Gene names have been marked by different fonts based on overlap between confidence intervals of effect estimates. **In bold:** nominal replication ($p < 0.05/8 = 6.3x10^{-3}$). *In italics:* CIs overlap and cover zero, and replication estimate is closer to zero than discovery. In roman: CIs do not overlap and replication estimate covers zero. Proxy SNP rs554715390 was used for replicating SNP rs183820095 (D'=1, r^2=1). Effect sizes and their standard errors are derived from two-sided linear regression, P-values are derived from two-sided Wald's test with one degree of freedom. GWAS effect size, standard error and p-value of each sentinel SNP are reported in Supplementary Data 1 for discovery, and in Supplementary Data 2 for replication cohort.

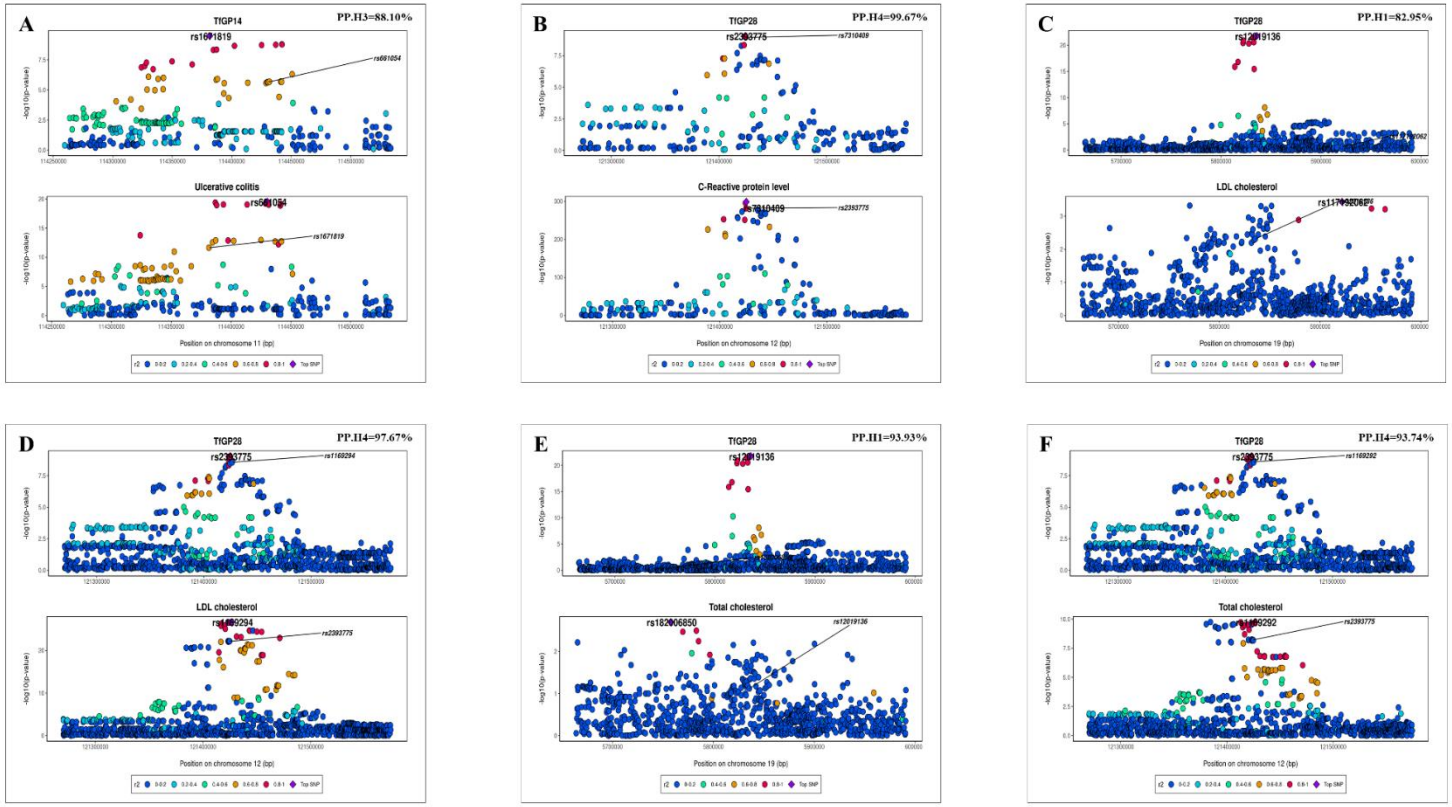


Supplementary Figure 3. **Difference in levels of (normalised and batch corrected) TfGP32 glycan trait by rs17855739 genotypes.** We grouped CROATIA-Korcula (CC genotype N= 859, CT genotype N = 79, TT genotype N = 0) and VIKING (CC genotype N= 890, CT genotype N = 65, TT genotype N = 0) samples based on their genotype at rs17855739 missense variant. rs17855739 reported a genome-wide significant association with TfGP32 glycan trait levels in both CROATIA-Korcula (p-value=4.54×10^{-18}) and VIKING (p-value = 2.00×10^{-16}) cohorts. As showed in the boxplot, the mean level of TfGP32 were lower for individuals heterozygous at rs17855739, compared to those having two C alleles. These results support that rs17855739 missense mutation significantly decrease the levels of TfGP32, which we suggest as a potential proxy of the activity of alpha-(1,3)-fucosyltransferase 6 enzyme (see Supplementary Results). In the plot, the middle line represents the median, lower and upper limits of the box

represent 1st and 3rd quartile, whiskers represent 1.5 interquartile range, points represent individuals with TfGP32 levels that are more than 1.5 interquartile distance away from the 3rd quartile.

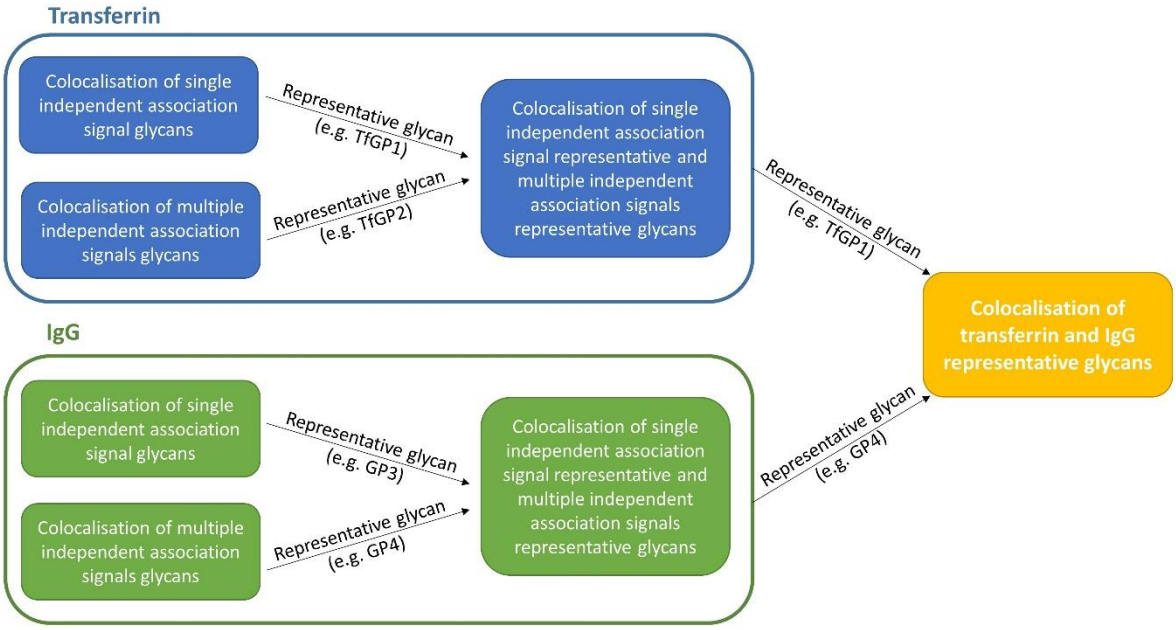


Supplementary Figure 4. Local association patterns of transferrin N-glycan TfGP32 and plasma glycosylation traits related to antennary fucosylation. A) Colocalisation of TfGP32 and total plasma glycan PGP32 - A4F1G3S[3,3+6,3+6]3, PP.H4 = 99.9% B) Colocalisation of TfPG32 and total plasma glycan PGP36 - A4F1G4S[3,3,3,6]4, PP.H4 = 99.69% C) Colocalisation of TfGP32 and total plasma antennary fucosylation (A-FUC), PP.H4 = 100%. Plasma glycosylation traits containing antennary fucose (PGP32 and PGP36) were taken from Sharapov *et al*[6]*.* and plasma glycosylation trait reflecting total antennary fucosylation (A-FUC) from Huffman *et al*.[7] For each pairwise colocalisation test, the hypothesis having the higher posterior probability is reported at the top right of the plot. PP.H4 – colocalisation: the two traits are regulated by the same underlying causal variant.

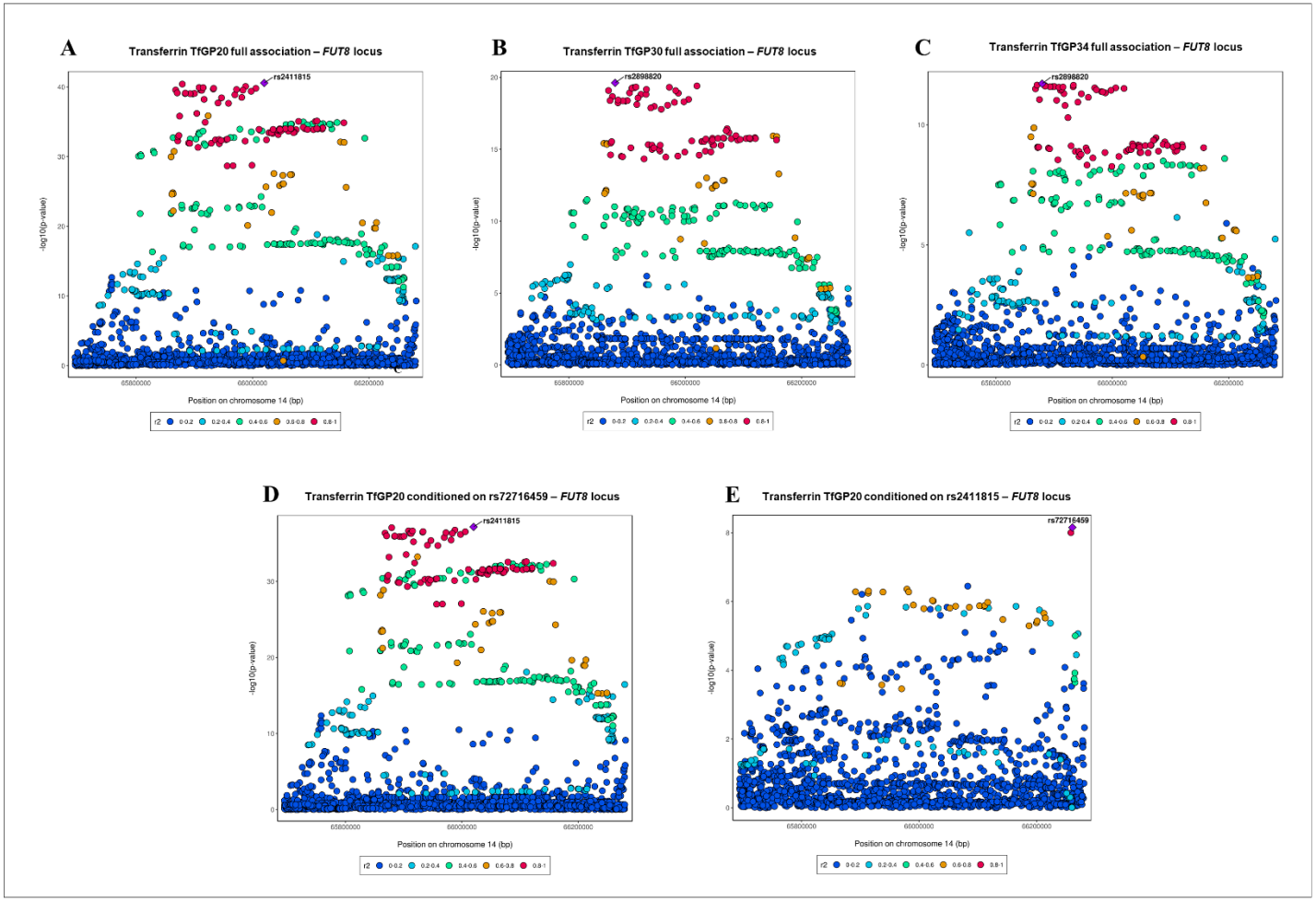


Supplementary Figure 5. Local association patterns of transferrin N-glycans tested for trait colocalisation with complex human traits. GWAS summary statistics of complex traits and diseases were taken from publicly available studies detailed at Supplementary Data 18. A) TfGP14 and Ulcerative colitis, in the *NXEP1/NXEP4* locus. PP.H3 = 88.1% B) TfGP28 and CRP, in the *HNF1A* locus. PP.H4 = 99.7% C) TfGP28 and LDL cholesterol, in the *FUT6* locus. PP.H1 = 83.0% D) TfGP28 and LDL, in

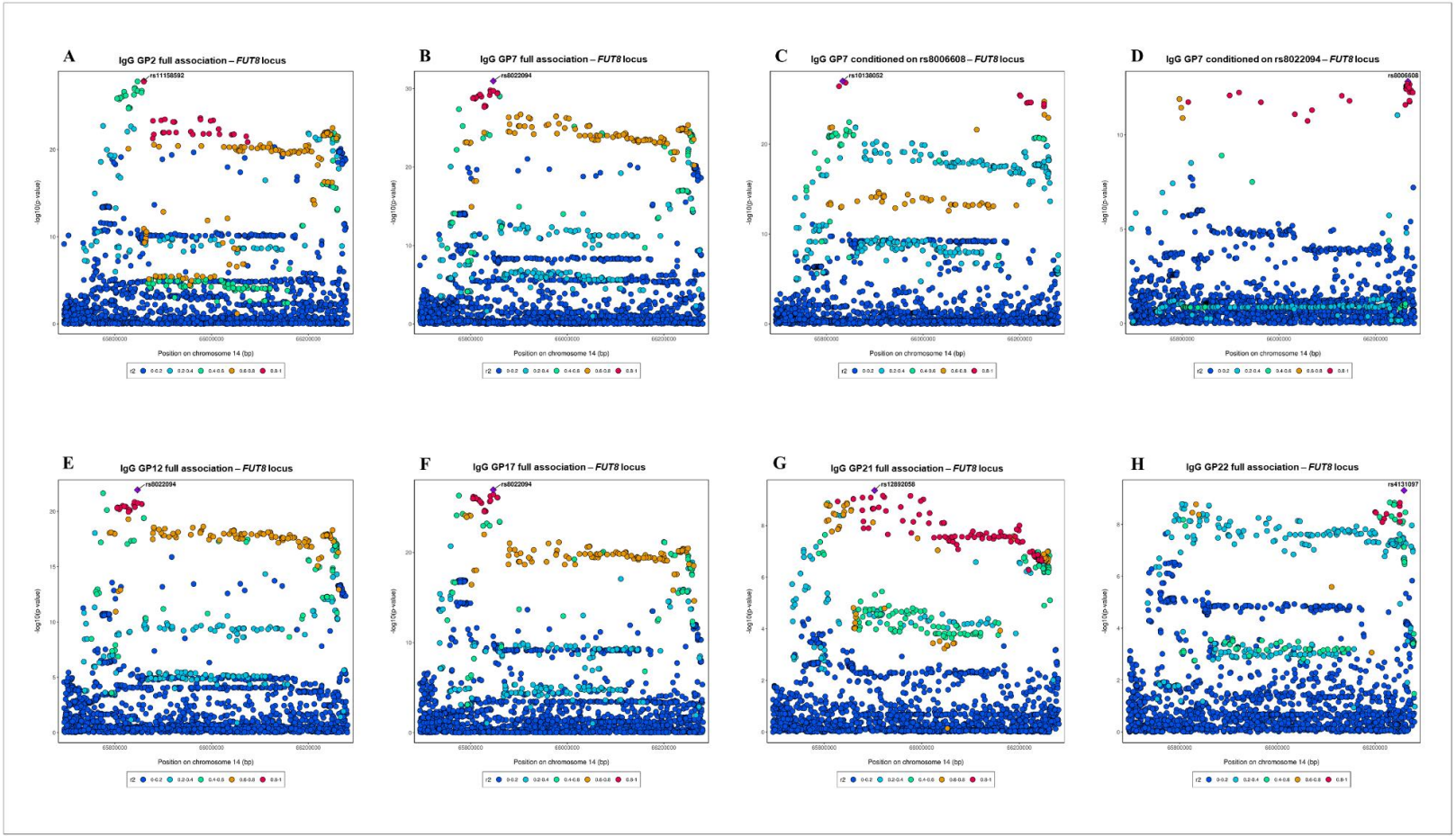
the *HNF1A* locus. PP.H4 = 97.7% E) TfGP28 and total cholesterol, in the *FUT6* locus, PP.H1 = 93.9% F) TfGP28 and total cholesterol, in the *HNF1A* locus. PP.H4 = 93.7%. The full colocalisation analysis results can be found in Supplementary Data 13. PP.H1 – association is observed only in one trait; PP.H3 – two traits are regulated by distinct underlying causal variants; PP.H4 – the traits are regulated by a shared underlying causal variant (colocalisation).



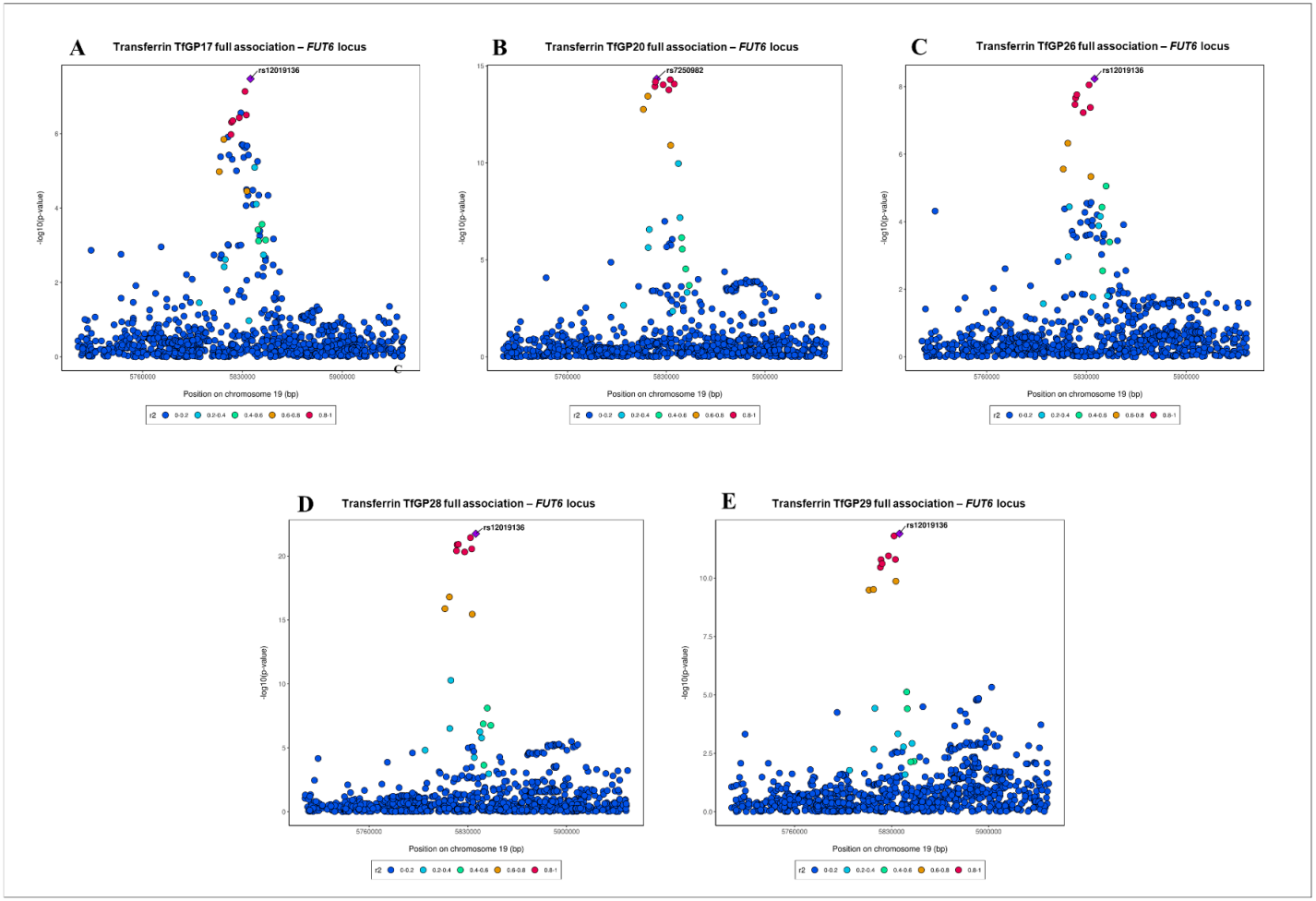
Supplementary Figure 6. Workflow applied for transferrin and IgG glycan traits colocalisation analysis. For each protein (i.e. transferrin and IgG) and each genomic region (i.e. *FUT8* and *FUT6*), the group of glycan traits showing multiple independent signals of association and, separately, the group of glycans carrying only one independent association signal at locus were pair-tested for colocalisation. Pairs of glycan traits with an ABF posterior probability for hypothesis 4 (suggestive of colocalisation) > 80% were pooled in the same colocalisation group. For each colocalisation group identified, the glycan trait reporting the lowest p-value was selected as group representative and carried on to the next step, where traits with single and multiple independent associations for each protein were tested for colocalisation. Similar to previous steps, glycan traits were grouped together on the basis of their colocalisation analysis results and the lowest p-value representative was chosen for the next step, where finally representative transferrin and IgG glycans at each locus were tested for colocalisation.



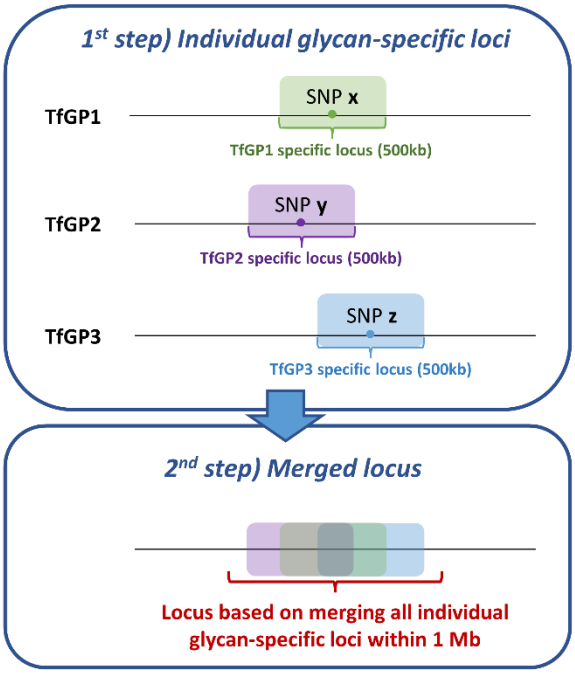
Supplementary Figure 7. Local association patterns of transferrin N-glycans tested for trait colocalisation at *FUT8* locus. Colocalisation analysis results for these glycan traits can be found in Supplementary Data 15 (A, B and C) and 16 (A, D and E). Since TfGP20 also has an independent secondary association signal (see Supplementary Data 5), local association patterns are reported also for conditioned summary statistics.



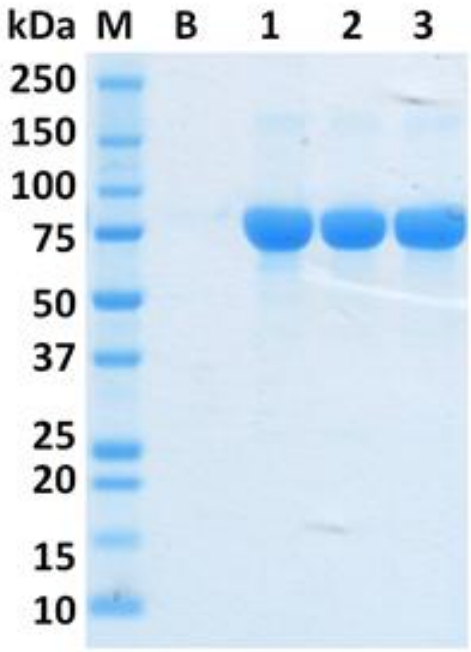
Supplementary Figure 8. Local association patterns of IgG N-glycans tested for trait colocalisation at *FUT8* locus. Colocalisation analysis results for these glycan traits can be found in Supplementary Data 15 (A, B, E, F, G and H) and 16 (B, C and D). Since GP7 also has an independent secondary association signal (see Supplementary Data 5), local association patterns are reported also for conditioned summary statistics.



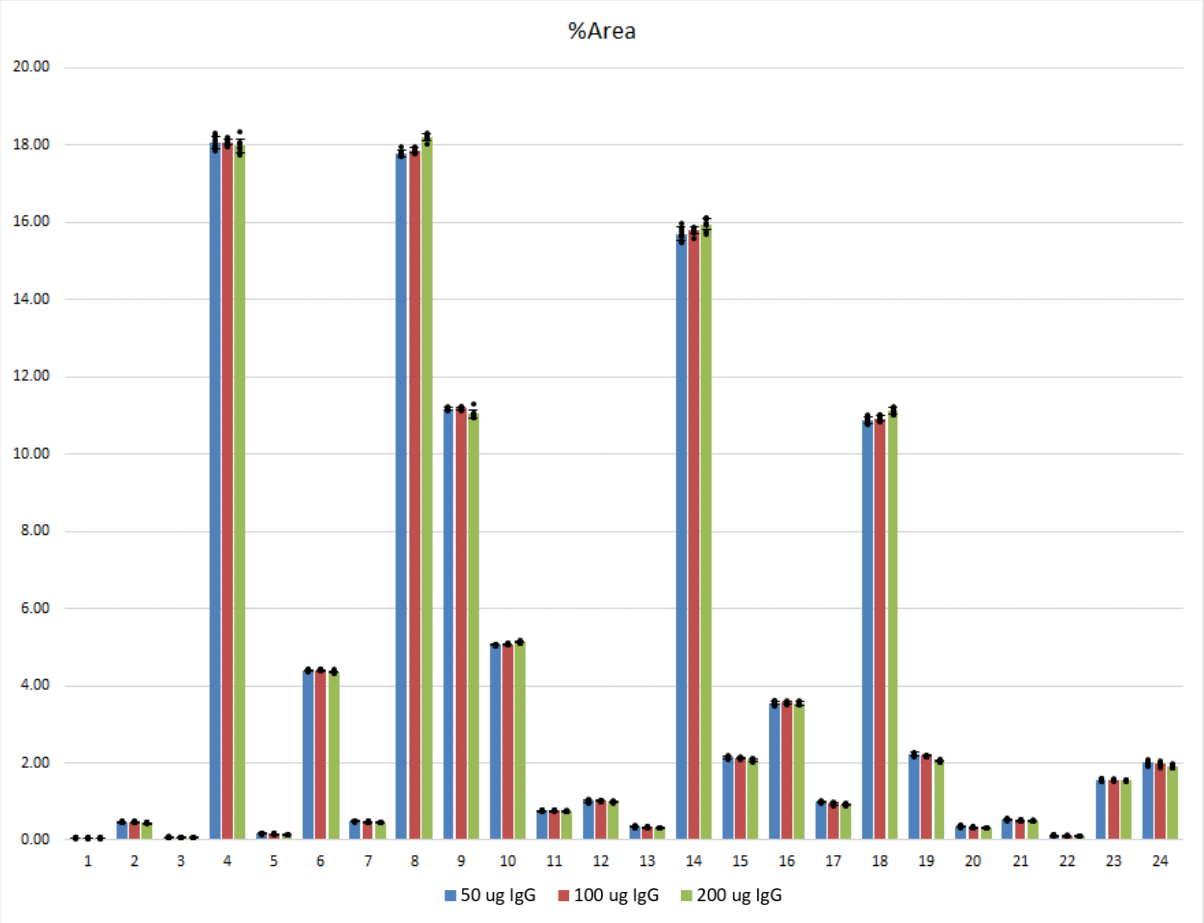
Supplementary Figure 9. Local association patterns of transferrin N-glycans tested for trait colocalisation at *FUT6* locus. Colocalisation analysis results for these glycan traits can be found in Supplementary Data 15 (A, B, C, D and E).



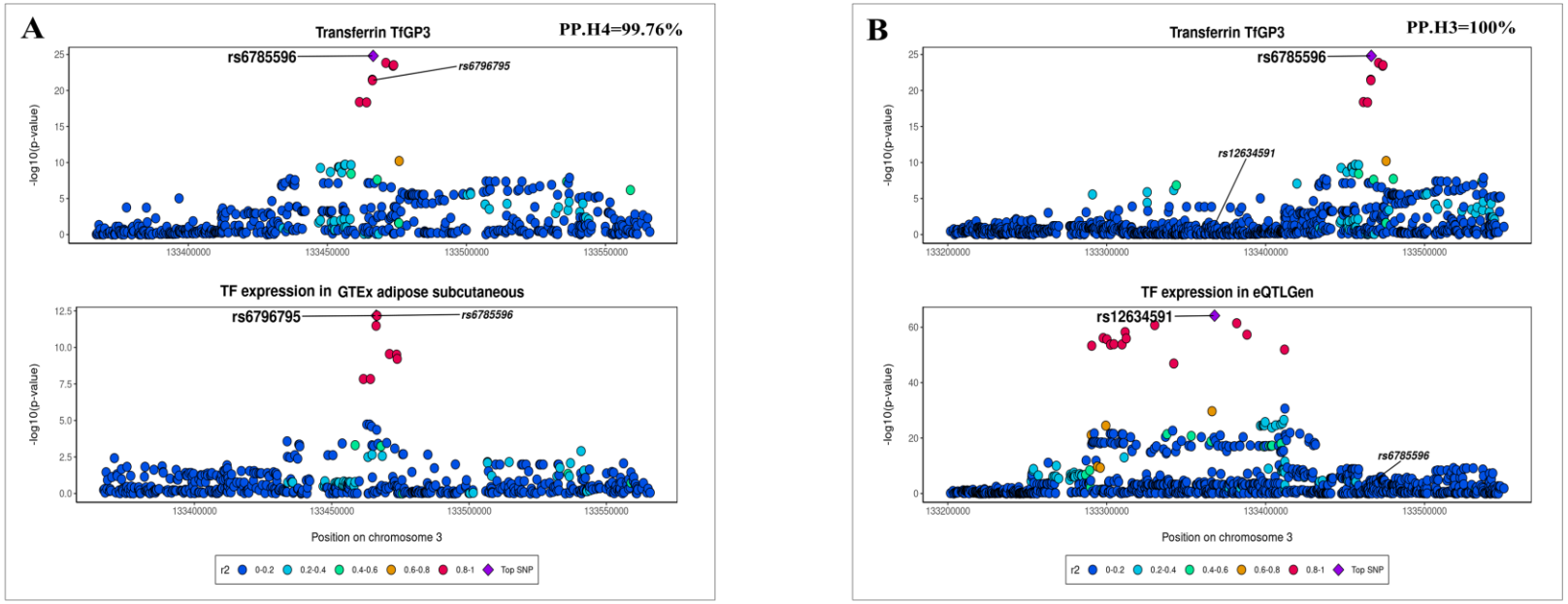
Supplementary Figure 10. Example of the workflow applied for loci definition. To define genomic regions significantly associated with N-glycan traits, we first grouped, for each glycan trait, all genetic variants located within a 500 kb window of the sentinel SNP (+/- 250 kb). With this 1st step we identified loci specific for each glycan trait. Then, to obtain a list of unique loci that are independent of the specific glycan trait, we grouped into a single locus all the glycan-sentinel SNP pairs that were overlapping within a 1Mb range. With this 2nd step we obtained a unique list of top glycan-sentinel SNP pairs, summarising the genomic regions most strongly associated with transferrin N-glycome across all traits.



Supplementary Figure 11. Elution fractions (7.5× concentrated) after transferrin isolation from IgG depleted plasma (lanes 1-3) analysed by SDS-PAGE using 4–12 % Bis-Tris gradient gels (1.0 mm thickness) under reducing conditions according to the manufacturer's instructions (Life Technologies). The gels were run at 200 V for 35 min using a MES SDS buffering system. Protein bands were visualized by GelCode Blue staining reagent. M – Precision Plus Protein Standards All Blue molecular mass standard (Bio-Rad). B – blank sample (10 × concentrated). The same experiment was repeated 32 times (once for each plate) with similar results.



Supplementary Figure 12. Levels of IgG glycan traits (GP1-24) measured for different initial amounts of isolated IgG protein. Data are presented as mean +/- standard deviation, n=8 technical replicates for each amount of IgG. Corresponding data points are overlayed as dots.



Supplementary Figure 13. Local association patterns of transferrin N-glycan TfGP3 and adipose tissue eQTLs (A) and blood (B) at TF gene locus. Adipose subcutaneous eQTLs were taken from GTEx v7[14], and gene expression data blood eQTLs were taken from eQTLGen consortium[15]. For each pairwise colocalisation test, the hypothesis having the higher posterior probability is reported at the top right of the plot. PP.H3 – two traits are regulated by distinct underlying causal variants; PP.H4 – the traits are regulated by a shared underlying causal variant (colocalisation).

Chapter 2 - Supplementary References

1. Trbojević-Akmačić, I. *et al.* Chromatographic monoliths for high-throughput immunoaffinity isolation of transferrin from human plasma. *Croat. Chem. Acta* **89**, 203–211 (2016).
2. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
3. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).
4. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014).
5. Klarić, L. *et al.* Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. *Sci. Adv.* **6**, eaax0301 (2020).
6. Sharapov, S. Z. *et al.* Defining the genetic control of human blood plasma N-glycome using genome-wide association study. *Hum. Mol. Genet.* **28**, 2062–2077 (2019).
7. Huffman, J. E. *et al.* Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults. *Hum. Mol. Genet.* **20**, 5000–5011 (2011).
8. Bermingham, M. L. *et al.* N-glycan profile and kidney disease in type 1 diabetes. *Diabetes Care* **41**, 79–87 (2018).
9. Benyamin, B. *et al.* Novel loci affecting iron homeostasis and their effects in individuals at risk for hemochromatosis. *Nat. Commun.* **5**, 4926 (2014).
10. Ziyatdinov, A. *et al.* lme4qtl: Linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics* **19**, (2018).
11. Karssen, L. C., van Duijn, C. M. & Aulchenko, Y. S. The GenABEL Project for statistical genomics. *F1000Research* **5**, (2016).

12. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
13. Kutalik, Z. *et al.* Genome-wide association study identifies two loci strongly affecting transferrin glycosylation. *Hum. Mol. Genet.* **20**, 3710–7 (2011).
14. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580–585 (2013).
15. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet. 2021 539* **53**, 1300–1310 (2021).

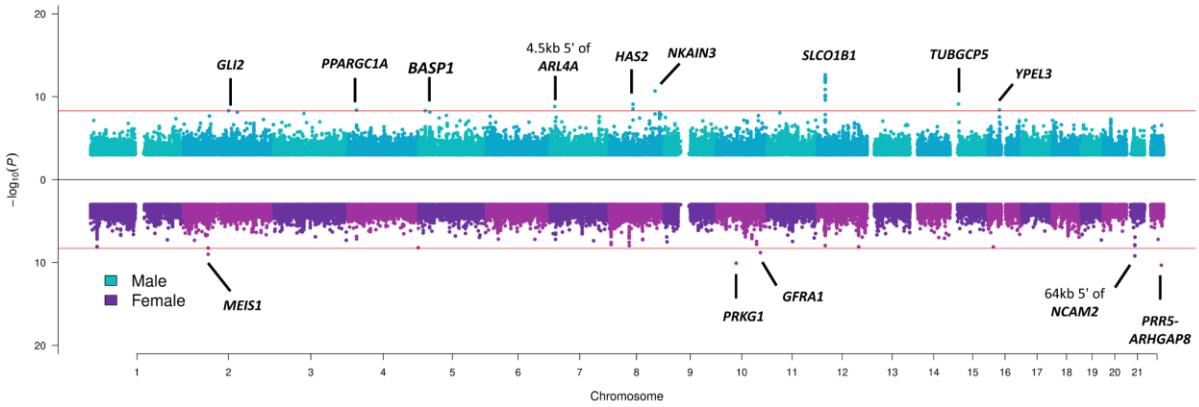
Chapter 2 - Supplementary Tables

The Supplementary Tables related to Chapter 2 “Genetic regulation of transferrin and IgG glycome” are available in the Chapter_2_Extended_Supplementary_Tables.xlsx file, downloadable at the following publicly available link: https://doi.org/10.7488/ds/7509.

Chapter 3 - Supplementary Tables

The Supplementary Tables related to Chapter 3 “Rare and low frequency variants contributing to variation in the protein glycome” are available in the Chapter_3_Extended_Supplementary_Tables.xlsx file, downloadable at the following publicly available link: https://doi.org/10.7488/ds/7509.

Chapter 4 -Supplementary Figure



Supplementary Figure 1. Miami plot pooling together sex-specific meta-analysis results obtained across 14 bile acid traits (quantitative traits with imputed LOD values, as described in Methods), for male at the top in blue, and for female at the bottom in purple. The pooling was performed by selecting the lowest *p* value (y-axis) from the 14 bile acids for every genomic position (x axis). The Bonferroni-corrected genome-wide significance threshold (horizontal red lines) corresponds to 5×10^{-9}. For simplicity, SNPs with *p* value > 1×10^{-3} are not plotted. *P* values are derived from the two-sided Wald test with one degree of freedom.

Chapter 4 - Supplementary Tables

The Supplementary Tables related to Chapter 4 “Genetic architecture of bile acid lipidome” are available in the Chapter_4_Extended_Supplementary_Tables.xlsx file, downloadable at the following publicly available link: https://doi.org/10.7488/ds/7509.