GENOME-WIDE ASSOCIATION STUDY OF THE HUMAN IMMUNOGLOBULIN G N-GLYCOME

Frkatović, Azra

Sveučilište u Zagrebu

FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGY

Azra Frkatović

GENOME-WIDE ASSOCIATION STUDY OF THE HUMAN IMMUNOGLOBULIN G N-GLYCOME

DOCTORAL THESIS

Zagreb, 2022.

Sveučilište u Zagrebu

FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGY

Azra Frkatović

GENOME-WIDE ASSOCIATION STUDY OF THE HUMAN IMMUNOGLOBULIN G N-GLYCOME

DOCTORAL THESIS

Supervisors:
Prof. Gordan Lauc, PhD
Frano Vučković, PhD

Zagreb, 2022.

Sveučilište u Zagrebu

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

BIOLOŠKI ODSJEK

Azra Frkatović

CJELOGENOMSKA ASOCIJACIJSKA STUDIJA N-GLIKOMA LJUDSKOG IMUNOGLOBULINA G

DOKTORSKI RAD

Mentori:

Prof.dr.sc. Gordan Lauc

Dr.sc. Frano Vučković

Zagreb, 2022.

This work was done in the Glycoscience Research Laboratory at Genos, Zagreb, under the supervision of Prof. Gordan Lauc and Dr. Frano Vučković. This thesis is submitted for review to Department of Biology at the Faculty of Science, University of Zagreb as a part of the Postgraduate doctoral programme of Biology to achieve the academic degree Doctor of Biology.

Supervisor biography

Gordan Lauc is the Professor of Biochemistry and Molecular Biology at the University of Zagreb, Director of the National Centre of Scientific Excellence in Personalised Healthcare, as well as honorary professor at the Kings College London, University of Edinburgh and a member of the Johns Hopkins Society of Scholars. Since 2017, he is one of the two co-directors of the Human Glycome project which was initiated in the same year. In 2007, he established Genos, a company that is a current global leader in high-throughput glycomics. His research team specializes in high-throughput glycomics and exploiting the biomarker potential of glycans in the field of precision medicine. They focus on understanding the role of glycans in normal physiology and disease by combining glycomic data with genetic, epigenetic, biochemical and physiological data in a systems biology approach. Professor Lauc co-authored over 250 research articles that are cited over 6,000 times. He was PI and co-PI in NIH, FP6, FP7, H2020, and ESI Funds-funded projects, five of which he coordinated.

Dr. Frano Vučković graduated Molecular Biology at the University of Zagreb, Faculty of Science in 2012, and obtained doctoral degree in Molecular Biology from the University of Zagreb in 2016. Since 2012 he is employed at Genos Ltd. working as a data analyst and a researcher in Glycoscience Laboratory. In 2018 he was appointed Head of Data Analysis. His work has been oriented on developing new models for integrative analysis of glycomics and other -omics data as part of MIMOmics and IMforFUTURE projects. His main research interest includes development of normalization and batch correction methods for high-throughput glycomics data. He is also involved in research of regulation of IgG N-glycosylation in autoimmune diseases. He co-authored more than 40 research papers that are cited more than 2,200 times

ACKNOWLEDGEMENTS

I would like to thank my mentors and supervisors Lucija Klarić, Frano Vučković, and Gordan Lauc for their support, valuable advice, and an opportunity to learn from them.

My immense gratitude goes to all partners in IMforFUTURE project for the support they provided throughout the three years, especially Professor Jim Wilson from the University of Edinburgh and Professor Frances Williams from King's College London for having me as a guest in their lab groups. I thank all collaborators for agreeing to be part of the study and for providing their data.

I thank all early-stage researchers in IMforFUTURE project for their support and advice they shared whenever I needed help with my work, all the great and fun gatherings we had and the amazing memories I collected with them all over Europe.
Big thanks to all my colleagues and friends in Genos who made my work and stay in Croatia such an amazing experience.

I thank my parents, Sevla and Hasan, my sister Ajla for always believing in me and supporting me during my journey of becoming a scientist, my friends and family back in Bosnia and Herzegovina for their encouragement and help throughout my whole education.

University of Zagreb
Faculty of Science
Department of Biology

Doctoral thesis

GENOME-WIDE ASSOCIATION STUDY OF THE HUMAN IMMUNOGLOBULIN G N-GLYCOME

AZRA FRKATOVIĆ
Genos Glycoscience Research Laboratory

Glycosylation of the immunoglobulin G (IgG) is a complex biosynthetic process that affects protein stability and effector functions. Genome-wide association study of IgG N-glycan traits in seven cohorts of European descent was conducted to discover genomic regions associated with IgG glycosylation. The tested phenotypes were defined as the percentage of presence of individual sugar units in the total IgG glycome. Discovery meta-analysis resulted in 42 genome-wide significant loci, 13 of which were novel. Prioritization of the genes mapped in the associated regions resulted in 83 candidate genes which are enriched in cellular transport, B and T cell development and immune system, suggesting that not only the regulation of glycoenzyme gene expression but intracellular transport and clonal selection and proliferation shape the IgG glycopatterns. The findings provide the basis for functional studies to confirm the role of prioritized genes in IgG N-glycosylation.

142 pages, 34 figures, 24 tables, 233 references, original in English
Keywords: Genome-wide association study; immunoglobulin G; N-glycosylation; immune system; pleiotropy

Supervisors: Prof. Gordan Lauc, PhD
Res. Assoc. Frano Vučković, PhD
Reviewers: Prof. Olga Gornik Kljaić, PhD
Res. Assoc. Lucija Klarić, PhD
Prof. Kristian Vlahoviček, PhD

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Doktorski rad

CJELOGENOMSKA ASOCIJACIJSKA STUDIJA N-GLIKOMA LJUDSKOG IMUNOGLOBULINA G

AZRA FRKATOVIĆ
Laboratorij za glikobiologiju Genos

Glikozilacija imunoglobulina G (IgG) je kompleksna postranslacijska modifikacija koje utječe na stabilnost proteina i njegove efektorske funkcije. Provedena je cjelogenomska asocijacijska studija N-glikanskih svojstava ljudskog IgG-a u sedam skupina ispitanika europskog podrijetla s ciljem otkrivanja novih genomskih regija povezanih s glikozilacijom IgG-a. Testirani fenotipovi definirani su kao postotak prisutnosti dane jedinice šećera u ukupnom IgG glikomu. Početna meta-analiza identificirala je 42 statistički značajna lokusa, od kojih je 13 novotokrivenih. Prioritiziranje gena mapiranih u otkrivenim regijama rezultiralo je s 83 gena koja su obogaćena u skupovima vezanim za stanični transport, razvoj B i T stanica te imunološki sustav, što sugerira da pored regulacije ekspresije gena za glikoenzime, unutarstanični transport i klonska selekcija i proliferacija oblikuju glikoprofil ljudskog IgG-a. Rezultati ove studije predstavljaju osnovu za funkcionalne studije kako bi se potvrdila uloga prioritiziranih gena u N-glikozilaciji.

142 stranice, 34 slike, 24 tablice, 233 literaturna navoda, jezik izvornika engleski
Ključne riječi: Cjelogenomska asocijacijska studija; imunoglobulin G; N-glikozilacija; imunološki sustav; pleiotropija

Mentori: Prof. dr. sc. Gordan Lauc
Dr. sc. Frano Vučković, znan. sur.
Ocjenjivači: Prof. dr. sc. Olga Gornik Kljaić
Dr. sc. Lucija Klarić, znan. sur.
Prof. dr. sc. Kristian Vlahoviček

TABLE OF CONTENTS

1. INTRODUCTION

The complex function of proteins is largely determined by the post-translational modifications, such as the covalent attachment of a sugar chain to the polypeptide backbone called glycosylation. Nearly all membrane and secreted human proteins are glycosylated[1]. The attached sugar chains are referred to as glycans and glycoconjugate formed by sugars attached to the polypeptide backbone is called glycoprotein.

One such glycoprotein is immunoglobulin G (IgG), the most abundant immunoglobulin in the human serum accounting for 75% of all immunoglobulins and 10-20% of the total human plasma proteome[2]. IgG is involved in several humoral immune system pathways including antigen neutralization, target opsonization for phagocytosis, complement activation, antibody-dependent cell-mediated cytotoxicity (ADCC), complement-dependent cytotoxicity (CDC) and hypersensitivity reactions. Given its importance in multiple immune processes, IgG is being actively studied.

IgG is composed of four polypeptide chains, two identical heavy chains and two identical light chains linked by inter-chain disulfide bonds. The structure of IgG can also be divided into two regions based on the functional properties, Fc (constant region) and Fab (antibody binding region). The Fc portion of the IgG molecule contains a conserved N-glycosylation site on the Asn-297 on each heavy chain[3]. The attached glycans modulate effector functions of IgG via fine-tuning the Fc conformation[4].

All N-glycans that can be found on IgG have a conserved core structure containing two N-acetylglucosamine (GlcNAc) and three mannose units which branch into two antennae, commonly expanded by two GlcNAc residues [135]. The core structure is further expanded by the addition of bisecting GlcNAc, core fucose, galactose, and lastly N-acetylneuraminic (sialic) acid. High-throughput methods for the analysis of N-glycans attached to IgG using ultra-performance liquid chromatography (UPLC) and liquid chromatography coupled with mass spectrometry (LC-MS) were developed, thereby enabling analysis of large cohorts containing thousands of samples[5]. The availability of IgG N-glycan data facilitated epidemiological studies exploring the association of IgG N-glycan changes with a range of pathological and physiological states[6], as well as genome-wide association studies (GWAS) which test for the association of glycans and SNPs across the human genome to identify genetic factors involved

in the process of IgG N-glycosylation. A total of 29 genomic regions has been implicated in GWAS of IgG N-glycome[7–10]. Among the discovered genomic loci, there are four containing glycosyltransferase genes, *B4GALT1, ST6GAL1, FUT8,* and *MGAT3*, genes encoding enzymes that catalyse the addition of galactose, sialic acid, core fucose and bisecting GlcNAc residues to the glycan chain, respectively. The remaining loci harbour genes encoding transcription factors, co-factors, transport proteins, as well as additional genes with no apparent role in IgG glycosylation. The identified loci were also shown to be pleiotropic with inflammatory diseases including rheumatoid arthritis (RA), inflammatory bowel disease (IBD), ulcerative colitis (UC), Crohn's disease (CD), asthma, primary biliary cirrhosis (PBC) and Parkinson's disease (PD)[10]. The expansion of the number of loci associated with IgG N-glycosylation would allow for assessment of biomarker potential of IgG for mentioned but also other diseases and understanding its possible role in the disease pathophysiology.

When aiming to functionally test the findings of the GWA study, careful inspection of the genomic regions must be undertaken to prioritize and choose the genes with strong evidence for a role in the biological pathway. Therefore, prioritization includes a set of *in silico* methods such as exploring the functional consequences of identified variants, the pleiotropy with gene expression, the effect of the variants on distant genes via chromatin interaction and gene-based association analysis. Once candidate genes are chosen, the hypothesis can be set and functional follow-up can be performed in the appropriate biological system.

GWAS is a hypothesis-free approach and is mainly used to generate knowledge that will enable the setting of new hypotheses. Previous GWA studies of the human IgG N-glycome provided a list of candidate genes that have a potential role in IgG glycosylation. In this study, we increase the number of samples and increase the power to detect novel genomic loci. Therefore, we set the hypothesis for this research as follows:

- Increased sample size in genome-wide association meta-analysis of the human IgG N-glycome increases statistical power to detect novel candidate genes associated with IgG N-glycosylation and allows setting of hypotheses that are testable in the functional study.

The main aim of the thesis is to provide a deeper understanding of the genes involved in the process of IgG N-glycosylation by means of GWAS conducted in a large number of participants. The workflow can be broken down into four objectives:

1. Find an appropriate pre-processing and harmonization method of IgG N-glycan values measured by UPLC and LC-MS to enable joint analysis

2. Perform GWAS and meta-analysis of GWAS summary statistics

3. Prioritize genes in discovered genomic loci

4. Set hypothesis for functional studies based on the results of post-GWAS *in silico* methods

2. LITERATURE REVIEW

2.1 Glycosylation

There is a common assumption that the function of the protein is defined by the structure encoded by the corresponding gene. However, the biologically complex functions of the proteins are largely defined by the post-translational modifications including the covalent attachment of sugars or sugar chains called glycosylation. Nearly all membrane and secreted proteins are glycosylated[1]. The attached sugar molecules are referred to as glycans. When present on the cell surface, glycans play a role in various events, such as cell-cell, cell-matrix, or cell-molecule interactions important for the development and function of multicellular organisms. Glycans can also be specifically bound to a protein where they can serve as regulators of protein functions, play a role in signalling, transport, and protein-protein interactions[11].

Glycoconjugate formed by the protein and sugars attached to its polypeptide backbone is called glycoprotein. The common classification of glycans is defined by the linkage to the protein, either N- or O-linkage. In this work, we focus on N-glycans attached to the human IgG. N-linked glycan or N-glycan is a sugar chain covalently attached to an Asn residue of polypeptide chain which usually involves a GlcNAc (N-acetylglucosamine) residue and a peptide sequence as Asn-X-Ser/Thr, where X is any amino acid except Pro.

As opposed to protein sequences which are directly encoded in genes, glycan structures are not primary gene products but rather the secondary gene product. There are hundreds of genes in the human genome that code for enzymes and transporters involved in glycan synthesis[12]. Glycan represents numerous combinatorial possibilities, as determined by many competing acting enzymes and the assembling process in the Golgi apparatus (GA) and endoplasmic reticulum (ER) of eukaryotic cells. In the case of N-glycans, the partial assembly occurs on the cytoplasmic side of ER follow by the flip across the membrane where the assembly is continued and transferred to the target protein. The oligosaccharide is further trimmed and the addition of monosaccharides occurs as protein travels through ER and GA[13]. The synthesis of sugar donors from the precursors occurs in cytosolic or nuclear compartments followed by the transport across membrane bilayer into ER lumen and GA. The addition of monosaccharides to the growing glycan chain is catalyzed by enzymes called glycosyltransferases[11].

The complex regulatory network behind protein glycosylation comprises hundreds of components including enzymes, transcription factors, transporters and other proteins. The synchronized action of these components in glycan biosynthesis and attachment is highly dependent on the genetic sequence but also regulatory mechanisms controlling the expression of involved genes. Nonetheless, evidence suggests that complex interaction between environment and genetic sequence has a vast impact on glycan biosynthesis resulting in immediate glycan change or creating lasting modifications that are maintained through epigenetic mechanisms[14].

The effects of altered glycosylation can range from being undetectable to a complete loss of function. The defects in N-glycan biosynthesis result in disorders that can manifest across multiple systems, including visual, hepatic, nervous and immune systems. The complete loss of N-glycans is lethal, thus making the congenital disorders of glycosylation (CDG) rare [22]. Many other nonmendelian human diseases can be caused by acquired changes in glycan biosynthesis and signalling including immune, nervous, cardiovascular, gastroenterological, haematological, and nervous system disorders, as well as cancers and infectious diseases[11].

2.2 Immunoglobulin G

Antibodies also called immunoglobulins are part of the defense mechanism our body uses to fight infection by pathogens like viruses and bacteria. The binding of immunoglobulins to the pathogens activates the complement, a system of white blood cells and blood proteins, which then act together to inactivate and remove the invaders. Immunoglobulins are synthesized by B cells and all immunoglobulins produced by the same cell contain the same antigen-specific binding site. Naïve B cells produce immunoglobulins and express them on their surface, while the antibody-secreting plasma B cells produce and secrete immunoglobulins without presenting any on the surface. There are five immunoglobulin classes produced in humans: IgA, IgG, IgM, IgE and IgD, each having a specific function in the downstream immune response. The focus of this study is IgG[15].

IgG is a complex protein with an essential role in the humoral immune response in humans. IgG is highly abundant in human blood accounting for 10-20% of the plasma proteins, and one of the most studied glycoproteins. IgG is further subdivided into four classes: IgG1, IgG2, IgG3 and IgG4[16]. IgG subclasses are 90% identical in the amino acid sequence, however, each of the

subclasses has unique properties related to their function, such as antigen binding, complement activation, effector cells activation, placental transport and formation of immune complexes[17].

IgG comprises of two identical heavy chains and two identical light chains linked by inter-sulfide bonds (Figure 1). Each heavy chain contains N-terminal variable domain (VH) and three constant domains (CH1, CH2, CH3), and a hinge region between CH1 and CH2. The light chains are composed of one N-terminal variable domain (VL) and one constant domain (CL). The light chain is linked to VH and CH1 domains, thereby forming a fragment antigen-binding or Fab region. Lower hinge region and CH2-CH3 domains together form Fc or fragment crystalline[17].

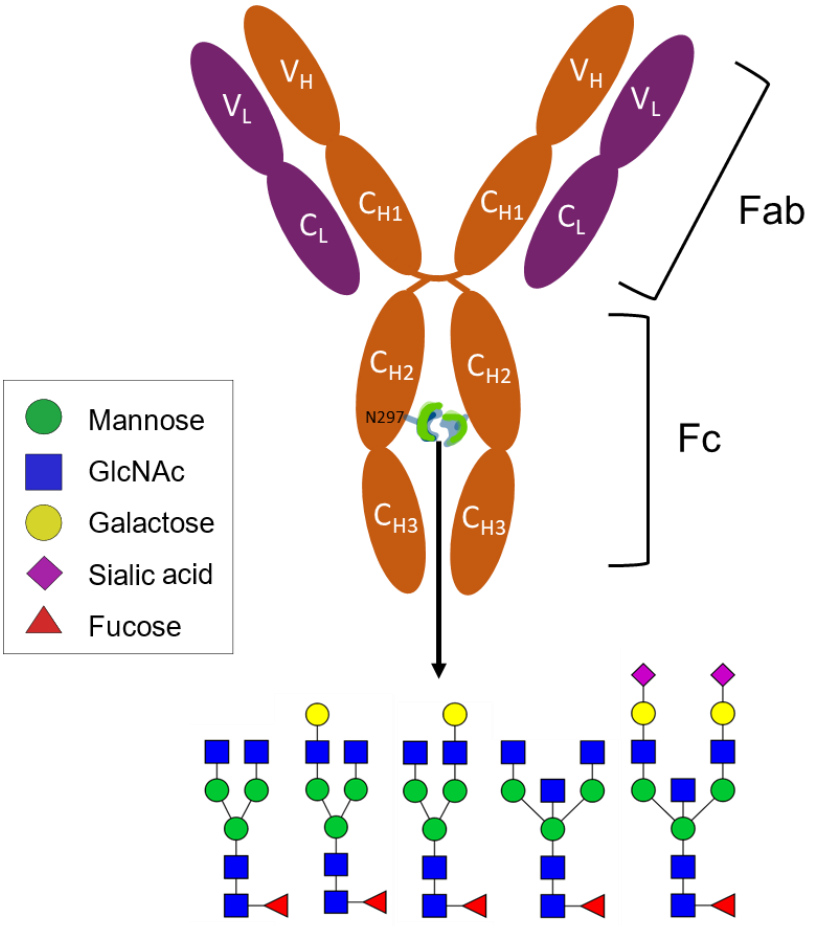


Figure 1: Structure of human IgG with examples of glycan structures that can be attached to Fc region.

2.3 N-glycosylation of immunoglobulin G

Each CH2 domain of the Fc region on IgG carries a single covalently bound N-glycan at the conserved Asn-297 residue. Additionally, around 15-25% of Fab regions have N-glycosylation

site[18]. The core structure of IgG N-glycans is composed of GlcNAc and mannose residues which can further be expanded by the addition of galactose, sialic acid, core-fucose, and bisecting GlcNAc units. When comparing Fab and Fc glycans, the differences are commonly observed in higher levels of bisecting GlcNAc, galactose and sialic acids in Fab glycans and lower levels of fucose. The noted differences can partially be explained by the restricted access for the glycosyltransferases and glycosidases to the Fc region as compared to the Fab region.

The change in the composition of the attached N-glycans on IgG can modulate the structural conformation, thus resulting in changes in the effector function of IgG. One of such changes is the addition of fucose residue to the core of the glycan structure, often referred to as core-fucose. This modification results in a change of conformation of the Fc region, thereby reducing the ability of IgG to bind FcγRIIIa[19,20,21], an activating Fc receptor expressed in natural killer (NK) cells. The core-fucose is present in over 95% of the circulating IgGs where it acts as a safety switch to reduce the initiation of antibody-dependent cellular cytotoxicity (ADCC) which results in the destruction of the targeted cells[22].

Changes in bisection levels of IgG were observed in antigen-specific responses of IgG[23,24], however, they are hard to distinguish from the effects of altered fucosylation levels due to the reciprocal manner that the fucosylation and bisection occur in[25,26].

In the case of galactosylation, decreased levels have been observed in multiple autoimmune diseases[27], with the initial discovery in rheumatoid arthritis (RA) where decreased levels of galactosylated structures were observed[28]. Lack of galactose residues facilitates the activation of complement via mannose-binding proteins[29]. Recent findings have also indicated the changes in galactosylation with aging[30], however, the underlying mechanism and function are still unknown.

The addition of sialic acid to the end of IgG N-glycans has been shown to have the most prominent effect on the Fc region as it closes the binding site for Fc receptors[31] and in parallel, opens the binding sites for DC-SIGN in the CH2-CH3 region[32]. The resulting change in conformation converts the IgG from proinflammatory to anti-inflammatory agent[33]. The sialylation of IgG has been used for the preparation of intravenous immunoglobulin (IVIG) which is utilized in therapy for numerous autoimmune diseases.

2.4 High-throughput measurement of IgG N-glycans

The set of glycan structures that are expressed in a specific cell type or organism is referred to as glycome. Glycomics represents the systematic characterization of glycome of a given cell, organism or protein, usually consisting of the release of glycans from the given entity and their characterization using mass spectrometry. The bottleneck for any large-scale epidemiological study of IgG N-glycome has been the isolation of the protein from the large set of samples, therefore, the development of a high-throughput quantification method based on ultra-performance liquid chromatography (UPLC) enabled such studies. The method includes isolation of IgG from serum or plasma sample using protein G plates, the enzymatic release of the N-glycans from its surface and their quantification[5]. The resulting profile consists of 24 peaks, each representing at least one glycan structure which can be either released from Fc or Fab region on IgG. Another high-throughput approach based on liquid chromatography coupled with mass spectrometry consists of the enzymatic digestion of IgG protein to obtain subclass-specific Fc glycopeptide species which are separated by liquid chromatography and submitted to quantification by mass spectrometry[34]. The main difference between the two approaches lies in the subclass-specific measurements obtained by LC-MS as opposed to total IgG measurements by UPLC. Additionally, UPLC data includes Fab glycans, while LC-MS data is limited to the glycans from the Fc portion of IgG.

2.5 Genome-wide association studies

2.5.1 Single nucleotide polymorphisms

The primary goal of human genetics is to identify genetic factors underlying common and rare diseases in the population, as well as indicate the main genetic players involved in specific processes in the human body. One of the widely used approaches is the genome-wide association study or GWAS which aims to identify the common genetic variations associated with the phenotype of interest[35]. GWAS relies on the measured genetic variation across the human genome, such as single nucleotide polymorphisms (SNPs). SNPs represent a single nucleotide change in the DNA which occurs in high frequency in the genome, and as such represent the most frequent type of the genetic variation[36]. Being the common type of genetic variation, most of the SNPs are present in a large proportion of the population[37]. Commonly, the frequency of a SNP is denoted by the minor allele frequency (MAF), the allele which is less common in the studied population. Many SNPs simply represent the marker for the genomic region and are found in the noncoding region of the gene or between the genes. As

such, they don't have a clear impact on the encoded protein and downstream biological pathways. However, some of the SNPs can be found in the regulatory regions where they affect the expression of the gene or they can cause a change in the amino acid sequence, thus having significant functional consequences[38].

2.5.2 Genotyping

Large-scale genotyping utilizing the chip-based microarrays made GWAS possible as this technology enables assays of more than one million SNPs. The two competing platforms commonly used are Affymetrix (Santa Clara, CA) and Illumina (San Diego, CA).

The Affymetrix microarray enables the selection of variants to be genotyped from their database, followed by the printing of the DNA probes on the chip which recognize the allele at the specific genomic position. The allele in the sample DNA is identified by the match or mismatch with the probes for the targeted variant[39].

On the other hand, Illumina genotyping platform is based on the bead array technology which can genotype around 4 million variants per sample. The silica beads are placed in microwells coated with probes, each representing a specific genomic locus that ends just before the position of interest. The elongation of the DNA fragments with fluorescently labeled nucleotides enables the detection of the allele at the specific position[40].

2.5.3 Linkage disequilibrium

SNPs are non-randomly distributed across the human genome and are often correlated with the nearby SNPs representing the phenomenon called linkage disequilibrium (LD). During the meiosis process, the SNPs in high LD are inherited together. LD in humans is present in the regions on the same chromosome which have a low recombination rate[35]. LD measures can be expressed as D, D′ and r^2 which represent the difference between the observed frequency if two alleles co-occur in the population and the expected frequency if the two markers are independent. D' is a measure related to recombination events between markers and it is scaled between 0 and 1, where D'=0 indicates complete linkage equilibrium and D'=1 indicates complete linkage disequilibrium, meaning that there are no recombination events between the two markers in the population. For genetic analysis, r^2 measure is used to report LD between two SNPs and it is a measure of correlation where high r^2 values indicate that two SNPs carry

similar information in a way that one allele of the first SNPs is frequently observed with one allele of the second SNP within the population[35].

Tag SNPs represent SNPs that are selected to capture the variation of the SNPs in the surrounding stretch of LD. It is important to note that the LD patterns are population-specific due to multiple factors including the size of the population, number of founding populations, and number of generations, which all contribute to the LD decay. The tag SNPs are used to minimize the redundant information produced by genotyping the SNPs[41]. According to the data from the HapMap project[42], more than 80% of common SNPs in European populations are captured by genotyping the subset of 500 thousand to one million SNPs.

2.5.4 Genotype Imputation

Genotyping arrays cover only the limited number of SNPs across the whole genome. Genotype imputation is a procedure used to call the variants which remain ungenotyped to maximize the SNP number tested in the GWAS but also provide the same coverage for the study groups which were originally genotyped using different arrays[43]. Imputation is performed by utilizing the LD pattern of the SNPs in the haplotype block in the reference population. Commonly used software for genotype imputation includes SHAPEIT2[44], IMPUTE2[45], MACH[46] and Minimac3[47]. For the studies involving participants of European descent, there are several reference panels used for the imputation including 1000 Genomes[48], HapMap consortium[49] and Haplotype Reference Consortium (HRC)[50], the latter being the panel used for imputation of genotypes for participants in this study. The imputation quality measure (0-1) is derived by each software and it is later used to filter out the low imputation-quality variants [43,51].

2.5.5 SNP-phenotype association testing

GWAS is a hypothesis-free approach and is mainly used to generate knowledge that will enable hypothesis generation. It represents a series of tests that test each of the SNPs independently for the association with the phenotype of the interest. In the case of quantitative traits, linear regression is used for testing usually under the additive genetic model assumption for the SNPs where genotype is converted to the number of the reference alleles (0, 1 or 2)[52]. Additionally, statistical tests are adjusted for the factors known to influence the phenotype to reduce bias in the results. Spurious associations can also result from population stratification due to the presence of population substructure. Prior to the statistical testing, the presence of population stratification is examined and corrected for in the subsequent analysis[53,54].

2.5.6 Meta-analysis of genome-wide association studies

GWAS studies conducted in separate study groups can be meta-analysed to increase the sample size and improve power to detect significant associations. Meta-analysis allows the pooling of multiple studies without requiring the transfer of genotype data and protected clinical information of the study participants[55]. Importantly, the participating studies need to be independent, similar in their design and follow the same procedure with consistent SNPs, covariate adjustments, phenotype measurement and phenotype definition. Quality control of the individual-study summary statistics is performed to harmonize the participating study data and ensure that the results provided by each cohort are based on the same genomic build and reference allele to avoid false results and nullifying them in case of opposite allele reporting[56]. The measure of heterogeneity is derived to quantify the degree to which the studies differ[57]. I^2 index is one of the commonly derived coefficients representing the proportion of variability in coefficient resulting from meta-analysis which is attributed to the heterogeneity[58]. Once the quality control is done, the meta-analysis can be conducted using different approaches with fixed effects meta-analysis being the most widely used. Fixed-effects meta-analysis is performed under the assumption of the same magnitude of the risk allele effect across all studies[59]. A commonly used model for fixed-effects meta-analysis is inverse variance weighting where the inverse of the squared standard error is used as the weight for each study[60]. Afterwards, the significant results are usually defined as the associations with P-value < 5×10^{-8} which corresponds to Bonferroni correction of the 5% type I error rate for one million independent comparisons for common variants in the human genome[61].

2.5.7 Replication analysis

Following the discovery GWA meta-analysis, a replication study is conducted in an independent sample which is drawn from the same population with aim of confirming the GWAS results[62]. Replication study needs to be well powered to detect the same SNP-phenotype association and will largely depend on the phenotype definition and overall study design which should be similar to the design of the discovery study. Replication is considered successful if the similar effect for GWAS-identified SNPs or SNPs which are in high LD with GWAS-identified SNP is detected[63]. LD calculation is done in a reference population sample of the same descent as GWAS cohorts with the majority of the studies using 1000 Genomes project data[48] or UK Biobank reference sample[64].

2.5.8 Post-GWAS analysis

In GWAS, both direct and indirect associations are possible. Direct association indicates a SNP that is directly genotyped or imputed, tested in the study, and is found to be statistically significant for the phenotype of interest and directly influences it. On the other hand, indirect associations are present when the causal SNP is not genotyped but the SNP in high LD is found to be statistically associated with the phenotype of interest. For this reason, in majority of cases the associated SNPs cannot be assumed to be causal but rather be used to map the influential SNP[35].

Since SNPs in high LD with the associated SNP cannot be excluded, genomic risk regions are defined as they cover all the potentially causal SNPs (candidate SNPs) rather than just one SNP. Also, the majority of GWAS-identified SNPs are commonly located in non-coding parts of the human genome with an equal proportion of intergenic and intronic regions[65].

In some cases, the genomic region can cover dozens of genes and further prioritization has to be performed. The efforts to identify the causal gene in the region include determining the effect of the candidate SNPs on gene expression and potential deleterious effect on the protein's structure and function.

To investigate the potential functional effects of the candidate variants, the existing databases and algorithms can be used, such as Variant Effect Predictor (VEP) by Ensembl which uses SIFT[66] and Polyphen2[67] algorithms. The effect of the SNP on gene expression can be determined by colocalization analysis of the GWAS and gene expression summary statistics. A similar regional association pattern is considered as positive colocalization with a high probability of SNP having a pleiotropic effect, meaning that it affects both the phenotype of interest and expression of the gene[68]. The pleiotropy of the trait of interest and gene expression in relevant tissues provides powerful indications for setting hypotheses about the genes and pathways through which the associated variants might mediate their effects.

Determining which genes are influenced by the GWAS-identified SNPs is important for pinpointing the biological pathways which underlie the phenotype of interest[69]. Once potentially causal genes are identified, the protein-protein interaction network can be constructed, either based on the existing knowledge about their interactions from the databases[70] or the summary statistics from the conducted GWAS by exploring the SNP-SNP effects and their correlation[10].

2.6 Previous GWAS of IgG N-glycosylation

GWA studies of IgG N-glycosylation have previously been used to identify main glycosyltransferases, transcription factors, co-factors, as well as additional genes with no apparent role in IgG glycosylation[7–10].

The first-ever GWAS of IgG N-glycome was conducted by Lauc *et al.*[7] on 77 IgG N-glycome traits based on UPLC measurement, including both directly measured and derived traits, in four cohorts (n=2247) of European descent. Association testing resulted in nine genomic loci that passed the genome-wide significance level (P-value < $5x10^{-8}$), four of which contain genes encoding glycosyltransferases- enzymes catalyzing the transfer of sugar molecules to the N-glycan chain. *B4GALT1, ST6GAL1, MGAT3 and FUT8* genes encode enzymes that catalyze the addition of galactose, sialic acid, bisecting GlcNAc and core-fucose units to the N-glycan chain, respectively[71].

Five additional loci were discovered containing the following genes: *IKZF1, IL6ST-ANKRD55, SUV420H1, SMARCB1-DERL3* and *ABCF2-SMARCD3,* however, these genes currently do not have an apparent role in the IgG N-glycosylation process. The conducted study provided proof that the GWAS approach could not only identify the genes with a relevant and known role in glycosylation but also implicate the novel genomic loci containing genes with potential but still unknown function in the process. The conducted study focused mainly on pre-existing knowledge on functions of the genes in discovered loci to make inference about the most plausible candidate genes associated with IgG glycosylation. However, additional evidence for the prioritization of certain genes in discovered loci was obtained in GWA studies that followed.

The next attempt to discover additional genomic loci associated with IgG glycosylation consisted of a set of multivariate GWA studies of 23 phenotypes derived from the UPLC measurements of IgG glycans in the Orkney Complex Disease Study (ORCADES) cohort (n=1960)[8]. A multivariate approach was applied to address the correlation structure among omics traits that might be partially genetically regulated and, therefore, cannot be ignored[72]. In multivariate analysis, the association of multiple phenotypes with a genetic variant is tested, providing more power to detect new associations as opposed to univariate approach[72]. Directly measured glycan traits were grouped into nine groups defined by the structural and chemical glycan properties (e.g. fucosylation, galactosylation, bisection), as well as one group which contained all 23 IgG N-glycans denoted simply as N-glycosylation. In this study, five

previously identified IgG N-glycosylation loci[7] were replicated and five novel loci were detected harbouring *IGH, ELL2*, *FUT6-FUT3, HLA-B-C,+* and *AZI1* genes. Four of the newly discovered loci were not detectable by univariate analysis, thereby confirming the need for multivariate approaches in the genetic analysis for complex traits.

GWAS by Wahl *et al.*[9] was conducted using the IgG N-glycome measurements obtained by LC-MS in the KORA F4 cohort (n=1836). As LC-MS glycoprofiling provides glycan measurements specific for IgG subtypes (IgG1, IgG2/3 and IgG4), the study found the difference in bisection and fucosylation regulation between N-glycomes in different IgG subclasses. Besides that, the study replicated six of the known and discovered one novel association on chromosome 1 containing *RUNX3* gene.

The findings from the performed GWA studies implicate the polygenic nature of the genetic regulation of IgG N-glycosylation. Considering that, the aim of the subsequent GWAS study by Klarić *et al.*[10] was to increase the number of samples and improve power to detect additional genomic loci which affect IgG N-glycans. The tested phenotypes included 23 UPLC-measured glycan traits and additional 54 derived traits. Meta-analysis of summary statistics from four cohorts of European descent was performed with a total sample size of 8090. Thirteen previously known loci were replicated and fourteen novel loci were identified. Moreover, the study helped refine the list of IgG glycosylation-related genes by applying several gene prioritization strategies including exploration of variant effects, pleiotropy with gene expression, and obtaining evidence of a gene being functionally similar to genes from the gene-set representing relevant biological pathways[73]. The genes prioritized in each replicated genomic region are shown in Figure 2. The gene-set enrichment analysis showed that the genes were overrepresented in sets associated with processes such as glycosylation, immune system and transcription. Based on the constructed functional network, the study explored the potential regulation of *FUT8* by *IKZF1*, a gene that encodes a transcription factor involved in the regulation of gene expression in B cells. They show that IKZF1 regulates the gene expression of *FUT8* by binding to the regulatory regions of *FUT8* gene, while the *IKZF1* knockdown increases the levels of fucosylation through an increase in *FUT8* expression.

Additionally, the pleiotropic effects of the identified variants on the IgG glycosylation and inflammatory diseases were observed including CD, UC, RA, PBC, cholesterol, asthma, Parkinson's disease and HDL.

The latest GWAS of IgG N-glycome[74] was also conducted on the same set of samples as the previous study[10] but using the multivariate approach as described in the study by Shen et al[8]. The study replicated 26 of the associations from Klarić *et al*. study while one locus (rs12341905 near *SPINK4*) remained unreplicated. With additional three novel associations (*TNFRSF13B, OVOL1/AP5B1***,** *RNF168***),** the study increased the number of genomic loci associated with IgG N-glycome to 29.

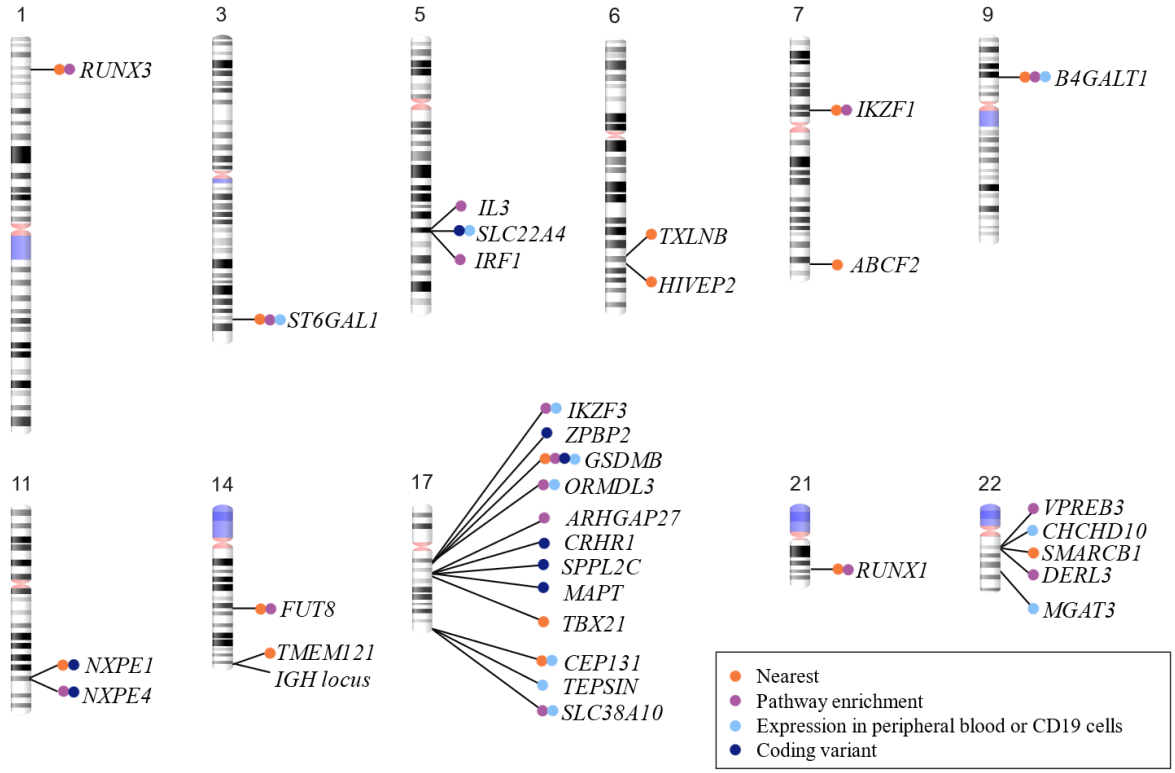


Figure 2:Chromosomal location of associations with IgG glycosylation across the human genome from Klarić *et al*. study; only loci passing the significance threshold in replication analysis are shown. The gene prioritization was based on the evidence denoted by coloured dots: nearest to the strongest association in the region (orange), biological pathway enrichment (purple), pleiotropy with gene expression in peripheral blood or CD19 B-cells (light blue) and missense mutation (dark blue). Image adapted from Pezer *et al.*[75]

3. METHODS

3.1 Studied cohorts

Recruitment of participants, sample collection, genotyping and phenotyping in the cohorts used in the study was performed by staff members at the King's College London, United Kingdom, German Institute of Human Nutrition, Germany, University of Zagreb, Croatia, University of Split Medical School, Croatia, University of Edinburgh, United Kingdom, Leiden University Medical Centre. Netherlands, Helmholtz Zentrum München – German Research Center for Environmental Health, Germany, Institute of Genetic Epidemiology, Freiburg University Medical Center, Germany and University of Tartu, Estonia.

IgG N-glycan quantification was performed by Genos Glycoscience Research Laboratory, Croatia and Leiden University Medical Centre, Netherlands.

TwinsUK

TwinsUK is a national registry of 12,000 volunteer twins in the United Kingdom. The cohort mostly consists of female subjects (83%) with an almost equal number of monozygotic (51%) and dizygotic (49%) twin pairs. With the primary goal to study the genetic background of healthy ageing and complex diseases, a subset of 7,000 twins was assessed for a wide range of clinical, biochemical, socioeconomic and behavioural characteristics. Furthermore, several omics' datasets are available including genome-wide SNP data which is being used in genome-wide association studies. The participants signed informed consent forms and ethical approval was obtained for academic and commercial use of the study[76]. A subset consisting of 4477 twins was used in this genome-wide association study of IgG N-glycome.

The European Prospective Investigation into Cancer and Nutrition Study

The European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam is a prospective cohort study that includes 27,548 participants who were recruited from the general population of Potsdam and surrounding area in the period between 1994 and 1998[77]. The age of participants at recruitment ranges between 35 and 65, and the number of female and male subjects is 16,644 and 10,904, respectively. The initial assessment consisted of anthropometric measurements and blood sample collection used for omics' data derivation. Questionnaires and face-to-face interviews were used for assessing the sociodemographic characteristics, lifestyle and current health status. Follow-up assessments were carried out by telephone and via questionnaires which were sent out every 2-3 years[78]. Ethical approval was obtained by the

ethics committee in Germany and all participants gave informed consent[77]. Based on the availability of both genetic and glycomic data, a subset of 2,406 subjects was used in this study.

CROATIA Vis, CROATIA-Split and CROATIA-Korcula

"10001 Dalmatians" is a study of Croatian island isolates which includes participants from six Adriatic islands (Korčula, Vis, Lastovo, Mljet, Susak, Rab) and the city of Split. The study aims to investigate genetic and environmental determinants in health and disease by using the advantage of genetically isolated populations. In this study, CROATIA-Vis, CROATIA-Korcula and CROATIA-Split sample groups were used. A total of 1008 participants aged 19-93 was recruited for CROATIA-Vis cohort in villages of Vis and Komiža during 2003 and 2004. Besides completing health, dietary and health questionnaire, participants were assessed for the number of anthropometric and physiological measurements, and they also donated overnight fasting blood samples which were used for DNA analysis, biochemical measurements and molecular marker assessment including glycomics, which is used in this study[79]. Data on CROATIA-Korcula subjects was collected from the island of Korčula, specifically from the town of Korčula and three villages including Lumbarda, Zrnovo and Račišće. The participants were aged 18-98 at the time of recruitment and the data was collected in the same way as for CROATIA-Vis cohort. CROATIA-Split cohort comprises 1,012 subjects aged 18-85 who were recruited in 2009-2010 in the city of Split. The data collection was carried out following the same protocol as in the other CROATIA cohorts. Ethical approval was obtained for each cohort from ethics committees in Croatia and Scotland. All participants provided signed informed consent.

The Orkney Complex Disease Study

The Orkney Complex Disease Study (ORCADES) is a family-based cohort collected with aim of identifying genetic risk factors for complex diseases in an isolated population of the Orkney Island in northern Scotland. The recruitment was initiated in 2005 and it lasted for 6 years during which data on 2080 subjects were collected. The subjects were recruited if they had at least two Orcadian grandparents. The initial visit included cardiovascular measurements and fasting blood sample collection, followed by additional visits for assessment of cognitive function, eye measurements and DEXA scans. The study was approved by ethics committees in Scotland and signed informed consent was obtained from all participants [80]. A subset of 1,720 subjects was used in this study.

Leiden Longevity Study

Leiden Longevity Study (LLS) is a family-based cohort from the Dutch population designed to study longevity. Nonagenarian siblings, individuals having a sibling older than 89 years for men and 91 years for women, and their offspring and offspring's spouses were recruited for the study if they were of European descent. Initial data collection started in 2002 and ended in 2006 during which blood samples were obtained for assessment of plasma parameters and DNA and RNA extraction. A total of 3,359 subjects was included: 944 long-lived proband siblings, 1,671 offspring and 744 controls (offspring spouses). The study was approved by the Ethical Committee of Leiden University Medical Centre. Written informed consent was obtained from all participants[81]. A subset of 1,190 participants including only offspring and their spouses was used in this study.

The Cooperative Health Research in The Augsburg Region F4

The Cooperative Health Research in The Augsburg Region (KORA) F4 is a population-based study conducted in 2006-2008 as a follow-up of the KORA S4 study which was conducted during 1999-2001[82]. Participants were individuals randomly selected from the population registry in the Augsburg region and two neighbouring counties. The data collection included standard medical and physical examinations. A total of 3,080 participants (1,594 females and 1,486 males) aged 32-86 years were included in the F4 follow-up[83] of whom 1,167 were used IgG N-glycome GWAS. Ethical approval was obtained from the Ethics committee of Bavarian Chamber of Physicians, Germany. All participants gave informed consent prior to entering the study.

The Viking Health Study - Shetland

The Viking Health Study - Shetland (VIKING) is an epidemiologic study aiming to discover the genetic basis for factors influencing risk for complex diseases including cardiovascular, chronic kidney and lung diseases, as well as glaucoma, diabetes and stroke. VIKING cohort consists of individuals from an isolated population of Shetland in the north of Scotland and the main criteria for participation was having at least two grandparents from Shetland. A group of 2,105 participants was recruited between 2013 and 2015. A large number of distant relatives makes the cohort suitable for the identification of rare genetic variants which influence disease risk[84]. Data on health-related phenotypes and environmental parameters was collected and participants donated a fasting blood sample. A total of 1,082 subjects was selected for IgG glycoprofiling and a subset of 1,071 was used in this study.

The Estonian Genome Center of the University of Tartu Biobank

The Estonian Genome Center of the University of Tartu (EGCUT) Biobank is a volunteer-based cohort of 52,000 adult subjects from the Estonian population (aged ≥ 18). The recruitment of the subjects was performed throughout the country via general practitioners and medical personnel in the period between 2002 and 2012. The participants donated blood samples and completed a questionnaire on topics such as lifestyle, diet and clinical diagnostics. The cohort was used in studies investigating over 200 traits including anthropometric traits, blood biochemistry, common and rare diseases, as well as lifestyle and personality traits. The data is being continuously updated through follow-up health checks using national electronic health registries and re-examinations[85]. A total of 1,108 subjects has data on IgG N-glycome but a subset of 483 was used in this study.

German Chronic Kidney Disease Study

German Chronic Kidney Disease (GCKD) study is an ongoing prospective observational study of kidney disease patients who are under the regular care of nephrologists in Germany[86]. The current sample size of 5,217 makes it the largest CKD cohort worldwide. The mean age of participants is 60 with 60% of the participants being male. The enrolment took place between March 2010 and March 2012 and was conducted by certified study members associated with the nephrologist practice or by outpatient units throughout different regions. Besides collecting information on sociodemographic factors, medical and family history, the participants donated blood samples which were later processed and shipped to the central laboratory for measurement of the core clinical parameters and samples were stored for any future analyses. IgG N-glycans for the cohort were measured in around 5,000 samples using UPLC and 4,933 were used in the current genome-wide association study.

3.2 Genetic analysis

Genotyping and Quality control

Genotyping in the cohorts used in the study was performed by staff members at the King's College London, United Kingdom, German Institute of Human Nutrition, Germany, University of Zagreb, Croatia, University of Split Medical School, Croatia, University of Edinburgh, United Kingdom, Leiden University Medical Centre, Netherlands, Helmholtz Zentrum München-German Research Center for Environmental Health, Germany and University of Tartu, Estonia.

Genotyping was performed using commercially available SNP genotyping arrays (listed in Table 1), followed by genotype calling in Illumina and Genome Browser software**.** Quality control (QC) was performed to exclude SNPs and samples with low genotyping quality. Quality control of SNPs removed 1) SNPs with low call rate, 2) SNPs violating the assumptions of Hardy-Weinberg Equilibrium (HWE) and 3) SNPs with low minor allele frequency (MAF) < 1%. Depending on the cohort, samples with call rates < 95%, < 97% or < 98% were removed.

Genotype imputation

Imputation of SNPs was performed to increase the number of SNPs tested in the association analysis, as well as increase the number of overlapping SNPs between cohorts due to different genotyping arrays. The overview of genotyping arrays, genotype quality control and imputation software is shown in Table 1. HRC[87] panel was used as a reference for imputation. All genotypes were mapped to Genome Reference Consortium GRCh37 (hg19).

Table 1: Overview of genotyping arrays and imputation

Cohort	**Genotyping platform(s)**	**ID call rate**	**SNP call rate**	**HWE p**	**MAF**	**N SNPs postQC**	Imputation Tool
TwinsUK	Illumina HumanHap300; Illumina HumanHap610Q	>95%	> 97% (MAF ≥ 5%); > 99% (1% ≤ MAF < 5%)	10^{-6}	≥1%	NA	MACH (Michigan Imputation Server v1.0.2)
EPIC-Potsdam	Human660W-Quad_v1_A[88]	>97% >99%	>95%	10^{-3}	NA	NA	Eagle2/minimac 3
	HumanCoreExome-12v1-0_B[88]	>98%	>95%	NA	NA	NA	Eagle2/minimac 3
	Illumina InfiniumOmniExpressExome-8v1-3_A DNA Analysis BeadChip[89]	>98%	>95%	NA	zCall threshold=7	NA	Eagle2/minimac 3
LLS	Illumina660 W; Illumina OmniExpress	>95%	>95%	10^{-4}	≥1%	296,619	IMPUTE2
CROATIA-Korcula 1,2,3	Illumina HumanHap s370CNV DUO/QUAD Phase 1(1); Illumina HumanOmniExpress Exome (2 & 3)	>97%	>98%	10^{-6}	≥1%	305,354 606,438	SHAPEIT2/Sanger
CROATIA-Split	Illumina HumanHap 370CNV QUAD Phase I; Illumina HumanOmniExpress Exome	≥97%	>98%	10^{-6}	≥1%	321,456	SHAPEIT2/Sanger
CROATIA-Vis	Illumina HumanHap300v1 BeadChip	>97%	>98%	10^{-6}	≥1%	272,930	SHAPEIT2/Sanger
VIKING	Illumina HumanOmniExpress Exome	>97%	>98%	10^{-6}	≥1% omni markers; ≥0.01%	611,836	SHAPEIT2/Sanger

					exome markers		
EGCUT	Illumina GSAv1.0, GSAv2.0, GSAv2.0_EST	>95%	>95%	10^{-4}	≥1%	NA	Beagle v.28Sep18.793
KORA F4	Affymetrix Axiom	>97%	>98%	$5x10^{-6}$	≥1%	508,532	SHAPEIT/IMPUTE
ORCADES	HumanHap300v2 Phase 1	>97%	>98%	10^{-6}	≥1%	278,618	SHAPEIT2/Sanger
GCKD	Illumina Omni2.5Exome BeadChip	>97%	>96%	10^{-5}	>1%	2,337,794	Eagle/minimac3

3.3 IgG N-glycome analysis

IgG N-glycome measurements were obtained using UPLC and LC-MS. The analysis of IgG N-glycome was performed in Genos Glycoscience Research Laboratory, Croatia for all cohorts except for LLS cohort for which the glycoprofiling was performed by the Centre for Proteomics and Metabolomics at the Leiden University Medical Centre, Netherlands. An overview of cohorts and corresponding platform used for IgG N-glycome measurement, as well as literature where the analysis was described in more detail, is shown in Table 2.

Table 2: Overview of platforms used for quantification of IgG N-glycans

Cohort	Platform	Reference
TwinsUK	UPLC	Menni *et al.* [90]
EPIC	UPLC	Not published
LLS	LC-MS	Wahl *et al.*[9]
CROATIA-Korcula	UPLC and LC-MS	Pučić *et al.*[5]
CROATIA-Split	LC-MS	Not published
CROATIA-Vis	LC-MS	Pučić *et al.*[5]
VIKING	UPLC	Landini *et al.*[91]
EGCUT	UPLC	Trbojević-Akmačić *et al.*[92]
KORA F4	LC-MS	Wahl *et al.*[9]
ORCADES	UPLC	Krištić *et al.* [93]
GCKD	UPLC	Not published

Glycan Quantification by Ultra-Performance Liquid Chromatography

Ultra-performance liquid chromatography is used for quantification of glycan structures attached to both Fc and Fab portions of IgG without the possibility to differentiate them. The detailed protocol for UPLC analysis is published elsewhere[5]. Briefly, IgG was isolated from blood plasma samples using Protein G plates (BIA Separations, Ajdovščina, Slovenia). After filtration, plates were extensively washed to remove unwanted proteins and IgG was released

from protein G monoliths using 0.1 M formic acid. Eluates were collected in a 96-well plate and neutralized with neutralization buffer (1 M ammonium bicarbonate) to pH 7.0 to maintain the stability of IgG. IgG samples were dried and denatured using SDS detergent and incubated at 65°C for 10 minutes. N-glycans from IgG were released using recombinant N-glycosidase F (ProZyme, Hayward, CA), followed by fluorescent labelling with 2-aminobenzamide dye. Hydrophilic interaction liquid chromatography (HILIC) based solid-phase extraction (SPE) was used to remove excess protein, reagents and fluorescent label, followed by clean-up with acetonitrile (ACN). Fluorescently labelled N-glycans were separated hydrophilic interaction UPLC on Waters Aquity UPLC H-class instrument (Waters, Milford, MA) with Waters bridged ethylene hybrid (BEH) glycan chromatography column. A linear gradient of 75 to 62% ACN in a 20-min analytical run was used to separate different glycan structures. The retention times for individual glycans were converted to glucose units based on hydrolysed and 2-AB labelled glucose oligomers which were used as external standards for calibration of the system.

Data processing was done in two ways depending on the cohort, 1) using Empower 3 software with an automated processing method with traditional integration algorithm, followed by manual correction of each chromatogram to maintain the same integration intervals in all samples or 2) automatic integration as described in Agakova *et al.*[94]. The resulting chromatograms were separated into 24 peaks where the amount of glycans was expressed as % of the total integrated area in the corresponding peak (GP1-GP24). Total separation of each glycan structure is not possible using the described method, thus resulting in multiple glycan structures being found under a peak. Ten of the 24 peaks contain more than one structure (shown in Figure 3). The output of UPLC measurement is represented by 24 values which are referred to as directly measured glycan traits. Glycan structures in each peak and percentages of the area in each peak are listed in Supplementary Table 1.

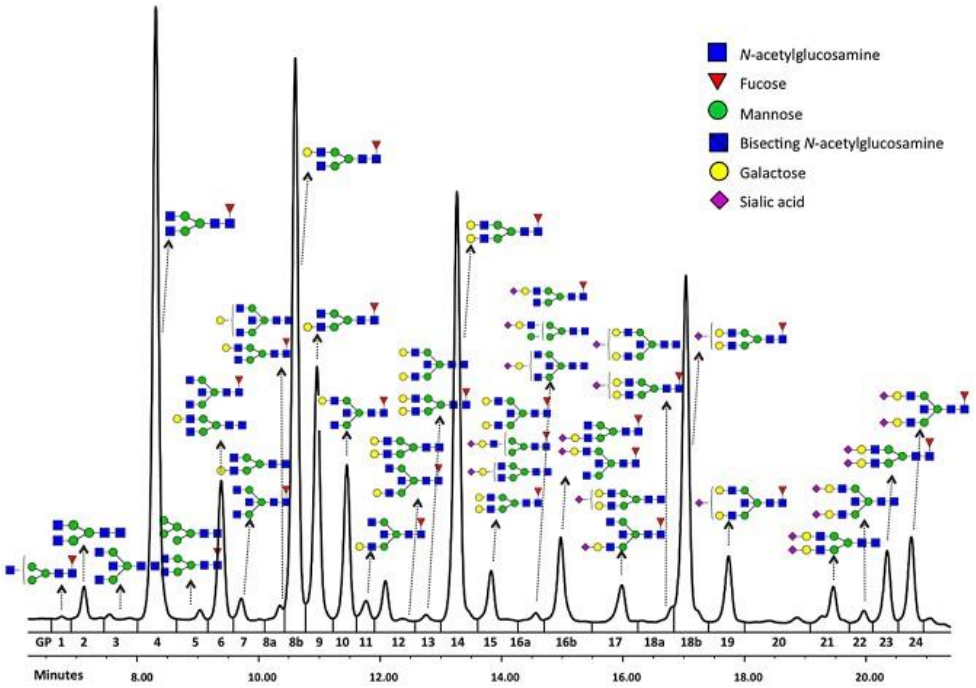


Figure 3: Chromatogram of UPLC N-glycan measurement of the human IgG. Image from Lauc *et al.*[7]

Glycan Quantification by Liquid Chromatography coupled with Mass Spectrometry

Liquid chromatography coupled with mass spectrometry enables quantification of glycan structures attached to the Fc portion of each of the IgG subclasses (IgG1, IgG2/3, IgG4). The full name of the method is reverse-phase nano-liquid-chromatography-sheath-flow-electrospray-mass spectrometry (LC-ESI-MS) but in this work, we refer to it as LC-MS. The detailed protocol for analysis of IgG N-glycans using LC-MS is described in Selman *et al.*[95]. Briefly, IgG was isolated by affinity chromatography binding to protein G 96-well plates (BIA Separation, Ajdovščina, Slovenia) and treated with trypsin overnight at 37ºC which allowed cleavage of IgG at specific amino acid sites. The cleavage by trypsin resulted in different glycopeptides due to the difference of amino acid sequence in different IgG subclasses, thereby enabling subclass-specific glycan measurements.

IgG subclass separation was performed using the Ultimate 3000 HPLC system (Dionex Corporation, Sunnyvale, CA). The SPE trap column was conditioned with mobile phase A and samples were loaded and separated on Ascentis Express C18 nano-LC column (Supelco, Bellefonte, USA) conditioned with mobile phase A and 95% ACN. For detection of separated subclass-specific glycopeptides, the HPLC system was coupled to a Dionex Ultimate UV detector and interfaced to a quadrupole-TOF-MS mass spectrometer (Bruker Daltonics, Bremen, Germany) with a standard ESI source (Bruker Daltonics, Bremen, Germany) and a

sheath-flow ESI sprayer (Agilent Technologies, Santa Clara, USA). The mass spectra were recorded in a range between 300 and 2000 m/z with two averages at a frequency of 1Hz. The analysis time for one sample was 16 minutes.

The calibration of LCMS datasets was done internally using a list of known glycopeptides and datasets were exported to the open mzXML format by Bruker DataAnalysis 4.0 software, followed by alignment to a master dataset of a typical sample. In-house software "Xtractor2D" was used to extract pre-defined features such as peak maximum or peak area in specific retention time and mass windows. Relative intensities of subclass-specific glycopeptides were obtained by integrating and summing three isotopic peaks. The obtained intensities were then normalized to the total IgG subclass-specific glycopeptide intensities. IgG2 and IgG3 subclasses have the same tryptic glycopeptide moieties, thus not enabling the separation of the subclass-specific glycopeptides. Here, obtained measurements are simply referred to as IgG2/3. LC-MS quantification results in 50 values which refer to 20 glycans measured on IgG1, 20 glycans on IgG2/3 and 10 glycans on IgG4. All glycans measured on IgG4 are fucosylated structures since the nonfucosylated glycans are hard to distinguish from the glycans found on IgG1[96]. The list of glycans measured by LC-MS and their description is listed in Supplementary Table 3.

3.4 Data harmonization

Previously, there were no GWA meta-analyses of the human IgG N-glycome using GWAS of both UPLC and LC-MS IgG N-glycome measurements, therefore making it necessary to first test how the two types of data correlate and what methods should be applied in pre-processing to make them comparable. For this, we used the CROATIA-Vis cohort (n=661) for which both UPLC and LC-MS glycan data were available for the same samples.

One important factor for comparing UPLC and LC-MS data is the definition of the traits which will ultimately be compared. Original glycan traits measured by the platforms are not directly comparable as they are based on different technologies and differ in the information they output. LC-MS output shows the abundance of glycan structures per each IgG subclass: IgG1, IgG2/3 and IgG4. UPLC gives output based on the total glycome regardless of the IgG subclass. We aimed to combine IgG subclass information from LC-MS in an appropriate manner to get information corresponding to whole IgG glycome values measured by UPLC.

Pre-processing of IgG glycome data consists of normalization of the data and batch correction to remove the effects of experimental variation. The term normalization describes the division of each data row by a normalization factor. The normalization procedure allows for the comparison of the samples by removing the unwanted variation between the samples[97]. Several types of normalization methods can be applied to glycan data and the following three were tested: total area normalization, largest peak normalization and median quotient normalization.

Total area normalization represents the normalization of each peak by the total area of the chromatogram. The normalization factor for each sample is calculated by summing all the features in the corresponding sample row. The total area normalization is applied under the assumption that the total concentration of the analyte remains unchanged across all samples. Analytes present in high concentrations will contribute to the normalization factor more than the low-concentration analytes. In case when the peak intensity of analyte with high concentration significantly changes, the normalization factor will be affected. Total normalization scales each sample in such a way that the sum of each row (sum of all glycans) is equal to 1, resulting in so-called compositional data[98]. Total area normalization is applied by using tanorm() function in "glycanr"[99] package in R.

The largest peak normalization represents the normalization of each peak with the highest peak of the chromatogram. The normalization factor for a sample is calculated by choosing the highest value in the corresponding sample row. The largest peak normalization can be applied by using refpeaknorm() function in "glycanr" package in R.

Median quotient normalization assumes that the signal intensity is a function of dilution only and uses the median as an estimator of the most probable quotient, a quantity that is used as a normalization factor. The normalization factor is computed as the most probable quotient between the corresponding spectrum and reference spectrum and it substitutes the total integral as a marker of sample concentration. The reference spectrum is calculated as the median spectrum based on all spectra in the study[100]. Median quotient normalization is applied by using medianquotientnorm() function in "glycanr" package in R.

Given that LC-MS measurement gives information per IgG subclass, we applied normalization both across the whole glycome and per IgG subclass.

Batch correction

Due to varying laboratory conditions during the experiment, it is necessary to perform batch correction to remove variation introduced by them. All procedure steps are performed using ComBat function in R package "sva"[101]. ComBat function implements empirical Bayes method for batch correction[102] which assumes a normal distribution of the data. The glycan data is mostly left-skewed making it necessary to first log-transform the data. Another reason for log transformation is the multiplicative nature of batch effects which will become normally distributed after the transformation and the model will be additive. The steps in batch correction are as follows: log transformation, batch correction with ComBat() and exponential transformation of the values to the original scale.

Derived trait calculation

We calculated derived traits from the initial traits to enable a more straightforward interpretation of the GWAS results so that the discovered genomic loci can be directly linked to the addition of one of the four sugar units which are found in IgG N-glycome: galactose, fucose, sialic acid and bisecting GlcNAc. The derived traits were calculated as the percentage of all structures from the measured glycans which contain a certain sugar unit(s) and the additional three traits representing a ratio of structures. The list of traits and formulas used to calculate them from UPLC and LC-MS data are listed in Supplementary Table 4.

Response factor

Given the nature of the LC-MS experimental procedure, we also incorporated an approximation of the IgG glycan subclass response factor to represent the true IgG subclass concentration relative to other subclasses. Previous experiments allowed for approximation of subclass response factors (RF): IgG1 with RF of 1, IgG2/IgG3 with RF of 2 and IgG4 with RF of 1 (not published). We incorporated the response factor by multiplying raw data values of the IgG subclass with the corresponding response factor. Response factors for each IgG subclass and relative concentrations of IgG subclasses were considered in the normalization, as well as in the calculation of the trait.

Relative concentrations of IgG subclasses

IgG subclasses are present in different abundances in human serum, so we attempted to maximize the correlation of LC-MS with UPLC data by incorporating relative concentrations of each subclass in the calculation of derived traits. We used the following relative measurements that have been previously indicated in the literature: 66% for IgG1, 30% for

IgG2/IgG3 and 4% for IgG4 [17]. The subclass-specific glycan measurements were weighted by the corresponding concentration prior to trait calculation.

3.5 Statistical analysis

3.5.1 Pre-processing of glycan data

Prior to genetic analysis, glycan measurements were normalized and batch corrected to reduce the impact of experimental variation on the downstream analysis. In the previous section, data harmonization was described where different normalization methods for glycan data were tested. The glycan data was pre-processed centrally in Genos in all cohorts except the LLS cohort for which the glycan data was pre-processed by a colleague from Leiden University Medical Centre. Also, glycan data for the CROATIA-Korcula cohort was obtained in three instances (2010, 2013 and 2017) and each dataset was separately pre-processed and treated as an individual cohort in downstream genetic analysis. TwinsUK cohort was analysed in four separate batches. Due to differences in methodology of sample collection, batches 1 and 2 were considered as one dataset and batches 3 and 4 were considered as the second dataset. The initial (raw) measurements of glycans represent the area under the peaks in the chromatogram. Extreme values in data were removed and considered as technical outliers if the values were in the 99.9th percentile. Next, based on the results of the previous test, for harmonization of the data, median quotient normalization was applied on both UPLC and LC-MS glycan data across 24 and 50 glycan values, respectively. Log transformation was applied to reduce the skewness of the data followed by batch correcting using empirical Bayes method[102] implemented in ComBat function in "sva" [103] package in statistical software R[104]. Each batch was represented as a 96-well plate on which samples were analysed. After batch correction, the values were exponentiated to the original scale followed by the calculation of the derived traits. Derived traits represent values to describe the group of glycans that contain certain sugar unit in their structure. For example, monosialylation represents the percentage of glycan structures in the total glycome which contain one sialic acid and is calculated by summing glycans which contain one sialic acid unit and dividing by the sum of all glycans. A total of eleven matching traits was defined in UPLC and LCMS to enable an integrative analysis of different datasets. Derived traits and formulas for their calculation are listed in Supplementary Table 4. Prior to the genetic association test, glycan data in all cohorts was transformed using rank-based inverse normal transformation (mean=0, standard deviation=1) which is commonly used in genetic studies, thereby ensuring that the phenotype values in all cohorts are on the same scale.

3.5.2 Genome-wide association study

To perform GWAS, an analyst for each cohort received the pre-processed glycan data and analysis plan explaining GWAS methodology. The list of analysts for each cohort is available in Supplementary Table 6 and a detailed analysis plan can be found in Supplementary Material. Different analysts used different software for association tests, but the same association method was applied for all cohorts. The list of cohorts and number of analysed samples is listed in Table 3. GWAS was performed on HRC imputed genotypes assuming an additive linear model of association. Derived trait values were adjusted for age, sex and cohort-specific covariates before fitting linear mixed models which also consider genomic kinship while testing the association between SNPs and phenotype. GWAS pipeline developed by the Wilson group from the University of Edinburgh was used to perform GWAS in ORCADES, CROATIA-Korcula (three datasets; one UPLC and two LC-MS datasets), CROATIA-Vis, CROATIA-Split and VIKING cohorts because the genetic data was directly available on the University of Edinburgh's High Performance Computing Server. The GWAS pipeline implements linear mixed modelling in three steps, 1) fits covariates (except genetic kinship) in a fixed effects linear model, 2) passes the residuals from the fixed effects linear model and the genetic kinship matrix to polygenic() from GenABEL package[105] in R and fits the linear mixed model, with kinship fitted as a random effect and 3) passes the residuals generated by polygenic() to REGSCAN[106] for the genome-wide association. Genome-wide association test in TwinsUK cohort (two datasets) was performed accounting for batch effects due to four different instances over the years in which glycan measurements in samples were obtained. Also, important to note is that GWAS for the EPIC cohort was run in four parts due to four available sample subgroups. The GWAS summary statistics files for each cohort were stored in text format containing SNP identifier, chromosome, base position, assessed allele, other allele, beta, standard error, number of samples analysed and imputation quality information. The results from collaborators were transferred via a secured site on the Lobsang server, University of Zagreb, where the subsequent analyses were performed.

Table 3: Cohorts included in the IgG N-glycome GWAS.

Cohort	N (glycans available)	N female	N male	median age	N in GWAS
TwinsUK	4624	4282	342	54	4477
EPIC	3601	NA	NA	NA	2406
CROATIA-Korcula	2478	1148	1330	57	2436
CROATIA-Split	973	383	590	52	920
CROATIA-Vis	683	394	289	57	675
VIKING	1080	644	436	52	1071
ORCADES	1786	1082	704	54	1720
LLS	1841	974	867	59	1190
KORA F4	1823	935	888	62	1167
EGCUT	1108	516	592	69	483
GCKD	4933	~1970	~2960	63	4933

3.5.3 Quality control of genome-wide association study

To increase the statistical power of genome-wide association studies, the summary statistics for multiple cohorts are pooled in a meta-analysis[107]. To ensure the maximum increase in power and avoid false positives, the summary statistics from individual cohorts need to be checked and variants with low quality need to be removed. The centrally performed pre-processing of phenotypes (except for the LLS cohort) was one of the ways to avoid discordance in data preparation which might lead to loss of power in meta-analysis. The analyst for the LLS cohort received a detailed plan for data preparation, thereby ensuring that the same phenotype preparation protocol was applied for all cohorts. Furthermore, a detailed analysis plan for GWAS was provided to all analysts in partner institutions who participated in the study. The analysis plan was based on the previous GWAS protocol used in Klarić *et al.*[10]. In this way, there was reassurance that the same phenotype transformation, statistical modelling, covariate adjustment and reference panel for imputation were applied for all individual studies and results were received in the same format which facilitated the follow-up meta-analysis. However, additional quality control (QC) steps were necessary before pooling the results from individual studies.

File-level QC

QC of GWAS summary statistics for individual cohorts (file-level QC) was done using EasyQC package in R as described in Winkler *et al.*[108]. The first part of the protocol checks for data format inconsistencies and allows the definition of uniform column names in all files. In cases where columns were missing, the study analyst was contacted and data was provided again.

Next, monomorphic SNPs (allele frequency =1 or =0), SNPs with missing alleles, P-value, effect estimate, standard error, allele frequency or sample size were excluded. SNPs with nonsense values were removed, such as SNPs with alleles other than 'A', 'C', 'G' or 'T', P-values < 0 or > 1, negative or infinite standard errors and infinite effect estimates or allele frequencies < 0 or >1. Additionally, SNPs with a low sample size (N < 30), low minor allele count (MAC ≤ 6), and low imputation quality were removed. Harmonization of allele coding and variant names was done. Allele frequencies from the HRC reference dataset were plotted against the allele frequencies provided by study partners to check for outlying SNPs and mismatches. In cases of deviation from the identity line, the study analyst was contacted and the source of the problem was clarified.

Meta-level QC

As suggested by Winkler *et al.*[108], meta-level QC was performed to check for any analytical issues in the genome-wide association scan in the individual cohorts. First, the SE/N plot was generated to check for issues with trait transformation, sample size or file-naming. The median standard error (SE) and maximum sample size (N) were used to generate a plot of c/median(SE) versus sqrt(max(N)), where c is a constant of proportionality and depends on the imputation panel. In cases of deviation from the identity line which might indicate unaccounted relatedness in the sample, the study analyst was contacted. Additional analytical issues were checked by producing the P-Z scatter plot. P-Z plot shows problems with effect estimates, standard errors and P-values, by plotting P-values calculated from a Z-statistic (Z = effect/SE(effect)) versus the P-values provided by the partners. To check for the presence of population stratification, the genomic control inflation factor (λ_{GC}) was calculated for each study and was plotted against the sample size. The plots were checked for high λ_{GC} and study files were considered for GC correction in meta-analysis.

3.5.4 Genome-wide Association Meta-analysis (Discovery)

GWAS summary statistics for seven studies (ORCADES, CROATIA-Korcula, CROATIA-Vis, TwinsUK, EPIC, CROATIA-Split, CROATIA-Vis and VIKING) were pooled using the inverse-variance weighted method in fixed-effect model implemented in METAL software[109]. The inverse variance-weighted meta-analysis approach summarizes effect sizes from multiple studies by deriving the weighted mean of the effect sizes using inverse variance to weight the effects from individual studies[110]. METAL software also allowed for the estimation of genomic

control to correct test statistics to account for relatedness or population stratification in the studies used in meta-analysis. The between-study heterogeneity was also estimated.

Genome-wide significance threshold

P-value is the measure of the probability of an observation being due to the chance or not. The widely used significance threshold for the statistical test is 0.05, meaning that 1 out of 20 tests will result in rejecting the null hypothesis when it is true. However, the threshold of 0.05 must be further corrected for the number of tests that are performed, otherwise, 1 out of every 20 tests will result in rejection of null hypothesis where it is true[111]. For GWAS, the suggested significance threshold is 5×10^{-8} which is calculated by dividing the traditional P-value threshold of 0.05 by one million tests[112]. The threshold of 5×10^{-8} is considered as conservative as less than 1 million SNPs are required to recover all tested common variants in the genome[113]. However, 5×10^{-8} is a widely accepted genome-wide significance threshold which is further corrected for the number of traits that are tested in the study. In the current study, eleven glycan traits were tested, thus requiring further correction of the significance threshold. Due to the high correlation between tested traits, principal components analysis was used to derive the number of principal components that explain 99% of the variance among eleven traits. The number of principal components was five further resulting in the p-value threshold $> 1 \times 10^{-8}$ ($5 \times 10^{-8}/5$). The analysis was done using prcomp() function in R using glycan data from ORCADES, CROATIA-Korcula, CROATIA-Vis and CROATIA-Split cohorts.

Genomic loci definition

Definition of genomic loci associated with IgG N-glycome was performed using FUMA v. 1.3.6[69]. FUMA stands for Functional Mapping and Annotation of Genome-Wide Association Studies and is an online platform that is used for annotation, prioritization and visualization of association results based on GWAS summary statistics.

SNP2GENE function in FUMA first identifies significant SNPs at the genome-wide significance level (p-value $< 1 \times 10^{-8}$) which are independent of each other ($r^2 < 0.6$). LD estimates were inferred from 1000G Phase 3 reference genome for European population[48]. Lead SNPs were selected from independent significant SNPs based on the pairwise LD ($r^2 < 0.1$). The maximum distance for merging LD blocks into a single genomic locus was 250 kb. Due to the merging of LD blocks, the defined genomic loci can contain multiple independent or lead SNPs. SNPs that are in LD with independent SNPs within 250kb distance were all selected as candidate SNPs and considered in further analysis.

Replication of discovered genomic loci

After the discovery GWAS, the validation of the results was necessary in an independent study by means of replication meta-analysis. The sample size for replication analysis was estimated based on the effect sizes and MAF for the top SNPs from the regions which should be replicated using the power of 80%. Quanto[114] software v. 1.2 was used for power calculations. Four independent cohorts of European descent were used: GCKD, KORA F4, LLS and EGCUT, with a total sample size up to 7,775. For the glycan traits, glycan measurements for KORA F4 and LLS cohorts were obtained on LC-MS, while GCKD and EGCUT glycan measurements were obtained on UPLC. The glycan-SNP pair of the strongest association in significant genomic loci were meta-analysed using the fixed-effect inverse variance method. The significance threshold was set to $p < 0.05/13 = 0.0038$, where 13 is the number of novel associations which we aimed to replicate. In cases where the top SNP association does not pass the significance threshold, we further explore the SNPs which are in high LD ($r^2 > 0.6$) with the top SNP, as well as the effect direction across cohorts.

Replication of genomic loci from previous GWAS of IgG N-glycome

The published GWAS on IgG N-glycome[7,8,9,10,74] were all conducted on a subset of participating cohorts so the replication of their finding is not considered as true replication. One of the main differences lies in the phenotypes used in the studies because previous studies were conducted either on LC-MS or UPLC IgG glycan measurements, both directly measured and derived traits that were different from the derived traits used in the current study. For instance, univariate association analysis in Lauc *et al.*[7] and Klarić *et al.*[10] studies were performed on 77 glycan values which include 23 directly measured glycan structures by UPLC and additional 54 derived traits, and Wahl *et al.*[9] conducted univariate GWAS on LC-MS-measured glycan data. On the other hand, Shen *et al.*[8] and Shadrina *et al.*[74] performed a multivariate association analysis which is based on the grouping of the directly measured glycan traits.

Additionally, meta-analysis in this study includes GWAS done on both UPLC and LC-MS-measured data. But given that all these studies aim to identify as many genomic regions involved in IgG N-glycosylation as possible, the overlap of the previously identified regions and the regions discovered here was checked by assessing the top SNP or SNP in high LD across all eleven glycan traits. The main focus was on the replication of 27 loci from Klarić *et al.* and novel loci from Shadrina *et al.* Additionally, three loci from previous studies which

were not replicated in Klarić *et al.* were assessed. One of the loci which was discovered by Shen et al. but not replicated in subsequent studies is located in the MHC region and the top SNP is rs116108880. However, this variant is not present in the tested SNPs nor does it have any SNPs in high LD.

3.5.5 Conditional analysis

The single-SNP model assumes that the maximum phenotypic variance in the region is captured by the strongest association which is in LD with an unobserved variant. Other significant associations are observed in the same region due to their correlation with the top variant. However, these assumptions may not be true for two reasons, 1) there is the possibility that a single SNP (genotyped or imputed) can capture all the phenotypic variance explained by the locus[115,116], 2) there may be more than one causal variant, hence, a single SNP may not be able to account for the LD between the unknown causal variant and tested SNPs (genotyped or imputed). Alternatively, conditional analysis can be used to identify secondary associations in the region[117,118] by performing association analysis while conditioning on the strongest association in the region. However, this approach is not feasible when individual-level genotype data is not available. Therefore, we use an approximate method called Conditional and joint analysis (COJO)[119] implemented in Genome-wide Complex Trait Analysis (GCTA) software[120] which uses GWAS summary statistics data as input and LD estimates from the reference sample to test for secondary associations. An iterative procedure is employed where association analysis is repeated while conditioning on the strongest association from the previous iteration and is continued until no significant associations are left. After, the GCTA-joint method is employed to estimate the joint effects of the independent variants. LD estimates were derived from the 10,000 randomly selected unrelated participants from the UK Biobank cohort[64] and a collinearity cut-off value of 0.9 was used. The analysis was performed by Arianna Landini from the Wilson group, University of Edinburgh. The GCTA version 1.91.4 was used.

3.5.6 Phenotypic variance explained

To calculate the variance in the eleven glycan traits explained by each independently associated SNP, the following formula was used:

$$\sigma_i = 2 * p_i * q_i * \beta_i^2 ,$$

where p_i and q_i are minor and major allele frequencies, respectively, and β_i is the effect estimate for the genetic variant in the meta-analysis for the given glycan trait. The total univariate explained variance in one glycan trait was calculated as the sum of variance explained for all independent genetic variants associated with the trait.

The total joint variance explained by all independently associated genetic variants for the glycan trait was calculated as the sum of the variance explained by each of the variants which was computed using the following formula:

$$\sigma_i^J = 2 * p_i * q_i * \beta_i^U * \beta_i^J ,$$

where β_i^U is the effect estimate for the variant from the univariate model and β_i^J is the effect estimate derived from the joint analysis.

The linkage disequilibrium score regression (LDSC) is an approach used to determine whether the observed inflation in GWAS summary statistics is due to population stratification or polygenicity[121] but it can also be employed to estimate SNP-based heritability and genetic correlation with other complex traits[122]. The LD score is obtained by summing the Pearson correlation coefficients between index SNP and surrounding SNPs. The resulting LD scores are further regressed against the chi-square statistic obtained in GWAS and the SNP-based heritability is defined as the slope of the regression line. The intercept of the regression is used to distinguish between population stratification and polygenicity as potential causes of observed inflation in summary statistics. LDSC is also used to determine the genetic correlation between two complex traits.

In this study, we use LDSC primarily to assess the SNP-based heritability for the eleven glycan traits. A subset of precomputed LD scores from 1000 Genomes EUR for HapMap3 SNPs (n=1,201,551) was used to estimate SNP-based heritability in LDSC software v. 1.0.0. integrated in LDHub[123]. SNP-based heritability is calculated as the ratio of variation in the observed additive effect of the SNPs to the total phenotypic variance. We compared the SNP-based heritability obtained by LDSC and phenotypic variance explained which was calculated using the abovementioned formula.

3.6 Previous associations of discovered glycosylation-associated loci

Phenoscanner[124] is a curated database of genotype-phenotype associations resulting from large-scale genome-wide association studies. The database facilitates “phenome scans” which enable

cross-referencing genetic associations across different phenotypes, including diseases and intermediate traits, thereby helping understand the potential biological mechanism behind a certain disease. Phenoscanner integrates NHGRI-EBI GWAS catalog[125], dbGaP[126] and NHLBI GRASP[127] catalogues of associations. In this study, we used Phenoscanner to look for shared genetic variants across IgG glycan traits and diseases and traits for which the GWAS results were available in the Phenoscanner database. Individual significant SNPs from GWAS for all eleven glycan traits as obtained by COJO and their proxies at $r^2 > 0.8$ were queried using phenoscanner() in phenoscanner v1.0 R package. Only phenoscanner GWAS association results at significance level 5×10^{-8} were retained.

3.7 Gene mapping

Genes in the significantly associated genomic loci were mapped using three approaches: positional mapping, eQTL (expression quantitative trait loci) mapping and 3D chromatin interaction mapping. Genes were positionally mapped based on ANNOVAR[128] annotations and the maximum distance between SNPs and genes (< 10 kb). The eQTL mapping was based on the eQTL datasets including B cell eQTL data from Database of Immune Cell Expression (DICE)[129] and eQTL catalogue datasets, Fairfax *et al.*[130,131] and CEDAR[132]. Only significant eQTL signals at false discovery rate (FDR) < 0.05 were used in the mapping of the SNPs to genes. Chromatin interaction mapping was performed using the Hi-C data derived from B cell line (GM12878)[133] and a suggested value of FDR < 1×10^{-6} was used[134].

3.8 Functional consequences of candidate SNPs from coding regions

The Combined Annotation-Dependent Depletion (CADD) score is the score of the deleteriousness of SNPs derived by integrating 63 functional annotations[135]. Deleteriousness refers to property that correlates with pathogenicity and molecular functionality. The CADD scores of 15 and higher scores were considered more deleterious because 15 is the median value for all non-synonymous variants and canonical splice sites in the first version of the CADD database. For the derivation of CADD scores, the integrated functionality in FUMA SNP2GENE was used. The CADD scores are derived for independent significant SNPs and SNPs that are in LD ($r^2 > 0.6$) with independent SNPs.

To further assess functional consequences of SNPs in the genomic regions significantly associated with IgG N-glycome traits, SIFT[66] and Polyphen-2[67] algorithms were used as implemented in Variant Effect Predictor (VEP) v97 by Ensembl[136]. SIFT uses the homology

of the sequence to predict if amino acid substitution will affect protein structure, hence its function and ultimately affect the phenotype of interest[137,138]. SIFT computes the probability that an amino acid at a specific position in the alignment is tolerated while conditioning on the most frequent amino acid at that position being tolerated. The substitutions with probability values of 0.05 and less are predicted to be deleterious. PolyPhen-2 predicts the damaging mutations by using three structure-based and eight sequence-based features. PolyPhen-2 computes the naive Bayes posterior probability of the given SNP being damaging and also outputs estimates of false positive and true positive rates. It also evaluates the SNP as being either benign (0.0-0.15), possibly damaging (0.15-0.85) and probably damaging (0.85-1.0). SIFT and Polyphen-2 scores were derived for independent significant SNPs and SNPs in LD ($r^2 > 0.6$) with independent SNPs.

3.9 Enrichment in cell-type-specific regulatory regions

We used FORGE2 to investigate cell-type-specific enrichment within DNase I-hypersensitive sites (DHS) as determined in ENCODE and 15 chromatin states as determined in the Epigenomics Roadmap Project[139]. Top SNPs in 42 loci along with additional independently associated SNPs as determined by COJO for each trait (n=84). FORGE2 assesses the overlap of GWAS SNPs with functional elements for each cell type sample as determined by ENCODE or Roadmap Epigenomics Project while counting the number of overlaps. Background sets of SNPs are obtained by picking the sets with the same number of SNPs as the input SNP set based on MAF, distance of the SNP to transcription start site (TSS) and GC content. Once matched background sets are obtained (default n=1000), they are overlapped with functional elements and the background distribution is derived. Then the -log10(p-value) of the test overlap count and background distribution is calculated. The enrichments at FDR < 0.05 are considered significant.

3.10 Pleiotropy with gene expression

One approach in prioritization of genes in genomic loci was determining whether there is shared pleiotropy between IgG N-glycosylation and expression of genes in discovered genomic regions. This approach is used to explore whether GWAS signals from gene expression (eQTL analysis) colocalize with GWAS signals in the disease or trait of interest, thereby indicating a potential molecular mechanism through which the disease or trait is regulated[140]. The colocalization test is applied under the assumption of a single causal variant in the region. If

the test is positive, the two traits are described as colocalized traits and there is a high probability that the traits share the same underlying causal variant[141,142].

The colocalization of IgG N-glycosylation GWAS signals and gene expression was estimated using Approximate Bayes Factor (ABF) method[143] as implemented in coloc package[68] in R which uses summary-level GWAS data as input. The method outputs 5 posterior probabilities (PP0, PP1, PP2, PP3, PP4), one for each of the five hypotheses: H0) no association with either of the two traits, H1) association with trait 1, but not with trait 2, H2) association with trait 2, but not with trait 1, H3) association with both trait 1 and trait 2, but two independent genetic variants and H4) association with both trait 1 and trait 2 and one shared genetic variant. The gene expression data was obtained from the publicly available eQTLgen[144] dataset (https://www.eqtlgen.org/cis-eqtls.html) which was derived from whole blood samples from 31,684 individuals across 37 cohorts. The posterior probabilities were computed for each of the genes found in IgG N-glycan GWAS loci and which has statistically significant *cis*-eQTLs in the gene expression data using the default values for prior probabilities as $p1 = 1 \times 10^{-4}$, $p2 = 1 \times 10^{-4}$ and $p12 = 1 \times 10^{-5}$. The method uses SNP p-values and MAF to derive the posterior probabilities. The threshold of 75% for PP4 (probability of the same shared variant for two traits) was used for positive colocalization and strong support for prioritization of the gene in the given genomic locus. Association patterns in the given loci were visualized using "locuscomparer" R package.

3.11 Pleiotropy with complex diseases and traits

To explore pleiotropy between IgG N-glycosylation and complex disease and traits, colocalization by ABF was also applied to IgG N-glycan trait GWAS and GWAS for multiple autoimmune diseases and traits for which there is previous evidence for the presence of aberrant IgG N-glycosylation[6] or if there is shared genetic variant as shown in the results from Phenoscanner. The GWAS summary statistics for diseases and traits derived from individuals of European descent were downloaded via GWAS Catalog. The details of each study and download links are listed in Supplementary Table 9. The input for IgG N-glycan GWAS was restricted to the summary statistics for the trait with the lowest p-value of association in the given locus. The default values for prior probabilities were used: $p1 = 1 \times 10^{-4}$, $p2 = 1 \times 10^{-4}$ and $p12 = 1 \times 10^{-5}$. The posterior probabilities by ABF were computed using beta and standard error values if available in the dataset, otherwise, SNP p-values and MAF were used. The threshold of 75% for PP4 (probability of the same shared variant for two traits) was used for

positive colocalization and evidence of high confidence for pleiotropy between IgG glycosylation and disease. The suggestive threshold of 50% was also used to assess the potential case of pleiotropy.

3.12 Genome-wide gene-based association test

FUMA implements Multi-marker Analysis of GenoMic Annotation (MAGMA)[145], a tool used for gene and gene-set analysis of GWAS data. The gene analysis is based on a multiple linear principal components regression model which is used to derive P-value for the gene[146]. The SNP matrix for a gene is projected onto its principal components (PCs) and PCs with small eigenvalues are pruned away while the remaining PCs are used as predictors in the linear regression. However, this requires genotype-level data, but MAGMA implements a method that can use summary statistics to perform gene analysis in a SNP-wise model. A gene test statistic is computed by combining the P-values of the individual SNPs in a gene. LD estimates from a reference dataset with similar ancestry are used to account for LD between SNPs. MAGMA v1.07 implements SNP-wise mean model which derives a mean χ^2 statistic and a p-value is obtained by using a known approximation of sampling distribution[147,148]. 1000G Phase 3[48] was used as a reference panel for the estimation of LD. The P-value threshold used for MAGMA results was 5.28×10^{-7} (0.05/18,934 tested genes/5 PCs which explain 99% of the variance in glycan traits).

3.13 Gene-set enrichment analysis

Gene-set enrichment analysis was performed by the GENE2FUNC tool in FUMA[69]**.** The list of prioritized genes was submitted as input. The tool performs hypergeometric tests to test whether prioritized genes are overrepresented in the previously defined gene sets. The GENE2FUNC analysis was based on Molecular Signature Database v7.0 (MSigDB) [149]. "All" was set as the option for background genes used in the enrichment tests with a total of 57,241 background genes used. To correct for the multiple testing in each category (canonical pathways, GO processes), Benjamini-Hochberg false discovery rate (FDR) procedure was used. The gene-sets were reported if at least two genes belonged to the gene set and the adjusted P-value for the gene set was < 0.05. To identify gene sets describing the higher-level biological pathways, a set of keywords for each pathway was used to filter the results of the gene-set enrichment analysis.

3.14 Network analysis

To construct a potential regulatory network of genes for IgG N-glycosylation, the GWAS summary statistics for top associated SNPs in each region across all eleven IgG N-glycan traits were used as described in Klarić *et al.*[10]. The effect sizes for all SNPs across the glycan traits were transformed into Z-scores, thereby accounting for uncertainty in effect size estimates. The Z-scores were calculated as Z-score= beta/SE, where beta corresponds to the effect of a SNP on a given glycan trait and SE is the standard error of the SNP effect. The resulting 11 vectors each consisting of 42 Z-score values were used to obtain a pairwise correlation matrix using Pearson's correlation. The correlation matrix was pruned for correlation values that had P-value > 5.80×10^{-5} (p-value > 0.05/861, where 861 is the number of unique pair-wise correlation tests (42x41/2)). The network was visualized using Cytoscape software[150] (v 3.7.2), where the nodes were represented by the prioritized genes in the region of top SNPs and edges were added based on the presence of significant correlation between the top SNPs. The colour of the edge corresponded to the squared Pearson's correlation coefficient (to account for directionality) between the two top SNPs.

The STRING protein-protein interaction (PPI) database (v. 11.0)[70] was used to construct a PPI network using the list of prioritized genes as input. The network nodes were represented by the proteins coded by the genes and edges were constructed if there is evidence for the functional association of the two proteins from 1) curated databases, 2) experimental evidence, 3) protein homology, and 4) co-expression of the two proteins. Only protein-protein associations with significant enrichment (FDR < 0.05) within each evidence category were constructed.

4. RESULTS

4.1 Harmonization of UPLC and LCMS data

Previously, there were no GWA meta-analyses of the human IgG N-glycome using GWAS of both UPLC and LCMS IgG N-glycome measurements, therefore making it necessary to first test how the two types of data correlate and what methods should be applied in pre-processing to make them comparable. For this purpose, the CROATIA-Vis cohort (n=661) was used since both UPLC and LC-MS glycan data were available in the same samples. The correlation values of corresponding traits defined in UPLC and LC-MS data after different normalization and subclass weighting approaches are shown in Table 4. The different weighting approaches for IgG subclass measurements are shown in total area normalized data since there are negligible differences between median quotient, total area and largest peak normalizations.

Table 4: Pearson's correlation coefficients for UPLC and LC-MS derived glycan values after applying different normalization types and weighting of the subclass-specific values in LC-MS measurements. Combination of normalization types as applied in UPLC- and LC- MS-derived data is shown on the top.

UPLC	Total area	Largest peak	Median quotient	Total area	Total area	Total area	Total area	Total area	Total area	Total area
LCMS	Total area	Largest peak	Median quotient	Total area with trait subclass average	Total area with weighted subclass average	Total area per subclass with weighted average	Total area per subclass	Total area per subclass with trait average	Total area with subclass concentration applied to raw values	Total area with response factor applied to raw values
Fucosylation	0.43	0.43	0.43	0.45	0.46	0.40	0.39	0.39	0.39	0.35
G1	0.78	0.77	0.78	0.75	0.80	0.79	0.75	0.75	0.79	0.73
G2	0.91	0.91	0.91	0.91	0.91	0.91	0.90	0.90	0.91	0.90
Galactosylation	0.96	0.96	0.96	0.95	0.97	0.97	0.95	0.95	0.97	0.95
Monosialylation	0.83	0.83	0.83	0.84	0.84	0.83	0.83	0.83	0.83	0.81
Bisecting	0.85	0.84	0.85	0.83	0.85	0.85	0.84	0.84	0.85	0.83
Monosialylation without bisecting GlcNAc	0.85	0.85	0.84	0.85	0.87	0.86	0.84	0.84	0.85	0.82

The highest correlation values between UPLC and LC-MS traits were obtained in galactosylation traits (G1, G2, total galactosylation). The traits are denoted by capital letters but are also written in small letters (g1, g2) throughout the text. For example, when applying median quotient normalization the values range from 0.78 for G1 to 0.96 for total

galactosylation. The lowest correlation was observed in the fucosylation trait where Pearson's correlation coefficient ranged from 0.35 to 0.46.

Including response factor values and concentrations of the IgG subclasses did not improve the correlation. The additional validation was done by comparing the GWAS output obtained by different harmonization approaches. For the sake of generalization and easier interpretation, we decided to use median quotient normalization (across whole IgG N-glycome in both UPLC and LC-MS) for our purposes. In Supplementary Table 5, the descriptive statistics for the derived glycan traits are shown across all cohorts used in the genome-wide association meta-analysis.

4.2 Quality control of genome-wide association studies

Discovery meta-analysis of IgG N-glycome was performed in seven cohorts of European descent (ORCADES, TwinsUK, CROATIA-Korcula, CROATIA-Vis, CROATIA-Split, VIKING, EPIC) in a total of 13,705 samples. Prior to meta-analysis, the individual GWAS summary statistics were checked for potential issues with the quality of the tested SNPs or analytical issues. Number of SNPs excluded due to nonsense values for beta and standard error, monomorphic state, low minor allele count (MAC ≤ 6), low imputation quality, allele mismatch and allele frequency outliers, are listed in Supplemetary Table 7. The largest number of SNPs excluded due to invalid beta and SE values and monomorphic state was in the TwinsUK cohort since no QC was applied to the HRC imputed data prior to the genome-wide association scan. The number of SNPs that passed QC ranged between 7,021,984 in EPIC (sample subset 4) and 14,255,310 in TwinsUK (batch 3&4). The allele frequencies were checked against 1000G Phase 3 EUR reference panel and outliers were removed (Supplementary Figure 2). The reported allele in the LLS cohort was opposite of the allele in the reference panel and other cohorts. The corresponding analyst was contacted and GWAS summary statistics with the appropriate allele was provided.

The inflation factor λ_{GC} ranged between 0.98 and 1.03 which indicated no inflation in summary statistics for individual cohorts due to population stratification (Supplementary Figure 1). The SE-N plots, PZ plots and QQ plots were checked for potential analytical problems due to trait transformation, samples size or fitted statistical model. A slight departure from the identity line in SE-N plots was noted (Supplementary Figure 3), thereby indicating potential analytical

issues. The cleaned GWAS summary statistics for all cohorts were included in the meta-analysis for eleven glycan traits.

4.3 Genome-wide significant genomic loci (Discovery)

The correction of genome-wide significance threshold P-value < 5 x 10^{-8} was applied due to the testing of eleven glycan traits. Given the high correlation between the tested traits, principal components analysis was used to derive the number of principal components that explain 99% of the variance among tested traits. Glycan data from ORCADES, CROATIA-Korcula and CROATIA-Vis cohorts was used. The number of principal components that explain 99% of the variance in the tested traits was five (Figure 4)**,** thus the applied genome-wide significance threshold was 1 x 10^{-8} (5 x 10^{-8}/5).

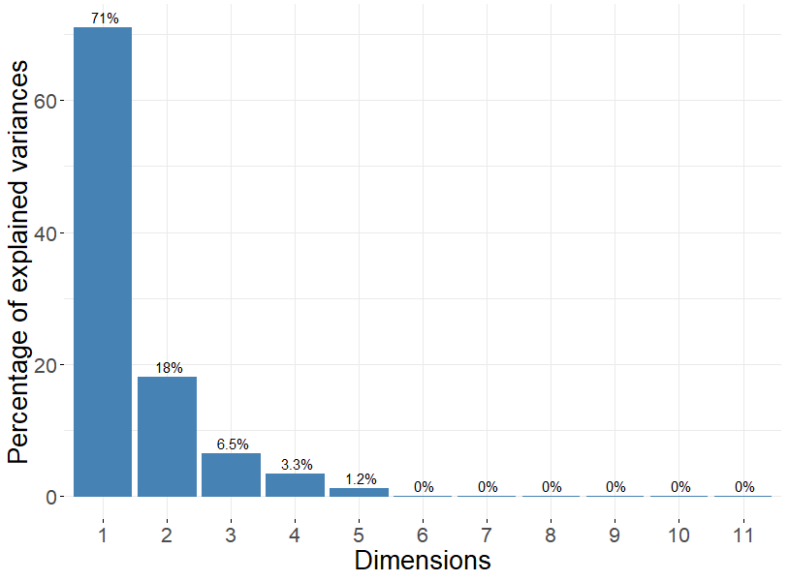


Figure 4: Number of principal components explaining 99% of the variance in eleven IgG N-glycan traits in ORCADES, CROATIA-Korcula and CROATIA-Vis samples.

The GWAS summary statistics of meta-analysis for eleven glycan traits were merged to create input for FUMA SNP2GENE[69] function. Additionally, FUMA SNP2GENE output was created for each trait separately using the same parameters. First, the significant SNPs were identified using a genome-wide significance threshold of 1 x 10^{-8} and 1000G Phase3 reference panel for the European population. SNPs with LD r^2 < 0.6 were considered as independent SNPs (n=444), furthermore, independent significant SNPs with $r^2 < 0.1$ were identified as lead SNPs (n=128). Start and end positions for genomic risk loci were defined by merging lead SNPs if they were found in the 250kb window. For the downstream analysis, all SNPs in LD ($r^2 \geq 0.6$)

with one of the independent significant SNPs were considered as candidate SNPs (n=12,348). A total of 42 genomic loci across nineteen chromosomes were identified.

Seventeen genomic regions were associated with one glycan trait only, eight regions with two traits, two regions with three traits, four regions with four traits, three regions with five traits, and the remaining eight regions were associated with six to ten glycan traits.

The strongest association is the association between s1_g2 (ratio of monosialylation and digalactosylation) trait and genetic variant (rs11710456; beta=-0.536, SE=0.012, p-value=1.44 x 10^{-444}) in region on chromosome 3 which harbours the *ST6GAL1* gene.

4.4 Replication of genomic loci from previous GWAS of IgG N-glycome

The previous GWAS of IgG N-glycome were performed either on LC-MS or UPLC derived glycan phenotypes and in univariate or multivariate association analysis. However, given that all IgG N-glycome GWA studies aim to identify as many genomic regions involved in IgG N-glycosylation as possible, the overlap of the previously identified regions and the regions discovered here was checked by assessing the top SNP or SNP in high LD across all eleven glycan traits.

There are 27 genomic loci identified by Klarić *et al.*[10], 25 of which are replicated in the current meta-analysis (Table 5). One of the regions which were not significantly associated (rs11748193; min P-value =8.97 x 10^{-5}; fuc) with any of the glycan traits in the current analysis was previously associated with only one of the glycan traits, IGP2, the percentage of A2 glycan in total IgG glycans. As such, IGP2 glycan structure is not captured by any of the derived traits used in this study, thus explaining why the association was not replicated. Also, nonreplicated locus (P-value= 0.03612) on chromosome 19 (lead SNP rs874232) harbours *FUT6* gene which codes for the fucosyltransferases enzyme involved in antennary fucosylation which was not tested in the current study. It is important to note that the two loci on chromosome 9 (top SNPs rs10813951 and rs12341905) are merged into one locus in the new study as they overlap due to high LD.

Table 5: Comparison of p-values for lead SNPs from Klarić *et al.* and the new meta-analysis.

Variant name	Chr:pos (hg19)	Prioritized gene(s)	Klarić *et al*. glycan trait	Klarić *et al*. P-value	New P-value	New glycan trait
rs10903118	1:25294878	*RUNX3*	IGP74	5.14E-13	9.52E-14	bisecting
rs7621161	3:186727170	*ST6GAL1*	IGP29	4.65E-276	1.63E-435	s1_g2
rs7700895	5:95273410	*ELL2*	IGP35	1.20E-14	4.60E-08	s1_gal_total
rs11748193*	5:131725329	*IRF1; IL3, SLC22A4*	IGP2	4.31E-10	8.97E-05	fuc
rs3099844	6:31448976	*HLA region*	IGP15	1.12E-13	2.94E-09	s1_no_bis
rs9385856	6:139625074	*TXLNB*	IGP70	5.05E-19	2.27E-34	bisecting
rs7758383	6:143169723	*HIVEP2*	IGP13	9.61E-14	8.42E-25	s1_g2
rs6964421	7:6520676	*DAGLB*	IGP14	5.31E-11	7.84E-13	bisecting
rs6421315	7:50355207	*IKZF1*	IGP62	4.70E-27	3.64E-34	fuc
rs7812088	7:150919829	*ABCF2*	IGP2	2.06E-22	5.38E-20	s1_no_bis
rs10096810	8:103545436	*ODF1*	IGP77	9.52E-11	2.43E-10	g0
rs10813951	9:33128021	*B4GALT1*	IGP17	8.84E-34	1.41E-59	s1_g1
rs12341905	9:33205136	*SPINK4*	IGP53	1.46E-09	4.52E-21	g2
rs481080	11:114442265	*NXPE1; NXPE4*	IGP29	1.05E-16	3.87E-16	s1_g2
rs11847263	14:65775695	*FUT8*	IGP42	1.13E-58	1.05E-120	fuc
rs4074453	14:105998544	*TMEM121*	IGP48	3.82E-29	2.30E-20	s1_g2
rs250555	16:23444268	*GGA2; COG7*	IGP26	6.76E-10	5.20E-12	s1_gal_total
rs7216389	17:38069949	*ORMDL3; GSDMB; IKZF3; ZPBP2*	IGP59	1.17E-15	1.39E-24	fuc
rs199456	17:44797919	*CRHR1; SPPL2C; MAPT; ARHGAP27*	IGP14	6.76E-14	4.23E-16	s1_g2
rs11651000	17:45835278	*TBX21*	IGP59	2.66E-12	3.59E-11	fuc
rs2725391	17:79192430	*SLC38A10; CEP131; TEPSIN*	IGP24	9.91E-16	6.18E-29	s1_gal_total
rs874232*	19:5843609	*FUT6*	IGP12	7.85E-13	0.03612	fuc
rs7257072	19:19267990	*RFXANK*	IGP9	1.59E-13	3.44E-10	bisecting
rs2745851	20:17829280	*MGME1*	IGP38	4.61E-13	1.67E-10	s1_g2
rs7281587	21:36565278	*RUNX1*	IGP45	1.13E-13	6.78E-21	bisecting
rs17630758	22:24136542	*SMARCB1; CHCHD10; VPREB3*	IGP66	1.20E-41	5.71E-71	bisecting
rs5750830	22:39840828	*MGAT3*	IGP40	7.74E-69	1.21E-86	bisecting

*not replicated in the new meta-analysis

In the comparison of the P-values for the two loci in Lauc *et al.*[7] which were not replicated in any of the subsequent GWAS of IgG N-glycome, it is shown that the genome-wide significance is not reached in the current study (Table 6). A genetic variant in the *IL6ST-ANKRD55* locus on chromosome 5 was also identified in the current GWAS but was not in high LD with rs17348299, a genetic variant from the same locus reported in Lauc *et al.* (LD $r^2 < 0.001$).

Table 6: Comparison of p-values for previously nonreplicated loci from Lauc *et al.* and the new meta-analysis

Variant name	Chr:pos (hg19)	Prioritized gene(s)	Lauc et al. glycan	Lauc et al. P-value	New P-value	New trait
rs17348299*	5:55322895	*IL6ST-ANKRD55*	IGP53	6.88E-11	3.93E-05	g2
rs4930561*	11:67931761	*SUV420H1*	IGP41	8.88E-10	0.01078	g0

*not replicated in the new meta-analysis

Three out of six novel loci (rs4561508 and rs11895615) in Shadrina *et al.*[74] multivariate GWAS are replicated in the current study, with the current lead variant in the locus on chromosome 11 being in high LD ($r^2 = 0.92$) with Shadrina *et al.* variant from the same locus (Table 7)..

Table 7: Comparison of p-values for novel loci from Shadrina *et al.* and the new meta-analysis

Variant name	Chr:pos (hg19)	Prioritized gene(s)	Shadrina et al. trait	Shadrina et al. P-value	New P-value	new trait
rs479844**	11:65551957	*OVOL1*	N-glycosylation	1.97E-13	2.68E-06	g1
rs12049042*	1:246288812	*SMYD3*	Galactosylation	1.20E-09	0.126	bisecting
rs4561508	17:16848750	*TNFRSF13B*	N-glycosylation	1.38E-10	1.26E-16	g1
rs11895615	2:26113120	*ASXL2*	Bisecting GlcNAc	5.69E-10	2.29E-09	g1
rs1372288*	3:142901537	*CHST2; SLC9A9*	N-glycosylation	8.73E-11	0.0812	g2
rs12635457*	3:196203979	*RNF168*	N-glycosylation	1.61E-13	0.0001952	fuc

*not replicated in the new meta-analysis; **in high LD with lead SNP in current GWAS (rs10896045; P-value= 2.61E-09)

Table 8: List of genome-wide significant loci in IgG N-glycome GWAS. Top SNP position- Position of the SNP with the strongest association with the glycan trait denoted by chromosome: base pair coordinate in GRCh37 (hg19) build; rsID- rsID identifier for the top SNP; Strongest associated trait- trait with the lowest p-value in association with the top SNP in the genomic locus; Chr-chromosome; Start- starting position for the genomic locus; End- end position for the genomic locus; EA- effect allele for which the effect is reported; OA- non-effect allele; P-value- lowest P-value for the top SNP in the locus; Beta- effect estimate for the effect allele in the top SNP; SE- standard error of the effect estimate; Associated traits- all traits that are significantly associated with the genomic locus; Repl- Beta, SE and P-value for replication analysis; Effect, SE and P-value for replication analysis reported for a SNP in LD with top SNP in the region are shown in italic.

Top SNP position	rsID	Strongest associated trait	Chr	Start	End	EA	OA	EAF	P-value	Beta	SE	Associated traits	Repl Beta	Repl SE	Repl P-value
1:25291697	rs188468174	bisecting	1	23526335	25903455	T	C	0.012	1.42E-134	1.419	0.058	bisecting;g0;g1;g2;gal_total;s1_g1;s1_gal_total;s1;s1_no_bis	1.325	0.079	7.15E-63
1:39302020	rs7548054	fuc	1	39302020	39380385	T	C	0.337	9.26E-10	0.079	0.013	fuc	0.027	0.017	1.14E-01
1:233723112	rs6689354	s1_g2	1	233715572	233740757	A	G	0.481	5.43E-09	-0.068	0.012	s1_g2	-0.016	0.015	2.77E-01
2:26139430	rs111919630	g0	2	26109539	26149988	T	C	0.336	3.80E-10	-0.062	0.01	g0;g1;gal_total	-0.049	0.015	1.03E-03
2:101991907	rs10186962	s1_g1	2	101930890	101991907	A	G	0.600	2.38E-09	0.066	0.011	s1_g1	-0.006	0.016	7.29E-01
2:158469050	rs77539041	fuc	2	158413902	158477773	C	G	0.055	1.25E-09	0.166	0.027	fuc	0.140	0.036	1.19E-04
3:186725887	rs11710456	s1_g2	3	186607935	186819448	A	G	0.282	1.44E-444	-0.536	0.012	s1_g1;s1_g2;s1_gal_total;s1;s1_no_bis	-0.483	0.016	2.92E-196
4:103519487	rs3774964	g2	4	103390496	103554821	A	G	0.625	1.56E-11	-0.065	0.01	g0;g2;gal_total;s1;s1_no_bis	*0.061*	*0.015*	*4.93E-05*
5:55438851	rs10065637	s1_g2	5	55436851	55444683	T	C	0.216	2.74E-10	-0.09	0.014	s1_g2	-0.094	0.019	1.36E-06
5:95240996	rs11741563	s1_no_bis	5	95217242	95324375	T	C	0.263	1.87E-09	-0.067	0.011	s1_gal_total;s1;s1_no_bis	-0.090	0.017	4.80E-08
6:22053674	rs113557827	g2	6	22053674	22053674	A	G	0.988	5.77E-10	0.633	0.102	g2;s1_no_bis	0.050	0.153	7.46E-01
6:31351764	rs2442752	s1_gal_total	6	30798697	32879471	T	C	0.624	4.05E-12	0.083	0.012	bisecting;fuc;g0;g2;gal_total;s1_g1;s1_g2;s1_gal_total;s1;s1_no_bis	*0.095*	*0.018*	*1.11E-07*
6:74230859	rs3822960	s1_g1	6	74168723	74285118	T	C	0.668	2.17E-10	0.073	0.012	g2;s1_g1	0.039	0.017	2.15E-02

6:139629524	rs4543384	bisecting	6	139617590	139636003	T	C	0.422	2.93E-37	-0.146	0.011	bisecting	-0.139	0.015	5.30E-20
6:143169723	rs7758383	s1_g2	6	143088071	143206826	A	G	0.502	8.42E-25	-0.119	0.012	g0;g1;g2;gal_total;s1_g1;s1_g2	-0.152	0.015	3.43E-24
7:6531268	rs7786067	bisecting	7	6497501	6550403	T	C	0.322	6.77E-13	0.086	0.012	bisecting;s1_g2	0.063	0.016	9.05E-05
7:50352695	rs7789913	fuc	7	50325563	50362999	T	C	0.389	1.69E-34	0.153	0.013	bisecting;fuc	0.188	0.016	1.20E-31
7:150942349	rs113745074	s1_no_bis	7	150902419	150969535	T	C	0.887	6.98E-22	-0.153	0.016	bisecting;fuc;g0;g2;gal_total;s1_g1;s1_gal_total;s1;s1_no_bis	-0.142	0.022	1.81E-10
8:103545983	rs13250010	g0	8	103542538	103550211	T	G	0.358	7.71E-11	0.063	0.01	g0;gal_total	0.062	0.015	3.47E-05
9:33124872	rs12342831	s1_g1	9	32932194	33385427	T	C	0.744	1.26E-59	0.197	0.012	g0;g1;g2;gal_total;s1_g1;s1_g2;s1_gal_total;s1;s1_no_bis	0.230	0.018	4.96E-39
10:94446635	rs10786052	fuc	10	94336963	94495241	T	C	0.481	3.43E-09	-0.072	0.012	fuc	-0.109	0.017	5.28E-11
11:65555524	rs10896045	g1	11	65555524	65555524	A	G	0.289	2.61E-09	0.079	0.013	g1	0.079	0.017	2.42E-06
11:114381448	rs1671819	s1_g2	11	114298893	114450529	A	G	0.471	1.03E-16	-0.096	0.012	s1_g2;s1_gal_total	-0.081	0.015	6.08E-08
12:121202664	rs9431	s1_gal_total	12	121188641	121351934	A	C	0.484	2.93E-11	0.077	0.012	s1_g1;s1_gal_total;s1;s1_no_bis	0.072	0.016	8.05E-06
14:65775695	rs11847263	fuc	14	64709913	66303683	T	G	0.632	1.05E-120	0.288	0.012	fuc	0.280	0.017	1.92E-59
14:106113281	rs10444775	s1_g2	14	105877057	106270813	C	G	0.452	1.08E-36	0.216	0.017	s1_g2	0.230	0.024	2.06E-21
16:23412310	rs30017	s1_gal_total	16	23397113	23613191	A	G	0.206	2.24E-13	0.108	0.015	s1_g1;s1_gal_total;s1;s1_no_bis	0.063	0.019	6.48E-04
17:16842991	rs34562254	g1	17	16820099	16875636	A	G	0.118	1.48E-18	0.166	0.019	g1	0.092	0.024	9.95E-05
17:38072727	rs2872516	fuc	17	37579383	38215117	T	C	0.554	9.34E-27	0.131	0.012	fuc	0.086	0.016	3.98E-08
17:44331214	rs554899842	s1_g2	17	43463493	44865603	T	C	0.808	4.82E-18	0.137	0.016	s1_g2;gal_total	0.126	0.020	3.90E-10
17:45809822	rs11650451	fuc	17	45766846	45874272	A	G	0.162	6.57E-12	-0.114	0.017	fuc;g2	-0.076	0.021	3.36E-04
17:56410041	rs2526377	s1_no_bis	17	56398006	56417002	A	G	0.565	2.64E-11	0.067	0.01	g2;s1_g1;s1_gal_total;s1;s1_no_bis	0.068	0.015	4.33E-06
17:79218714	rs2659005	s1_gal_total	17	79140505	79275406	T	C	0.451	1.57E-39	0.153	0.012	g0;g2;gal_total;s1_g1;s1_g2;s1_gal_total;s1;s1_no_bis	0.135	0.017	2.70E-15

19:1657741	rs72989754	bisecting	19	1576098	1658699	T	G	0.148	4.44E-15	-0.136	0.017	bisecting	-0.118	0.025	1.58E-06
19:19294091	rs68147405	s1_g2	19	19260586	19298099	T	C	0.550	9.93E-12	0.079	0.012	bisecting;s1_g2	0.058	0.015	1.01E-04
20:4115720	rs56260553	bisecting	20	4115720	4115975	T	C	0.214	3.60E-12	-0.098	0.014	bisecting	-0.045	0.019	1.84E-02
20:17831618	rs2618590	bisecting	20	17820309	17833534	T	C	0.582	6.31E-11	0.076	0.012	bisecting;s1_g2;s1_gal_total	0.062	0.015	5.28E-05
20:50077482	rs4809845	fuc	20	50054190	50077482	T	C	0.760	4.03E-09	0.085	0.015	fuc	0.083	0.019	1.21E-05
20:61598731	rs7271712	s1_no_bis	20	61573062	61639750	T	C	0.026	1.54E-11	-0.222	0.033	s1_g1;s1_gal_total;s1;s1_no_bis	-0.211	0.050	2.63E-05
21:36564553	rs8129053	bisecting	21	36524140	36787961	T	C	0.750	6.20E-22	0.126	0.013	bisecting;g0;g2;gal_total;s1;s1_no_bis	0.121	0.018	6.90E-12
22:24179922	rs3177243	bisecting	22	23951753	24193924	C	G	0.154	4.43E-76	-0.284	0.015	bisecting	-0.304	0.020	1.02E-50
22:39845898	rs1005522	bisecting	22	39590915	39973151	A	C	0.743	1.16E-91	-0.261	0.013	bisecting	-0.327	0.018	3.77E-78

4.5 Replication analysis

Replication analysis of the genome-wide significant SNPs was performed on derived glycans traits which were calculated from UPLC or LC-MS measured glycans from four European cohorts: GCKD, EGCUT, KORA F4 and LLS. When looking into the same glycan-SNP association as in discovery analysis, 34 genome-wide loci were replicated at the significance threshold of 0.001 ($P \leq 0.05/42$) (Table 1). Additional two associations were replicated when considering SNPs which are found in the defined genomic region: rs9266231 in chr6:30798697-32879471 region **(**fuc; P-value = $1.11x\ 10^{-7}$**)** and rs11097786 in chr4:103519487-103390496 (g2, P-value = $4.93x10^{-5}$)

Given that the replication study was underpowered compared to the discovery analysis (two times smaller sample size) directional consistency of the effect estimates between the two analyses was compared. The additional two non-replicated loci (rs56260553 and rs3822960) were significant at nominal significance threshold < 0.05 and the direction of the effects was the same as in the corresponding discovery SNP, thus we expect them to replicate in a larger sample size. For rs7548054 and rs6689354, the P-value was > 0.05, however, the effect direction was consistent with the direction in the discovery cohort, so we expect it might replicate in a larger sample as well.

4.6 Trait specific associations

Given that GWAS was performed for eleven glycan traits which enable a more straightforward interpretation of the results, trait-specific associations are further shown. Structures with bisecting GlcNAc and core fucose are each represented by only one traits. The glycan structures with galactose are represented by g0, g1, g2 and gal_total traits and monosialylation phenotype is described by s1, s1_no_bis, s1_g1, s1_g2 and s1_gal_total traits.

Bisecting GlcNAc

There are thirteen genomic regions associated with the trait describing the addition of bisecting GlcNAc to N-glycans on IgG (Figure 5). The strongest association is present on chromosome 1 (rs188468174; p-value=1.42×10^{-134}) in the region where *RUNX3* gene is located. The second strongest association (rs1005522; 1.16×10^{-91}) is found in the region on chromosome 22 which harbours *MGAT5* gene. Five of the associated regions are specific to the bisecting trait: chr6:139617590-139636003 (rs4543384; p-value=2.93×10^{-37}), chr:19-1576098-1658699 (rs72989754, 4.44×10^{-15}), chr20:4115720-4115975 (rs56260553; p-value=3.6×10^{-12}),

chr22:23951753-24193924 (rs3177243; 4.43 x 10^{-76}) and chr22:39590915-39973151 (rs1005522, 1.16 x 10^{-91}).

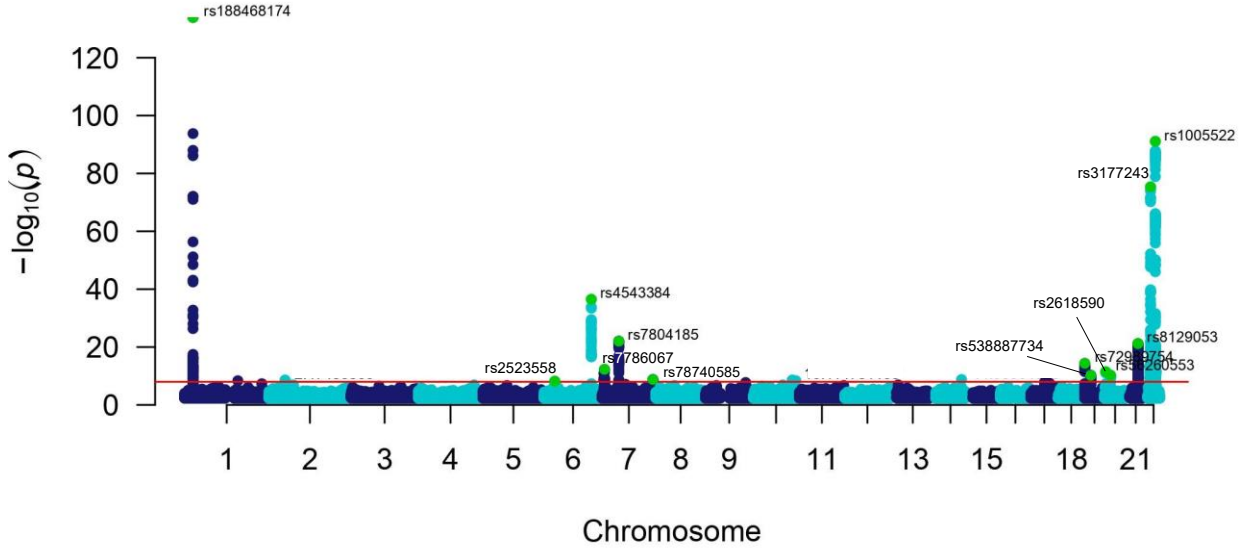


Figure 5: Manhattan plot for the bisecting trait. Plot shows -log10(P-values) of association on y-axis and SNPs ordered by chromosomal location on x-axis. Red line indicates the genome-wide significance threshold (1x10^{-8}).

Fucosylation

A total of ten genomic regions were associated with fucosylation trait, six of which are specific for fucosylation: chr1:39302020-39380385 (rs7548054; P-value=9.26 x 10^{-10}), chr2:158413902-158477773 (rs77539041; P-value=1.25 x 10^{-9}), chr10:94336963-94495241 (rs10786052; P-value=3.43 x 10^{-9}), chr14:64709913-66303683 (rs11847263, P-value=1.05 x 10^{-120}), chr17:37579383-38215117 (rs2872516; P-value=9.34 x 10^{-27}) and chr20:50054190-50077482 (rs4809845; P-value=4.03 x 10^{-9}). The strongest association (rs11847263; P-value=1.05 x 10^{-120}) is present on chromosome 14 in the region which harbours *FUT8* gene (Figure 6).

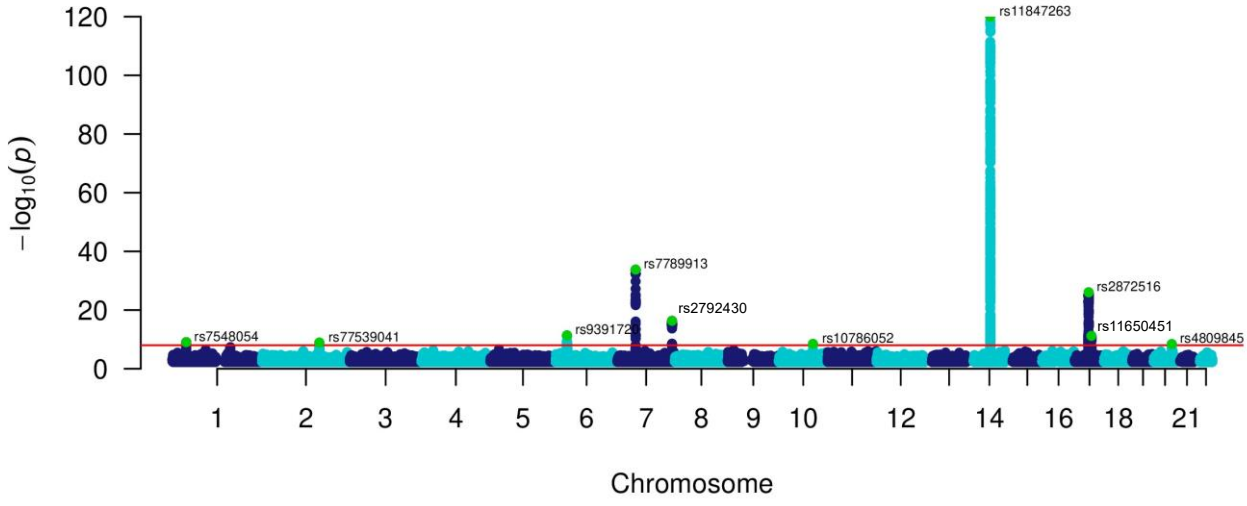


Figure 6: Manhattan plot for fuc trait. Plot shows -log10(P-values) of association on y-axis and SNPs ordered by chromosomal location on x-axis. Red line indicates the genome-wide significance threshold (1x10^{-8}).

Galactosylation

To capture the genetic variation associated with the addition of galactose unit to the IgG N-glycans, four traits were derived and tested including: agalactosylation (g0), monogalactosylation (g1), digalactosylation (g2) and total galactosylation (gal_total). The g0 and gal_total are corresponding traits given that, by definition, their sum equals 1. A single associated region (chr8:103542538-103550211; rs13250010; 7.71 x 10-11) is specific to g0 and gal_total traits (Figure 7). Monogalactosylation (g1) is the phenotype with the least number of genome-wide significant associations (n=6; Figure 8) among which is an association on chromosome 11 (rs10896045; 2.61 x 10^{-9}) which is specific for this trait. A total of twelve genomic regions are associated with digalactosylation (Figure 9). The lack of galactosylation-specific associations is explained by the fact that the presence of galactose is the prerequisite for sialylation to occur, hence, the same associations are captured by sialylation phenotypes.

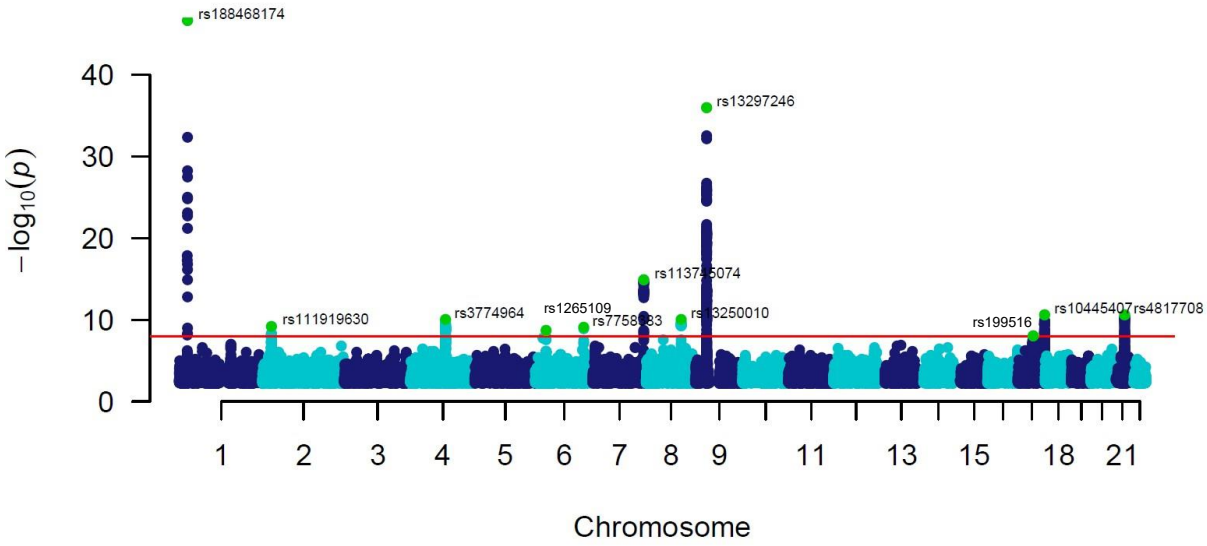


Figure 7: Manhattan plot for total galactosylation. Plot shows -log10(P-values) of association on y-axis and SNPs ordered by chromosomal location on x-axis. Red line indicates the genome-wide significance threshold (1x10^{-8}).

The strongest association for galactosylation traits, except gal_total, is on chromosome 9 in the *B4GALT1* locus. However, two different SNPs in *B4GALT1* locus are shown to be top SNPs for galactosylation traits: rs3780490 (P-value=7.66 x 10^{-52}) for g1 trait and rs13297246 for gal_total (P-value=1.08 x 10^{-36}) and g2 (P-value=4.20 x 10^{-55}) traits. Based on the 1000G Phase 3 reference panel for European population, LD r^2 value for rs3780490 and rs13297246 is 0.1056.

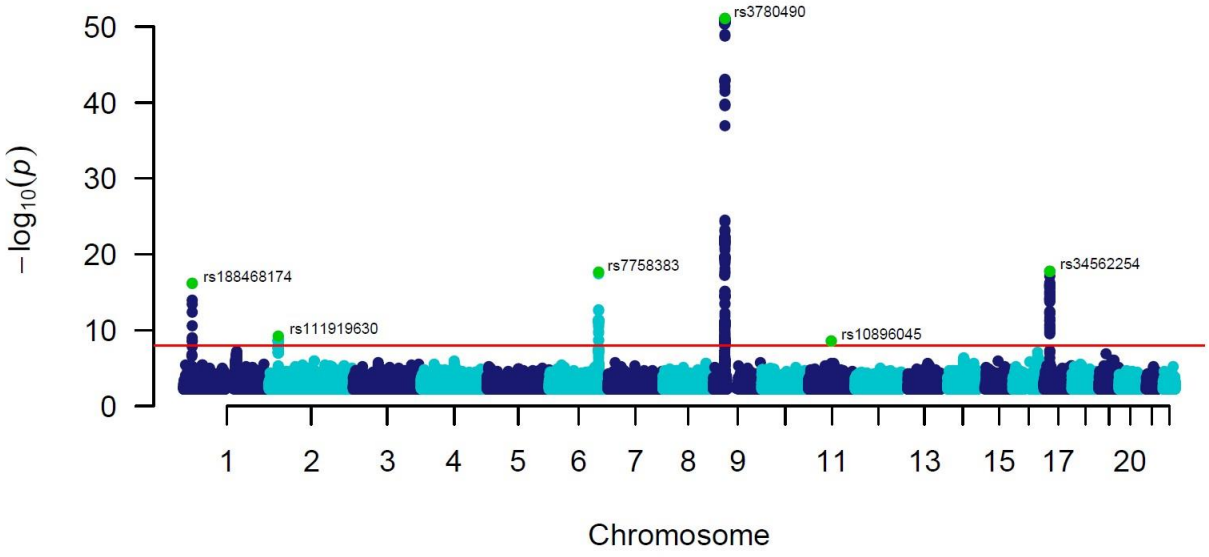


Figure 8: Manhattan plot for g1 trait. Plot shows -log10(P-values) of association on y-axis and SNPs ordered by chromosomal location on x-axis. Red line indicates the genome-wide significance threshold ($1x10^{-8}$).

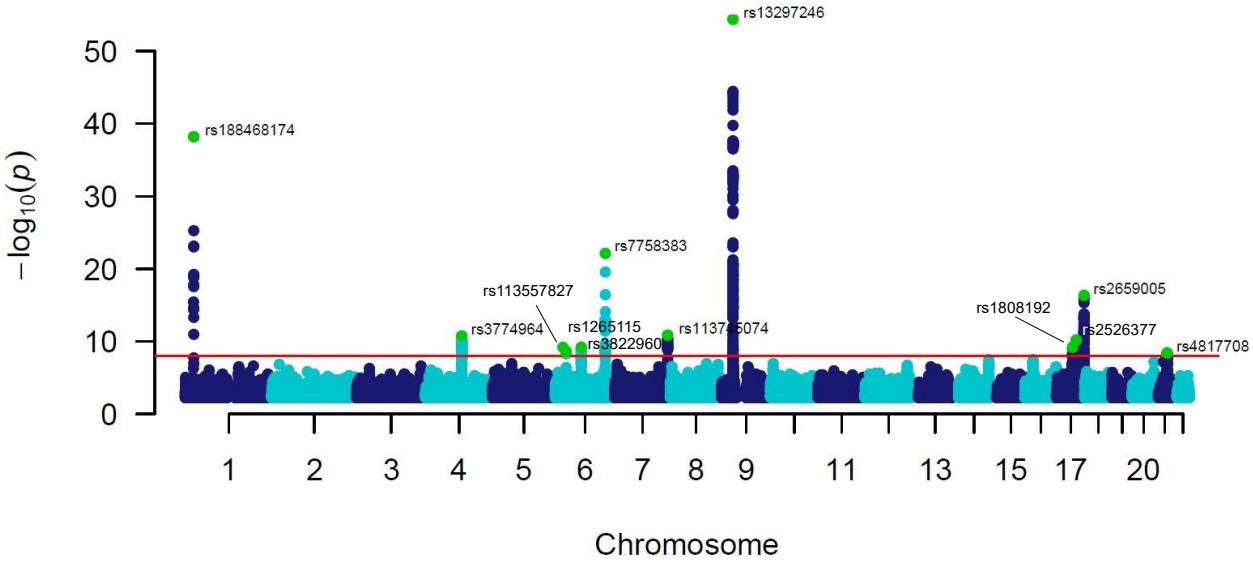


Figure 9: Manhattan plot for g2 trait. Plot shows -log10(P-values) of association on y-axis and SNPs ordered by chromosomal location on x-axis. Red line indicates the genome-wide significance threshold ($1x10^{-8}$).

Sialylation

To link the addition of sialic acid to the genomic regions, five glycan traits were defined: monosialylation (s1), monosialylation without bisecting GlcNAc (s1_no_bis), the ratio of monosialylation and monogalactosylation (s1_g1), the ratio of monosialylation and digalactosylation (s1_g2) and the ratio of monosialylation and total galactosylation (s1_gal_total). Due to the use of LC-MS measured glycan data and thus, lack of measurements

for glycan structures that contain two sialic acids, the GWAS of the sialylation process is limited to the traits describing monosialylation.

The strongest association of sialylation phenotypes is with *ST6GAL1* locus on chromosome 3 with association p-value = 1.44 x 10^{-444} for s1_g2 trait. Ten of the 42 genomic regions are associated specifically with one or more of the sialylation phenotypes. Since sialylation is conditioned on the presence of galactose units on IgG N-glycans, traits including s1_g1, s1_g2 and s1_gal_total were derived to isolate the sialylation specific signals. The trait s1_g1 (ratio of monosialylated and monogalactosylated structures) was associated with thirteen regions, however, the signals on sialylation-specific loci did not become stronger. This could be explained by the very low number of both monogalactosylated and monosialylated structures in the IgG N-glycome, therefore, the s1_g1 trait is not able to capture the monosialylation-specific variation. However, there was one association on chromosome 2 (rs10186962; p-value=2.38 x 10^{-9}) specific for the s1_g1 trait. On the other hand, s1_g2 (ratio of monosialylated and digalactosylated IgG N-glycan structures) was associated with 13 regions (Figure 10), three of which were trait-specific: chr1:233715572-233740757 (rs6689354; p-value=5.43 x 10^{-9}), chr5:55436851-55444683 (rs10065637; p-value=2.74 x 10^{-10}) and chr14:105877057-106270813 (rs10444775; p-value= 1.08 x 10^{-36}). The s1_gal_total trait is associated with thirteen genomic regions all of which are discovered in the remainder of the monosialylation phenotypes.

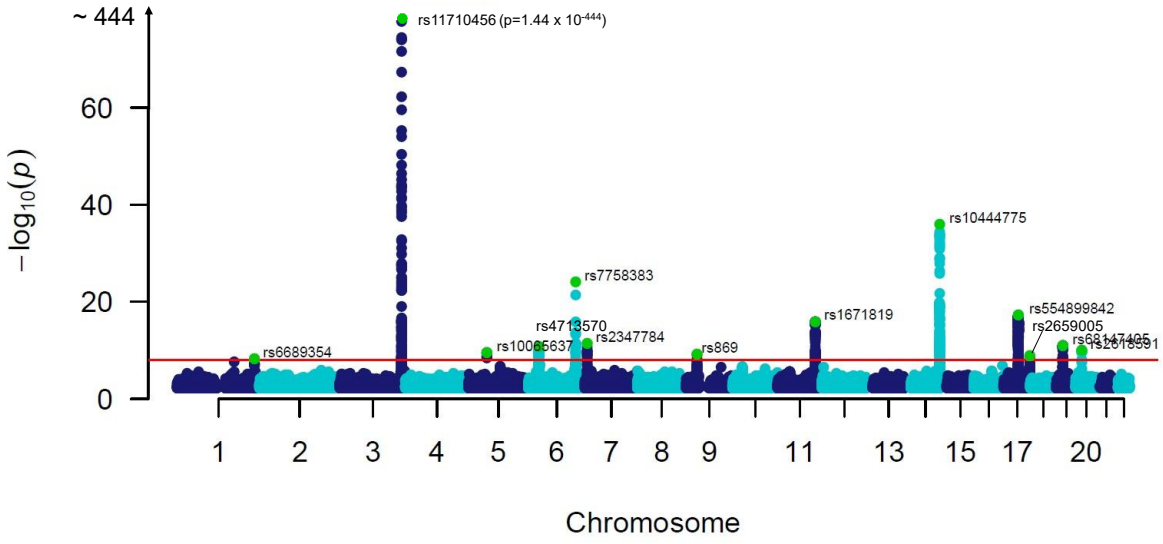


Figure 10: Manhattan plot for s1_g2 trait. Plot shows -log10(P-values) of association on y-axis and SNPs ordered by chromosomal location on x-axis. Red line indicates the genome-wide significance threshold (1x10-8). For simplicity, the y axis is trimmed at -log10(P-value)=100.

4.7 Conditional analysis and variance explained

The strongest association (1.44 x 10^{-444}) was found between rs11710456 in *ST6GAL1* locus on chromosome 3 (chr3:186607935-186819448) and s1_g2 trait representing the ratio between monosialylated and digalactosylated structures in total IgG N-glycome. Klarić *et al.*[10] indicated a different SNP (rs7621161) with strongest association in this locus, however, the two SNPs are in high LD (r^2=0.99). Monosialylation trait (s1) had the strongest association with rs6764279 (P-value=1.1 x 10^{-81}) also in *ST6GAL1* locus and this SNP is in high LD with rs11710456 (r^2= 1). Rs6764279 had slightly higher P-value than rs11710456 for s1_g2 trait (p=8.6 x 10^{-444}), hence rs11710456 was picked by the software as the strongest. Other sialylation phenotypes including s1_g1, s1_gal_total and s1_no_bis, had the strongest association with rs6764279 with P-values of association being 3.34 x 10^{-71}, 4.08 x 10^{-216} and 6.74 x 10^{-89}, respectively. Fucosylation trait (fuc) had the strongest association with rs11847263 (P-value=1.05 x 10^{-120}) in genomic region on chromosome 14 (chr14:64709913-66303683) where *FUT8* gene is located. A genetic variant located in region on chromosome 1 (chr1:23526335-25903455) had a stronger association with bisecting GlcNAc trait (rs188468174; P-value= 1.42 x 10^{-134}) in comparison to *MGAT3* locus (rs1005522; P-value= 1.16 x 10^{-91}) encoding glycosyltransferase which acts in addition of bisecting GlcNAc to the glycan structure.

Digalactosylation trait (g2) had the strongest association (P-value=4.2 x 10^{-55}) with genetic variant rs13297246 in *B4GALT1* (chr9:32932194-33385427) locus, while the monogalactosylation (g1) had the strongest association with variant rs3780490 (P-value=7.66 x 10^{-52}) in the same locus. The two variants are not in a strong LD (r^2=0.1) indicating potentially distinct causal variants for the two galactosylation phenotypes. Considering *B4GALT1* locus, total galactosylation (gal_total) had the strongest association with rs13297246 (P-value=1.08 x 10^{-36}). However, the strongest association among all regions and gal_total was with rs188468174 (P-value=2.1 x 10^{-47}) located in *RUNX3* locus on chromosome 1. Only considering the marginal effects of the top associated variants in all loci for each trait, the phenotypic variance explained ranged from 3.4% for g0 to 18.06% for s1_g2.

The SNP-heritability estimates were derived using LDSC which calculates SNP-based heritability as the proportion in phenotypic variation in the population which is explained by additive effects of the SNPs. The estimates for glycan traits ranged from 0.093 (±0.032) for monogalactosylation (g1) trait to 0.222 (±0.10) for fucosylation trait (Figure 11). Glycan traits

describing the galactosylation phenotypes (g0, g1, g2 and gal_total) have the lowest heritability estimates, while the fucosylation trait has the highest heritability estimate h2g=0.2225 but with high standard error (SE=0.1007).

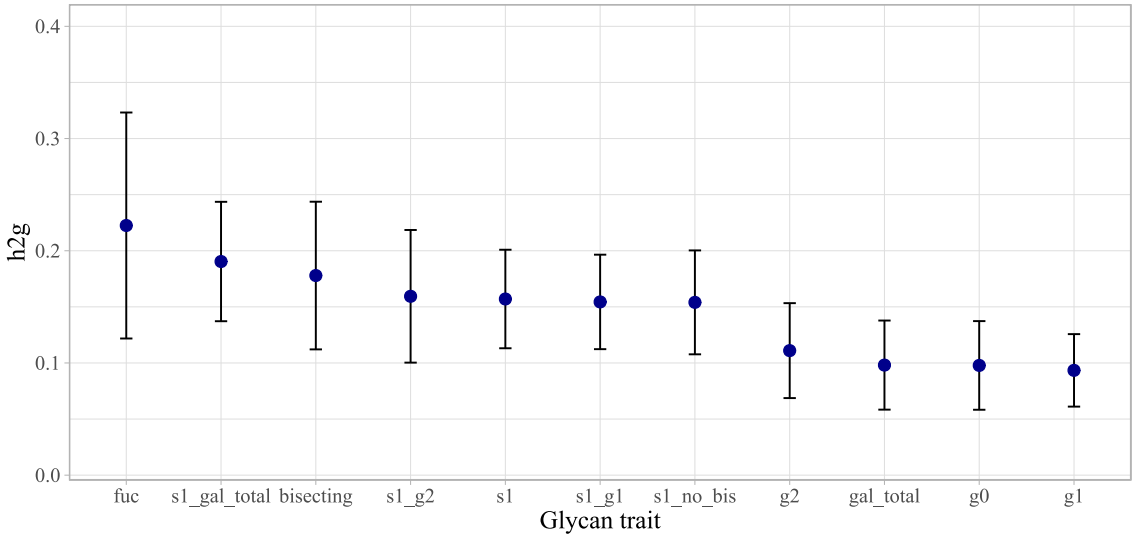


Figure 11: SNP-based heritability values (h^2_{snp}) for 11 IgG glycan traits. Error bars denote ±SE values for the h^2_{snp} estimate

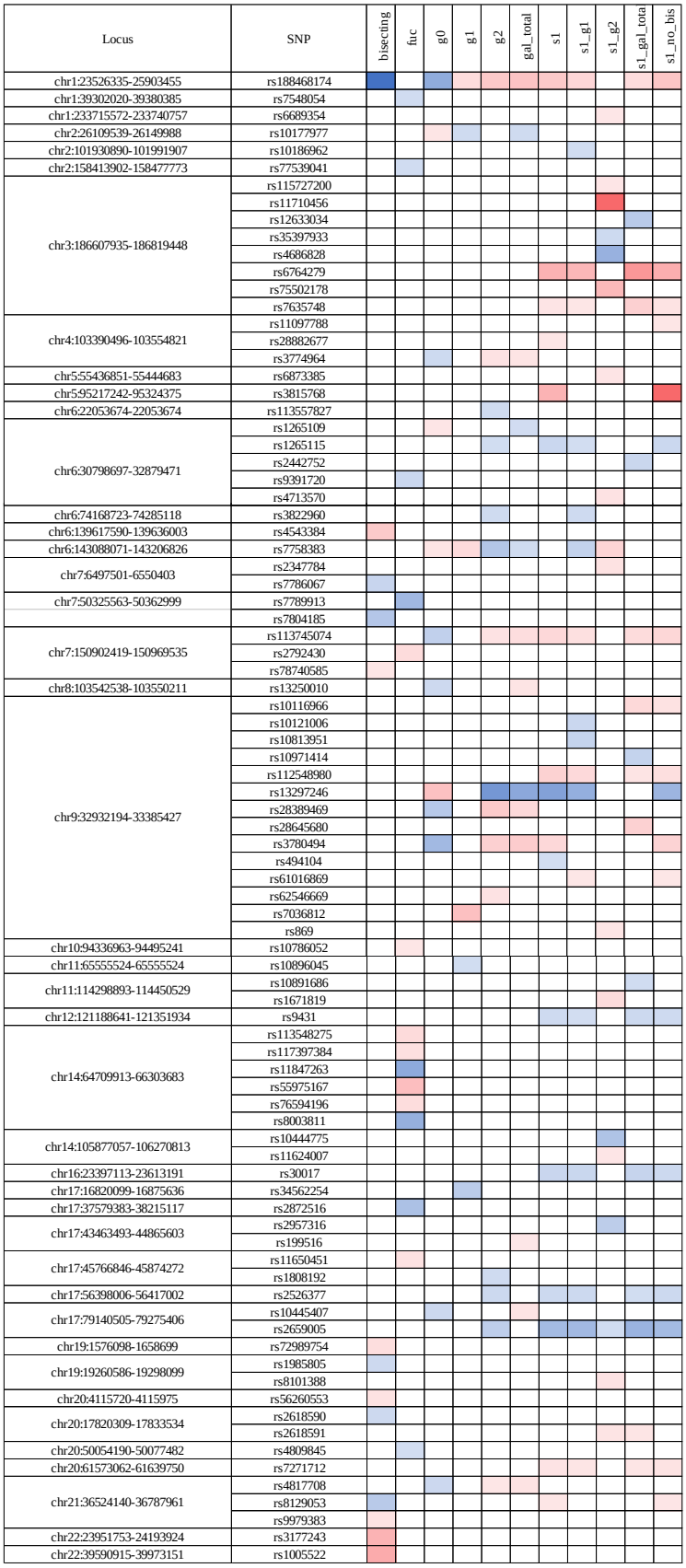


Figure 12: Joint analysis of the secondary signals. Colours represent the joint effect direction of each independent SNP; the effect is represented by the Z score which is calculated by dividing the effect estimate in the joint analysis by the standard error of the effect estimated in the joint analysis (Z score values range from -36 to 24)

COJO analysis was conducted to detect independent secondary signals associated with glycan traits and estimate their joint effects. The approximate method was applied to GWA meta-analysis summary statistics for each trait and LD for the SNPs was estimated using a subsample of unrelated UK Biobank participants. Ten out of eleven traits had the secondary signal detected in at least one associated genomic region. A total of five regions contained secondary signals with two to six independent variants being associated with a single trait.

In total, sixteen loci contain more than one variant associated with glycan traits. Different glycan traits were in stronger association with different variants in the same locus but some of the variants are in high LD.

Largest number of independent associations (n=6) in a single locus was identified in GWAS for fucosylation trait in region harbouring *FUT8* gene (chr14:64709913-66303683). The direction of the effect was negative for two of the SNPs (rs11847263, rs8003811) as opposed to positive effect direction for four remaining SNPs (rs113548275, rs117397384, rs55975167 and rs76594196). The *ST6GAL1* locus contained varying number of independent associations across different traits: s1_g2 (n=5), s1_gal_total (n=3), s1 (n=2), s1_g1 (n=2) and s1_no_bis (n=2). Locus on chromosome 9 which harbours *B4GALT1* gene had the highest number of trait associations with multiple independent associated SNPs: s1_g1 (n=5), s1_no_bis (n=5), g2 (n=4), s1 (n=4), s1_gal_total (n=4), g0 (n=3) and gal_total (n=3). Additional two regions contained multiple independent associations: chr14:105877057-106270813 (s1_g2; rs10444775 and rs11624007) and chr21:36524140-36787961 (bisecting; rs8129053 and rs9979383).

The observed difference in the joint analysis was due to the overestimation and underestimation of SNP effects in the single-SNP analysis. *FUT8* locus contains six independent SNPs which are associated with the fucosylation trait. The joint estimates for these six SNPs can be seen in Table 9. Upon jointly modelling the effects of all six SNPs, there was an increase in power for rs76594196 ($p_{single\text{-}SNP} = 1.83\times10^{-6}$ to $p_{joint} = 1.71\times10^{-15}$) and only a slight increase for rs113548275 ($p_{single\text{-}SNP} = 5.13\times10^{-17}$ to $p_{joint} = 1.34\times10^{-17}$), while the effects of other SNPs in the region were overestimated in the single-SNP analysis including the lead SNP rs11847263 where $p_{single\text{-}SNP} = 2.2 \times 10^{-118}$ decreased to $p_{joint} = 7.33 \times 10^{-47}$.

Table 9: Joint analysis of SNPs in *FUT8* locus and fucosylation trait.

SNP	EA	freq	b	se	p	bJ	bJ_se	pJ	LD_r
rs113548275	T	0.9640	-0.3003	0.0354	5.13E-17	-0.3197	0.0370	1.34E-17	0.2041
rs11847263	T	0.6318	0.2879	0.0123	2.20E-118	0.2060	0.0142	7.33E-47	-0.2193
rs55975167	T	0.0347	-0.6445	0.0361	1.13E-69	-0.5881	0.0382	2.33E-52	-0.0319
rs76594196	A	0.0204	-0.2443	0.0506	1.83E-06	-0.4112	0.0510	1.71E-15	-0.0178
rs117397384	T	0.0308	-0.4880	0.0436	1.93E-28	-0.3264	0.0444	3.70E-13	-0.1479
rs8003811	T	0.6031	0.2407	0.0122	1.14E-84	0.1846	0.0136	6.51E-41	0.0000

SNP-variant rsID; EA-effect allele; freq-frequency of effect allele in the discovery single-SNP meta-analysis; b-effect estimate in the discovery single-SNP meta-analysis; se- standard error of the effect estimate in the meta-analysis; p- P-value of the single-SNP analysis; bJ, bj_se and pJ- effect, standard error and P-value in the joint analysis; LD_r-correlation of the SNP with the SNP in the next row.

For *ST6GAL1* locus (Table 10), there were 3 glycan traits (s1, s1_g1 and s1_no_bis) with the same two independent SNP associations: rs6764279 and rs7635748. In all three cases, there was an overestimation of the SNP effects in the single-SNP analysis. The glycan trait describing the ratio of monosialylated and galactosylated structures, s1_gal_total, has three independent associations in *ST6GAL1* locus, while the s1_g2 trait has five independent SNP associations. The only SNP with underestimated effect in the single-SNP analysis was rs75502178 where the negative effect estimate increased from -0.4697 to -0.6228, hence the P-value decreased from $p_{single\text{-}SNP}= 1.44 \times 10^{-38}$ to $p_{joint}=1.59 \times 10^{-62}$.

Table 10: Joint analysis of SNPs in *ST6GAL1* locus and fucosylation trait. SNP-variant rsID; EA-effect allele; freq-frequency of effect allele in the discovery single-SNP meta-analysis; b-effect estimate in the discovery single-SNP meta-analysis; se- standard error of the effect estimate in the meta-analysis; p- P-value of the single-SNP analysis; bJ, bj_se and pJ- effect, standard error and P-value in the joint analysis; LD_r-correlation of the SNP with the SNP in the next row.

trait	SNP	EA	freq	b	se	p	bJ	bJ_se	pJ	LD_r
s1	rs7635748	A	0.0694	-0.1711	0.0229	1.47E-13	-0.1455	0.0230	3.89E-10	0.0703
	rs6764279	A	0.2790	-0.2106	0.0110	5.81E-80	-0.2064	0.0111	6.28E-75	0.0000
s1_g1	rs7635748	A	0.0703	-0.1699	0.0243	3.83E-12	-0.1445	0.0244	4.00E-09	0.0703
	rs6764279	A	0.2823	-0.2091	0.0117	1.80E-70	-0.2049	0.0118	3.35E-66	0.0000
s1_g2	rs35397933	A	0.0953	-0.1435	0.0212	1.83E-11	0.1442	0.0224	1.70E-10	-0.0481
	rs75502178	T	0.0357	-0.4697	0.0359	1.44E-38	-0.6228	0.0370	1.59E-62	-0.0521
	rs11710456	A	0.2823	-0.5365	0.0119	1.44E-444	-0.5085	0.0141	7.38E-281	-0.3399
	rs4686828	T	0.2219	0.3461	0.0135	7.31E-143	0.2021	0.0152	5.14E-40	-0.1807
	rs115727200	A	0.9864	-0.5662	0.0551	1.98E-24	-0.3398	0.0562	1.97E-09	0.0000
s1_gal_total	rs7635748	A	0.0703	-0.3145	0.0261	7.18E-33	-0.2959	0.0265	1.80E-28	0.0703
	rs6764279	A	0.2817	-0.3873	0.0123	9.11E-214	-0.3375	0.0135	3.39E-135	-0.3390
	rs12633034	C	0.2218	0.2437	0.0137	1.49E-69	0.1340	0.0148	2.82E-19	0.0000
s1_no_bis	rs7635748	A	0.0694	-0.1782	0.0227	7.63E-15	-0.1517	0.0228	4.39E-11	0.0703
	rs6764279	A	0.2789	-0.2178	0.0109	3.88E-87	-0.2134	0.0111	1.88E-81	0.0000

Region on chromosome 9 which harbours *B4GALT*1 gene has the highest number of multi-SNP associations (n=7) with additional two single-SNP associations with g1 and s1_g2 traits. The power gain in joint analysis for three SNPs associated with g0 and gal_total increased in joint analysis, thus decreasing the P-values. Four-SNP association was identified for g2, s1 and s1_gal_total traits, however, the SNP sets differed between the traits. Five-SNP associations were detected with monosialylation phenotypes, s1_no_bis and s1_g1, with a differing set of SNPs that are not in high LD. Low LD values (derived from 1000G EUR reference data) might indicate differing causal variants for monosialylation (s1, s1_gal_total, s1_no_bis), monogalactosylation and digalactosylation phenotypes as different genetic variation might be causal for changes in IgG glycosylation pathway: addition of single galactose, the addition of second galactose unit or addition of two galactose units as a prerequisite for monosialylation. Both overestimation and underestimation of SNP effects in the single-SNP analysis as opposed to joint analysis can be seen in Table 11.

Table 11: Joint analysis of SNPs in *B4GALT1* locus and fucosylation trait. SNP-variant rsID; EA-effect allele; freq-frequency of effect allele in the discovery single-SNP meta-analysis; b-effect estimate in the discovery single-SNP meta-analysis; se- standard error of the effect estimate in the meta-analysis; p- P-value of the single-SNP analysis; freq_J- frequency of the effect allele in the reference sample; bJ, bj_se and pJ- effect, standard error and P-value in the joint analysis; LD_r-correlation of the SNP with the SNP in the next row.

trait	SNP	EA	freq	b	se	p	bJ	bJ_se	pJ	LD_r
s1_no_bis	rs10116966	A	0.3061	-0.1498	0.0107	1.06E-43	-0.0885	0.0127	5.11E-12	-0.2008
	rs3780494	T	0.9824	-0.3148	0.0409	2.50E-14	-0.4314	0.0417	1.29E-24	0.0502
	rs13297246	A	0.1646	0.2007	0.0134	9.17E-50	0.1936	0.0153	3.40E-36	0.2422
	rs112548980	A	0.7581	-0.1035	0.0127	7.03E-16	-0.1104	0.0146	5.98E-14	0.1100
	rs61016869	A	0.1091	-0.1353	0.0163	2.04E-16	-0.0975	0.0165	4.90E-09	0.0000
s1_gal_total	rs28645680	A	0.8574	-0.1701	0.0171	6.33E-23	-0.1935	0.0177	3.14E-27	0.1039
	rs10971414	T	0.0702	0.1757	0.0226	1.31E-14	0.1806	0.0238	5.32E-14	-0.1813
	rs10116966	A	0.3073	-0.1806	0.0124	3.18E-47	-0.1247	0.0138	3.09E-19	0.3646
	rs112548980	A	0.7587	-0.1132	0.0148	3.46E-14	-0.1060	0.0164	1.67E-10	0.0000
s1_g2	rs869	A	0.4072	-0.0746	0.012	6.79E-10	-0.0746	0.0120	7.13E-10	0.0000
s1_g1	rs10813951	A	0.7442	0.1974	0.0121	4.99E-59	0.1022	0.0137	1.24E-13	0.2655
	rs13297246	A	0.1663	0.2216	0.0144	1.01E-52	0.2204	0.0160	1.56E-42	-0.0492
	rs10121006	T	0.0173	0.2691	0.0444	1.76E-09	0.3092	0.0449	8.25E-12	0.0660
	rs112548980	A	0.7598	-0.1259	0.0137	7.11E-20	-0.1371	0.0153	7.28E-19	0.1100
	rs61016869	A	0.1077	-0.1543	0.0177	4.85E-18	-0.1052	0.0179	5.70E-09	0.0000
s1	rs3780494	T	0.9824	-0.315	0.0413	4.58E-14	-0.3809	0.0416	1.25E-19	0.0502
	rs13297246	A	0.1646	0.1924	0.0135	4.04E-45	0.2255	0.0140	7.97E-57	0.2422
	rs112548980	A	0.7579	-0.0962	0.0128	1.06E-13	-0.1433	0.0133	1.67E-26	-0.1133
	rs494104	A	0.8889	0.1345	0.0161	1.43E-16	0.0965	0.0163	4.93E-09	0.0000
gal_total	rs3780494	T	0.9825	-0.4146	0.0381	2.56E-27	-0.4660	0.0384	1.58E-33	0.0502
	rs13297246	A	0.1645	0.1582	0.0125	2.33E-36	0.1917	0.0129	2.04E-49	0.2417
	rs28389469	A	0.7627	-0.0642	0.012	1.02E-07	-0.1143	0.0124	3.42E-20	0.0000

g2	rs3780494	T	0.9824	-0.3595	0.0382	1.21E-20	-0.4294	0.0385	2.31E-28	0.0502
	rs13297246	A	0.1645	0.1954	0.0125	5.18E-54	0.2316	0.0130	1.77E-69	0.2417
	rs28389469	A	0.7624	-0.1092	0.012	2.10E-19	-0.1584	0.0125	3.12E-36	0.1135
	rs62546669	A	0.1105	-0.1422	0.0149	3.48E-21	-0.1020	0.0151	2.35E-11	0.0000
g1	rs7036812	T	0.6923	-0.1955	0.0129	2.17E-51	-0.1955	0.0130	1.06E-50	0.0000
g0	rs3780494	T	0.9825	0.4139	0.0382	4.27E-27	0.4654	0.0385	2.81E-33	0.0502
	rs13297246	A	0.1645	-0.1578	0.0125	3.50E-36	-0.1912	0.0129	3.31E-49	0.2417
	rs28389469	A	0.7627	0.0644	0.012	9.32E-08	0.1144	0.0124	3.18E-20	0.0000

Additional two genomic regions chr14:105877057-106270813 and chr21:36524140-36787961 had two-SNP associations with s1_g2 and bisecting traits, respectively. For both SNPs in region chr21:36524140-36787961, rs8129053 and rs9979383, the effects were slightly overestimated in single-SNP analysis (rs8129053; $p_{single\text{-}SNP}$= 1.62 x 10^{-21}; p_{joint}= 2.28 x 10^{-19}; rs9979383; $p_{single\text{-}SNP}$= 3.67 x 10^{-13}; p_{joint}=4.49 x 10^{-11}). Same trend in power loss was observed for the two SNPs in chr14:105877057-106270813 region: rs11624007 ($p_{single\text{-}SNP}$=3.56 x 10^{-20}; p_{joint}= 1.85 x 10^{-9}) and rs10444775 ($p_{single\text{-}SNP}$= 3.96E-36; p_{joint}= 7.44 x 10^{-25}). Subsequently, the additional independent SNPs associated with glycan traits were considered in the calculation of joint phenotypic variance explained which ranged from 3.9% for g1 to 22.11% for s1_g2 trait (Figure 13). In ten traits the phenotypic variance explained in joint analysis increased, while in g1 trait it did not change in comparison to the estimation of variance explained from the marginal effects of top SNPs only in single-SNP analysis. The biggest increase is observed for the fucosylation trait from 8.18% to 13.77%, while the smallest increase is seen for the bisecting trait, from 13.79% to 14.08%.

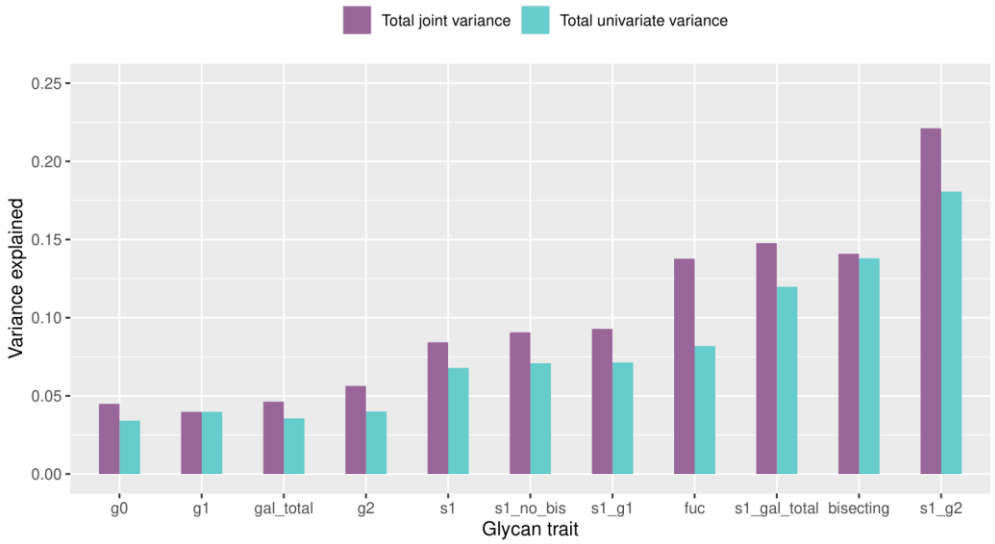


Figure 13: Phenotypic variance explained in the glycan traits by top independent SNPs in the univariate and joint analysis.

4.8 Gene mapping

The list of candidate SNPs derived by FUMA was further used to map genes in the defined genomic loci using three approaches, 1) positional mapping, 2) eQTL mapping and 3) chromatin interaction mapping. Positional mapping is used to map candidate SNPs based on their proximitiy to genes in a 10kb window. A total of 173 genes were found in 10kb windows around candidate SNPs. eQTL mapping is based on significant *cis*-eQTL associations (FDR < 0.05) for the candidate SNPs. Significant *cis*-eQTL association is present if the variation in the SNP is associated with the expression of a particular gene within 1 Mb window. A total of 82 genes were mapped based on significant eQTL association in eQTL data from DICE[129], CEDAR[132] and Fairfax[131] datasets, including data for B cells, CD4+ T cells, CD8+ T cells, neutrophils and monocytes. A subset of 37 genes was mapped using only B cell-specific eQTL data. Chromatin mapping or 3D chromatin interaction mapping is used to map SNPs to genes when there is a significant association of the SNPs in GWAS regions and nearby or distant genes via chromatin interaction. Hi-C data from the human lymphoblastoid cell line (GM12787) was used and 204 genes were mapped. The number of genes per mapping strategy is shown in Figure 14. The number of genes mapped per genomic region is shown in Figure 15 with the largest number of genes mapped in the region on chromosome 1 (chr1:23526335-25903455), followed by region on chromosome 17 (chr17:37579383-38215117) which is associated with the fucosylation trait.

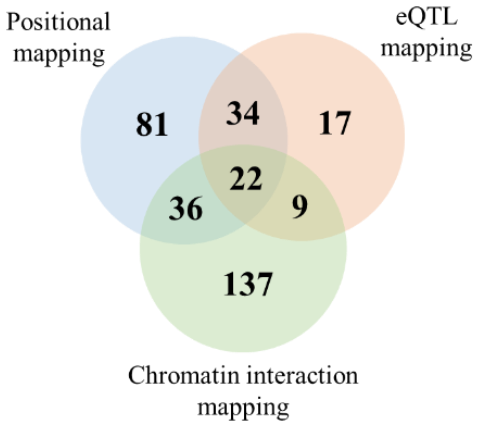


Figure 14: Number of genes mapped per mapping strategy

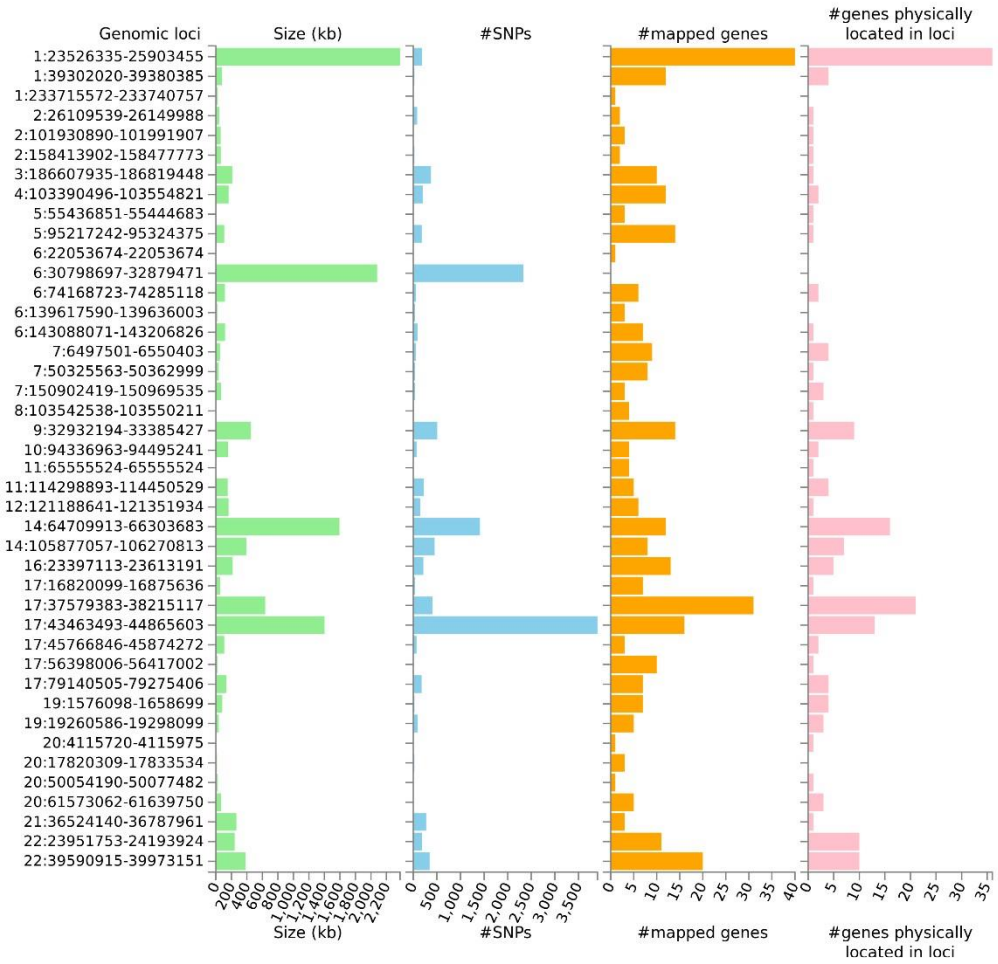


Figure 15: The histogram displays the size of genomic risk loci, the number of candidate SNPs, the number of prioritized genes and the number of genes physically located within the genomic locus.

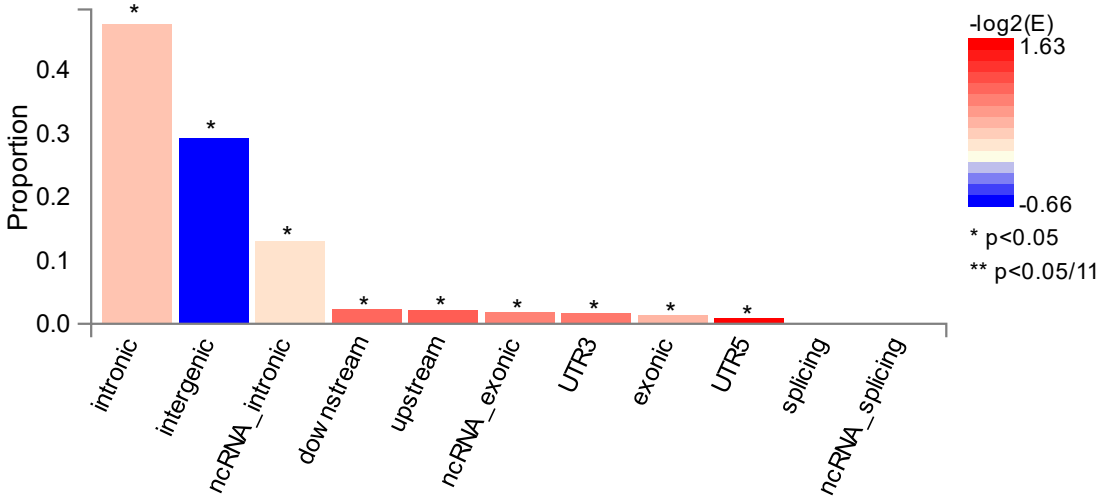


Figure 16**:** Positional annotation of candidate SNPs as assigned by ANNOVAR. Asterisk denotes a significant enrichment of the SNPs (P-value<0.05) and color of the bar denotes the log2(enrichment) relative to all SNPs in the 1000G Phase 3 EUR reference panel.

4.9 Previous genotype-phenotype associations of top independent candidate SNPs

Potential pleiotropic effects of the 42 associated loci were assessed with Phenoscanner to explore variants that were associated both with IgG N-glycosylation and other traits and diseases. The results from Phenoscanner were used to acquire a list of diseases and traits which were further tested in colocalization analysis. SNPs in 21 loci were significantly associated in previous studies ($p \leq 5 \times 10^{-8}$) with at least one trait in the PhenoScanner database. Given that the majority of the regions with pleiotropic effects are already discussed in previous IgG N-glycome GWAS study[10], the Phenoscanner results from four novel genomic regions (annotated by reference SNP) are listed in Supplementary Table 11.

4.10 Gene prioritization

Prioritization of the potential candidate genes was the next step following the identification of the genomic regions associated with IgG N-glycome. Most of the associated SNPs are found in the non-coding regions as shown in Figure 16 and cannot be directly linked to the positionally closest gene as gene regulation can be complex and both *cis* and *trans* effects are possible. Therefore, multiple gene prioritization approaches were applied: 1) genes in which glycosylation-associated SNP affected the amino-acid sequence of the resulting protein, 2) pleiotropy with gene expression in the whole blood, 3) pleiotropy with gene expression in immune cells, 4) genes in genome-wide gene-based association analysis (MAGMA) and ultimately, 5) positional mapped genes. Additionally, previous prioritization efforts by Klarić *et al.*[10] were considered, as well as the literature/known biology-based evidence for loci that code for genes with known roles in glycosylation processes, such as glycosyltransferases genes.

4.10.1 Functional consequences of candidate genetic variants

Variant effect predictor (VEP) was used to assess the functional consequences of candidate variants on genes, transcripts, protein sequence and regulatory regions using SIFT and Polyphen-2 algorithms. Besides listing the gene and transcript influenced by the variant, VEP annotated the variant by listing their location and MAF and type of consequence on the protein such as stop-gained, stop lost, missense or frameshift. The genetic variants with probability values < 0.05 were considered deleterious. Polyphen-2 predicts the damaging effect of a mutation by using eight sequence-based and three structure-based features with scores ranging from 0 (benign) to 1 (probably damaging).

A total of 12,348 candidate genetic variants were assessed. The predictions with severe coding consequences were considered including missense (n=86) and stop-lost (n=1) variants. Additionally, splice donor (n=2) and splice acceptor variants (n=1) were taken into account. A total of 77 variants were predicted as synonymous. Forty-six genes that are spread across fifteen genomic regions were predicted to be affected by the protein-coding mutations. SIFT and Polyphen2 scores were assigned to a subset of 45 variants in 30 genes which are found in ten genomic regions (Supplementary Table 8). The predicted functional consequences of the candidate variants were considered in the subsequent gene prioritization, excluding the variants from the MHC region (chr6:30798697-32879471).

Besides VEP, candidate SNPs (excluding SNPs in the MHC region) were also checked for their CADD scores via FUMA SNP2GENE tool. CADD score is a deleteriousness score based on 63 functional annotations and it was used as additional evidence for SNP functional consequence (Figure 17). At the cut-off > 15, there were 159 SNPs located across 39 genes with three SNPs with CADD score > 30.

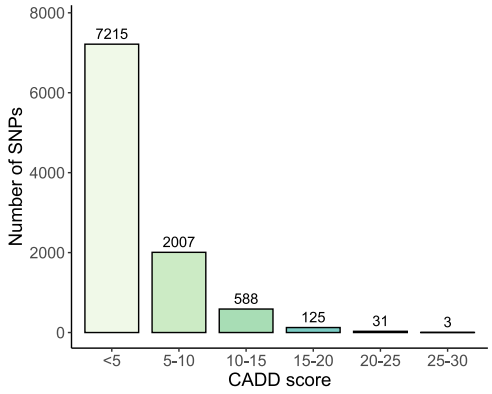


Figure 17: Distribution of CADD scores for candidate SNPs (excluding MHC region)

4.10.2 Pleiotropy with gene expression in whole blood

The summary statistics for *cis*-eQTLs in whole blood from the eQTLgen dataset were used to investigate whether the identified variants affect the expression of a nearby gene, thereby potentially affecting the N-glycosylation. The eQTLgen dataset was chosen as it is one of the biggest eQTL datasets to date derived on 31,684 blood samples, thus increasing the power to detect eQTL signals. The suggested threshold PP4 ≥ 75%[68] was used to support of eQTL/N-glycosylation colocalization hypothesis. In total, eQTLs for 21 genes in 18 genomic regions colocalized with variants from regions identified in IgG N-glycosylation GWAS. A subset of the associations was previously found in GWAS by Klarić *et al.*[10] but here all associations are

listed as the glycan phenotypes between the studies are different. Given the more straightforward derivation of traits in this study, it was possible to link the gene expression and addition of a specific sugar unit in a process of IgG N-glycosylation.

Fucosylation

On chromosome 1, there is a positive colocalization for the expression of *MYCBP* gene and fucosylation trait (PP4 = 99.4%). The associated region (chr1:39302020-39380385) is fucosylation-specific. The comparison between the regional association pattern for fucosylation and expression of *MYCBP* gene is shown in Figure 18.

Region on chromosome 2 (chr2:158413902-158477773) harbours *ACVR1C* gene for which eQTL signal colocalizes (PP4=92.8%) with association signals in the region for fucosylation trait. The SNP.PP.H4 value is the posterior probability of the variant being causal for the colocalized signal when the colocalization hypothesis (H4) is true. In this case, SNP with highest PP.H4 (rs10164853; 0.996) value is different than lead SNP in fucosylation GWAS (rs77539041).

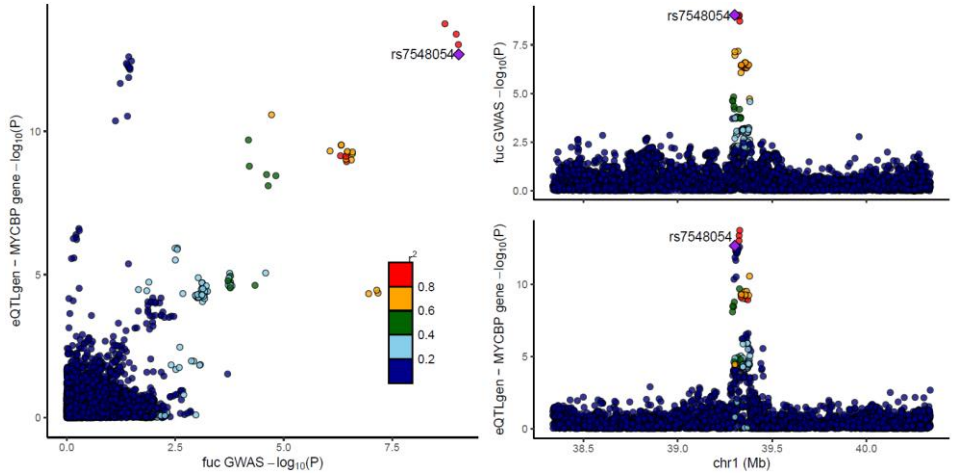


Figure 18: Regional plot of association in chr1:39302020-39380385 region with fucosylation trait (top right) and regional plot for *MYCBP* expression in whole blood in the same region (bottom right). Scatter plot of -log10(p-values) of SNPs in fucosylation GWAS and *MYCBP* expression (left).

Fucosylation GWAS signals also colocalize with eQTL signals for *KIF11* gene (PP4=99%) which is located on chromosome 10 (chr10:94336963-94495241) and *ORMDL3* gene (PP4=78.7%) in region on chromosome 17 (chr17:37579383-38215117) (Figure 19).

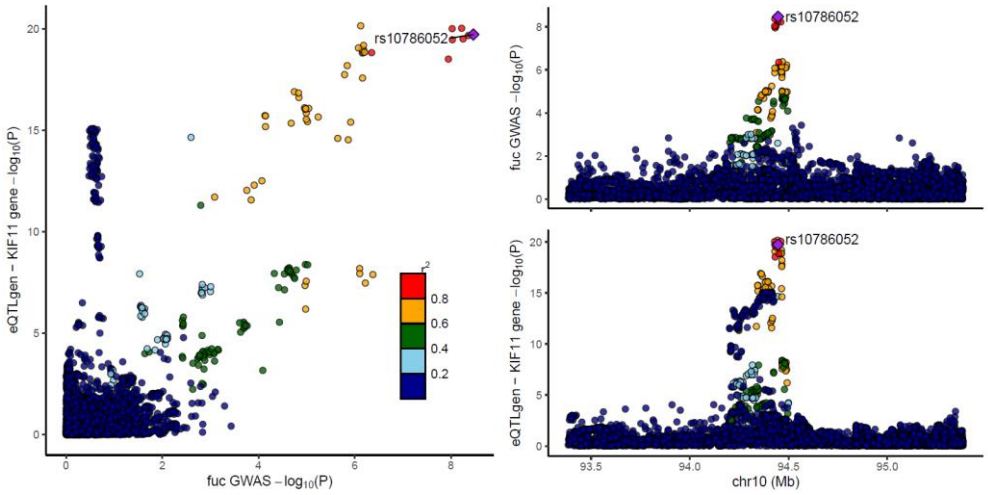


Figure 19: Regional plot of association in chr10:94336963-94495241 region with fucosylation trait (top right) and regional plot for *KIF11* expression in whole blood in the same region (bottom right). Scatter plot of -log10(p-values) of SNPs in fucosylation GWAS and *KIF11* expression (left).

Bisecting GlcNAc

There is a high posterior probability for colocalization between eQTL signals for *TCF3* gene (Figure 20) and GWAS association signals for the bisecting trait (PP4=96.2%) in a genomic region on chromosome 19 (chr19:1576098-1658699) and eQTL for *KDELR2* gene on chromosome 7 (chr7:6497501-6550403).

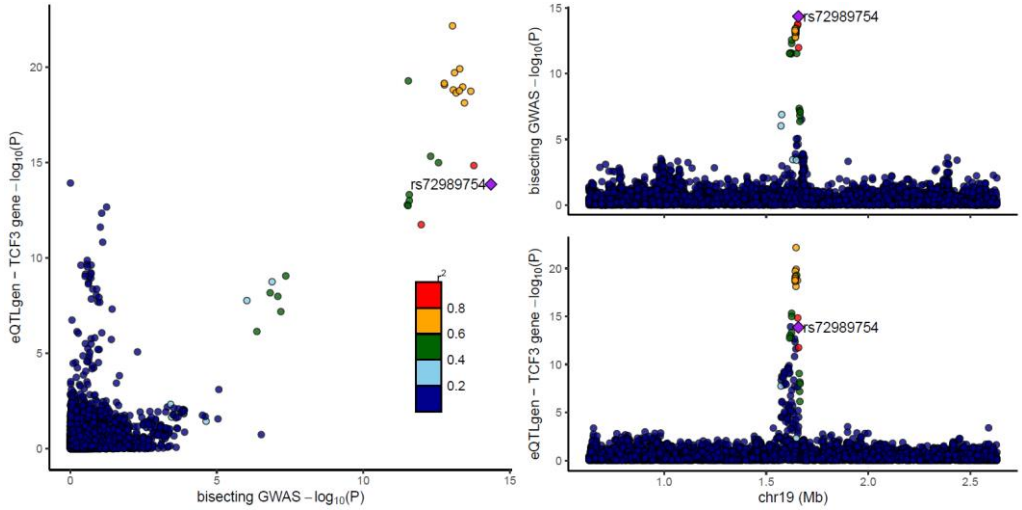


Figure 20: Regional plot of association in region chr19:1576098-1658699 with the bisecting trait (top right) and regional plot for *TCF3* expression in whole blood in the same region (bottom right). Scatter plot of -log10(p-values) of SNPs in bisecting GWAS and *TCF3* expression (left).

Positive colocalization is found for bisecting trait and expression of *MGAT3* gene (PP4 = 96.0%) pointing to a potential mechanism of control of the addition of bisecting GlcNAc to

IgG N-glycans as the same variant is potentially affecting the level of IgG N-glycan structures with bisecting GlcNAc and levels of *MGAT3* gene expression in whole blood (Figure 21).

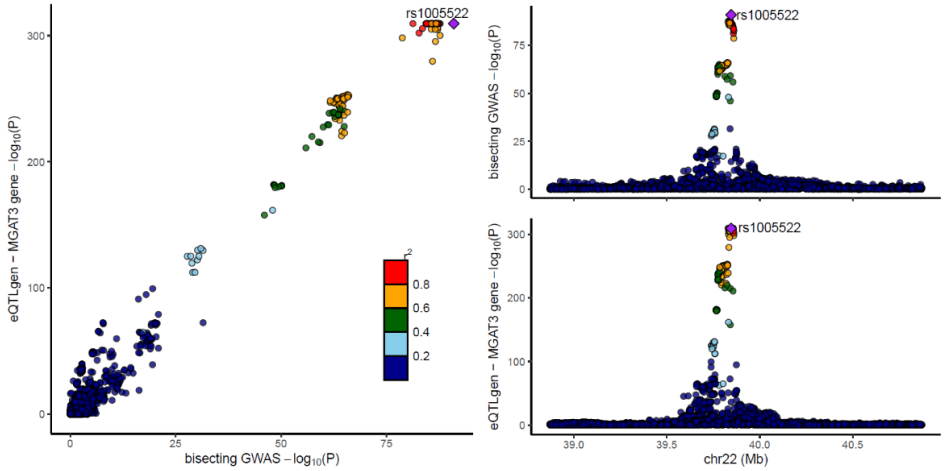


Figure 21: Regional plot of association in *MGAT3* locus with the bisecting trait (top right) and regional plot for *MGAT3* expression (eQTLs) in whole blood in the same region (bottom right). Scatter plot of -log10(p-values) of SNPs in bisecting GWAS and *MGAT3* expression (left).

Sialylation and galactosylation

Given that sialylation is conditioned on the presence of galactose on the IgG N-glycan structure, multiple regions were associated with both galactosylation and sialylation, thus there was a high posterior probability for colocalization for both traits with eQTLs in whole blood. Three genes had high PP4 including, *EEF1A1* (s1_g1; PP4=96.5%), *MTO1* (s1_g1; PP4=96%), and *NFKB1* (s1_no_bis; PP4=93%). Both *EEF1A1* and *MTO1* genes are found in the same region on chromosome 6 (chr6:74168723-74285118).

A total of six genes contain eQTLs in whole blood which colocalize with GWAS signals for sialylation-specific traits, including *ANKRD55* (s1_g2; PP4=97.5%), *COG7* (s1_no_bis; PP4=78.9%; Figure 22), *DCTN5* (s1_g1; PP4=84.4%), *ELL2* (s1_no_bis; PP4=93.2%; Figure 23), *IGHG2* (s1_g2; PP4=98.3%), *MEF2B* (s1_g2; PP4=97.3%) and *IL6ST* (s1_g2; PP4=95.7%), where trait names denote traits with highest PP4. The region chr16:23397113-23613191 harbours two of the colocalized signals, eQTLs for *DCTN5* and *COG7*, as well as the region on chromosome 5 (chr5:55436851-55444683) which harbours *IL6ST* and *ANKRD55* genes. The only galactosylation-specific region (chr2:26109539-26149988) with positive colocalization signal contains *KIF3C* gene with PP4=90% (g1). Additionally, eQTL signals for *hsa-mir-142* micro RNA colocalized with association signals for digalactosylation and monosialylation phenotypes with PP4=99.8% in locus on chromosome 17.

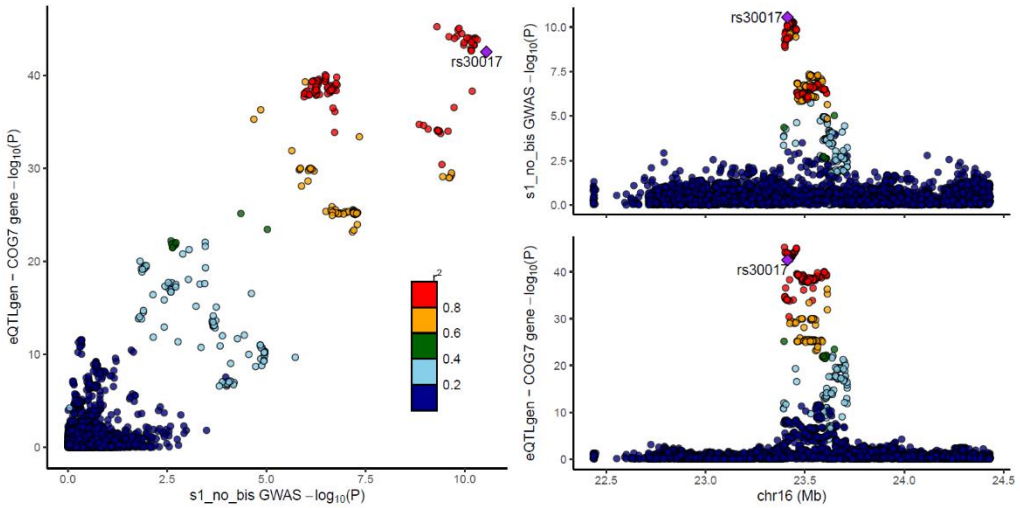


Figure 22: Regional plot of association in chr16:23397113-23613191 region with s1_no_bis trait (top right) and regional plot for *COG7* expression (eQTLs) in whole blood in the same region (bottom right). Scatter plot of -log10(p-values) of SNPs in s1_no_bis GWAS and *COG7* expression (left).

In the genomic regions where other glycosyltransferase genes (*FUT8*, *ST6GAL1* and *B4GALT1*) are found, the conditional analysis indicated more than one causal variant in the region, thus not meeting the single causal variant assumption to compute the posterior probability for colocalization of signals (PP4). Low PP4, in that case, would not indicate a lack of colocalization as the algorithm considers only the strongest association and the rest of the independent signals are not considered. Also, according to the conditional analysis, the region which harbours *IGHG2* gene harbours multiple independent associations, hence, the positive colocalization test does not exclude other candidate genes from the region. The positive colocalization was not taken as exclusive evidence for prioritization of the gene since it was performed using eQTL data for whole blood, instead, it was taken as one of the criteria to prioritize a gene.

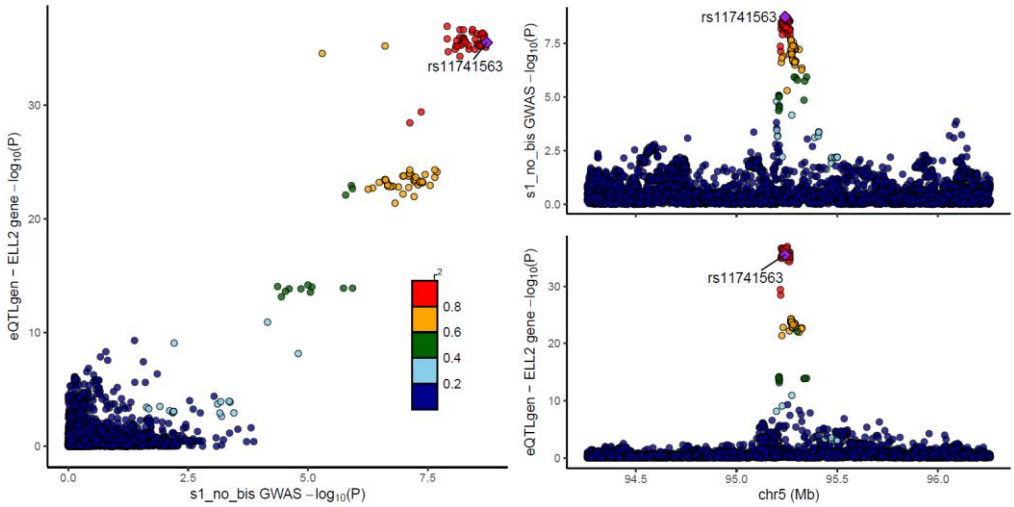


Figure 23: Regional plot of association in chr5:95217242-95324375 region with s1_no_bis trait (top right) and regional plot for *ELL2* expression (eQTLs) in whole blood in the same region (bottom right). Scatter plot of -log10(p-values) of SNPs in s1_no_bis GWAS and *ELL2* expression (left).

In total, 83 genes were prioritized across 42 genomic regions (Supplementary Table 8). The genes are shown in Figure 24 with evidence for prioritization besides each gene symbol. A subset of 22 regions have only one prioritized gene, 9 regions contain two prioritized genes, 5 regions contain three and the remaining regions have four or more candidate genes. In genomic regions containing glycosyltransferases genes -*ST6GAL1*, *B4GALT1*, *FUT8*, *MGAT3*- the prior knowledge of their function is taken as the main evidence for the prioritization of these genes, as well as the prioritization efforts by the previous GWA studies which were further supported by the evidence in the current study.

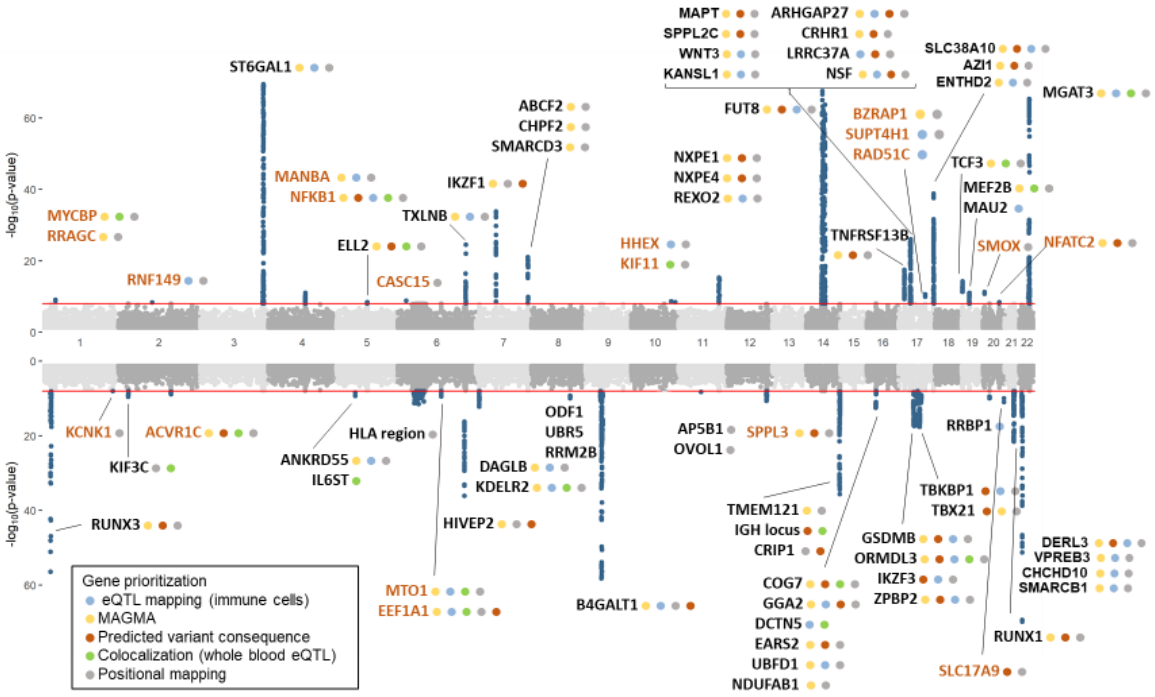


Figure 24: Manhattan plot of genome-wide significant associations in IgG N-glycome GWAS with prioritized genes in each locus. Plot shows -log10(P-value) of association on y-axis and SNPs ordered by chromosomal location on x-axis. Red line indicates the genome-wide significance threshold (1×10^{-8}). For simplicity, the y axis is trimmed at -log10(P-value)=60. Orange gene names indicate novel loci associated with IgG N-glycosylation.

4.11 Gene-set enrichment analysis

A gene-set enrichment test was performed using FUMA GENE2FUNC[69] option to test for overrepresentation of the 83 prioritized genes from IgG N-glycome GWAS in gene sets obtained from MSigDB c5[149] using hypergeometric tests. The candidate genes were tested against Canonical pathways and Gene Ontology (GO) terms. A total of 123 gene sets with an FDR < 0.05 were identified (Supplementary Figure 5). The results were further filtered based on keywords to identify gene sets that describe higher-level biological processes. Among the identified 108 GO sets, six gene sets were related to glycosylation process, three gene sets to ER-Golgi transport, 26 were related to B and T cell activation and proliferation, nine to immune response in general and seven were related to transcription and gene expression.

Fifteen gene sets were enriched in Canonical pathways, four of which were glycosylation-related, six were related to ER-Golgi transport, one to T cell development and one gene set was related to transcriptional regulation.

4.12 Functional network of IgG N-glycome associated loci

The functional network of genomic loci identified in the IgG N-glycome GWAS was constructed based on the summary statistics for the top SNPs across all 42 genomic regions and their correlation. The nodes represent the top SNP but are denoted by one of the prioritized genes in the locus, while the edges represent the squared Spearman's correlation of their effects on glycan traits (Z-scores). Only the correlations which passed the corrected p-value threshold ($0.05/((42*41)/2)= 5.8 \times 10^{-5}$) were taken into account when constructing the network. The sign of the correlation between SNP effects depends on the tested reference allele, therefore here it is not considered informative. Given the strict significance threshold, only 28 nodes and 26 edges were constructed in the network which is shown in Figure 25.

There are seven separate clusters in the network formed which can partially be explained by the low number of traits or Z-score values which enter the correlation analysis (n=11) thereby obtaining non-significant correlation even when the Spearman's correlation coefficient is high. There are two two-loci clusters: *KIF3C-OVOL1/AP5B1* and *HLA region-RRBP1*. Two three-loci clusters are comprised of *TCF3-TMEM121*/*IGHlocus/CRIP1-TXLNB* and *MEF2B/MAU2-NXPE1/NXPE4-ST6GAL1*.

The significant correlation values ranged from 0.927 to 0.981. The strongest correlation (Spearman's ρ=0.981; $p=8.4 \times 10^{-8}$) of SNP effects was observed for SNPs found in *RUNX3* (rs188468174; chr1:23526335-25903455) and *RUNX1* (rs8129053; chr21:36524140-36787961) loci. The *RUNX3* and *RUNX1* loci are associated with all trait categories except fucosylation. Among four glycosyltransferases, only *ST6GAL1* and *B4GALT1* were included in the network. The lead SNP in *ST6GAL1* locus, rs11710456, had a strong correlation (Spearman's ρ=0.927; $p=3.97 \times 10^{-5}$) with effects of lead SNP in *NXPE1-NXPE4* locus, rs1671819. The lead SNP in *B4GALT1* locus, rs12342831, was strongly correlated (Spearman's ρ=0.927; $p=3.97 \times 10^{-5}$) with top SNP in *SLC38A10-AZI1-ENTHD2* locus, rs2659005. It is important to note that majority of the nodes included are limited to monosialylation- and galactosylation-specific loci because there were multiple traits defined for both trait categories, thereby causing inflation in the correlation values as opposed to fucosylation and bisecting phenotypes which were limited to one derived trait.

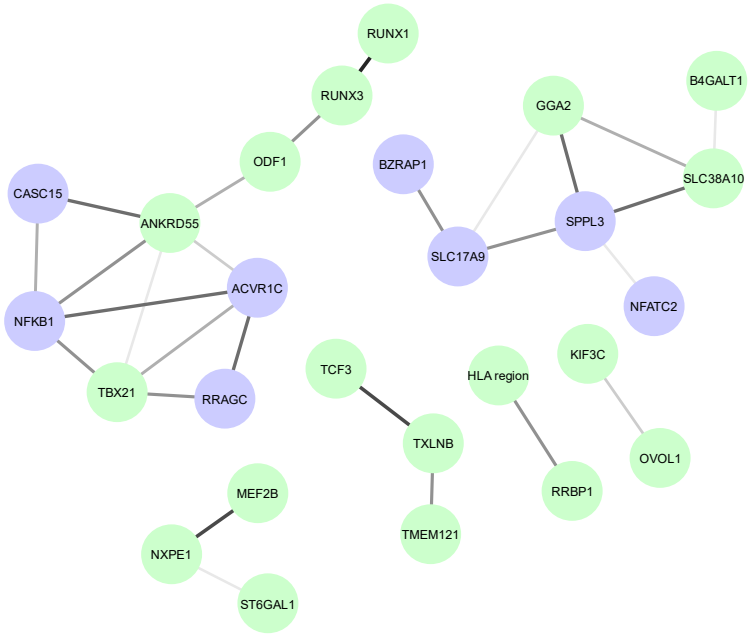


Figure 25: Functional network of the genomic loci identified in IgG N-glycome GWAS as represented by one of the genes prioritized in the locus. Purple nodes represent novel loci identified in the GWAS. The color of the edge represents the squared correlation value of the top SNP effects (Z-scores) across glycan traits which ranges from 0.85 to 0.96.

4.13 STRING protein-protein interaction network

String-PPI[151] database was used to construct a network based on the 83 prioritized genes (Figure 26). To obtain the score for each interaction, evidence was limited to experimentally determined evidence, evidence from curated databases, co-expression and protein homology. The scores range from 0 to 1, where 1 represents strong evidence. In total, 27 significant interactions (FDR < 0.05) are present between the proteins which form seven clusters in the network with average node degree=0.692 and PPI enrichment p-value=7.24 x 10^{-6} (Supplementary Table 10). The glycosyltransferases (MGAT3, B4GALT1, ST6GAL1, FUT8) form a separate cluster together with MANBA (Beta mannosidase) enzyme which is known for its role in glycosylation processes. The functional association between glycosyltransferases is supported by information obtained from curated databases (combined score=0.9) while the association between B4GALT1 and MANBA is further supported by the evidence for their co-expression (combined score=0.902). One of the clusters contains proteins which are mainly involved in the transport of proteins from ER to Golgi and normal function of Golgi complex (KIF11, KIF3C, KDELR2, DCTN5, NSF, COG7), with a combined score for interaction ranging from 0.9 to 0.904 as obtained from curated databases, co-expression and protein homology. The biggest cluster in the network consists of eight proteins (MEF2B, NFATC2,

SMARCB1, SMARCD3, RUNX1, RUNX3, TBX21 and TCF3) functionally enriched in processes such as chromatin remodelling, transcriptional activation and regulation, and B cell activation and proliferation. The associations between the proteins in the cluster are all supported by the data from curated databases and 7 out of 9 interactions were experimentally determined and three interactions were due to the co-expression of the proteins (RUNX1 and RUNX3; RUNX3 and TBX21; SMARCB1 and SMARCD3; NFATC2 and RUNX1). The combined score in the cluster ranged from 0.6 to 0.992. The remaining four clusters are each formed by two proteins: NXPE1 and NXPE4 (score=0.555), IKZF3 and IKZF1 (score=0.802), ELL2 and SUPT4H1 (score=0.9) and UBR5 and EEF1A1 (score=0.425).

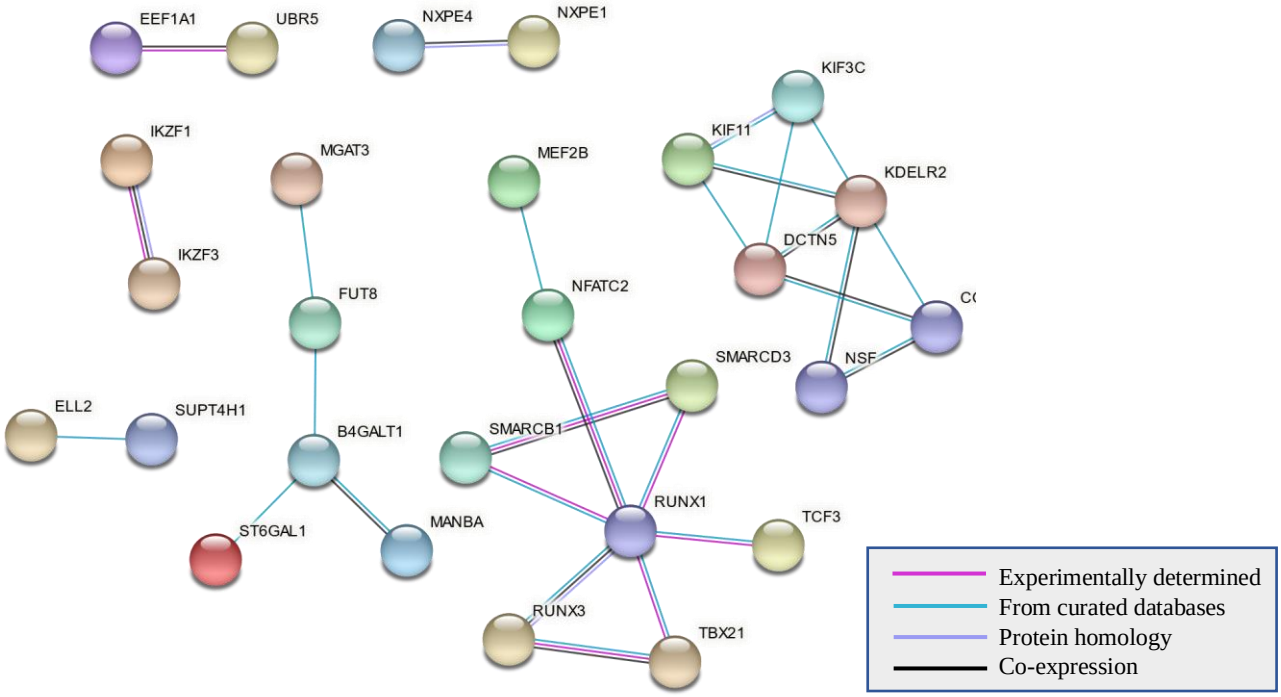


Figure 26: Interaction network of candidate genes obtained from STRING protein-protein interaction database. Nodes represent a protein encoded by the gene and edges are derived based on the evidence from various sources (See the grey box).

4.14 Pleiotropy with complex diseases and traits

Based on the results of Phenoscanner and indicated a potential role of IgG N-glycosylation in diseases, the colocalization analysis was performed to test the overlap in the regional association patterns between disease and IgG N-glycosylation, where the positive colocalization would suggest that the two traits share the same underlying causal variant. The

results are based on a colocalization test between disease and glycan trait which had the strongest association in the given locus (Supplementary Table 12).

The positive colocalization of glycan traits and disease with high confidence (PP4 > 75%) was found in 10 genomic regions where SNPs meet the suggestive significance threshold in the disease or trait GWAS ($p < 1 \times 10^{-5}$). Colocalized signal was obtained for the following diseases: RA, juvenile idiopathic arthritis (JIA), asthma, osteoarthritis, allergy, type 2 diabetes (T2D), adult-onset asthma (AOA), SLE, UC, IBD, schizophrenia, CD, PBC, Alzheimer's disease, breast cancer, as well as, a trait describing the percentage of lymphocytes in white blood cells (WBC). Additional 4 genomic regions had the positive colocalization result (PP4 > 75%), but it is important to note that the SNPs used in the analysis did not reach the suggestive significance threshold (1×10^{-5}) in disease GWAS. The list of loci with pleiotropic effect on glycosylation and disease or trait is illustrated in Figure 27.

Positive colocalization with lower confidence (50% < PP4 < 75%) was observed in 5 genomic regions associated with glycan traits and diseases including SLE, asthma, osteoarthritis, allergy, schizophrenia, breast cancer, as well as the percentage of lymphocytes in WBC. Additionally, 14 genomic regions had the posterior probability for colocalization > 50%, however, the suggestive significance threshold in GWAS for the disease was not reached.

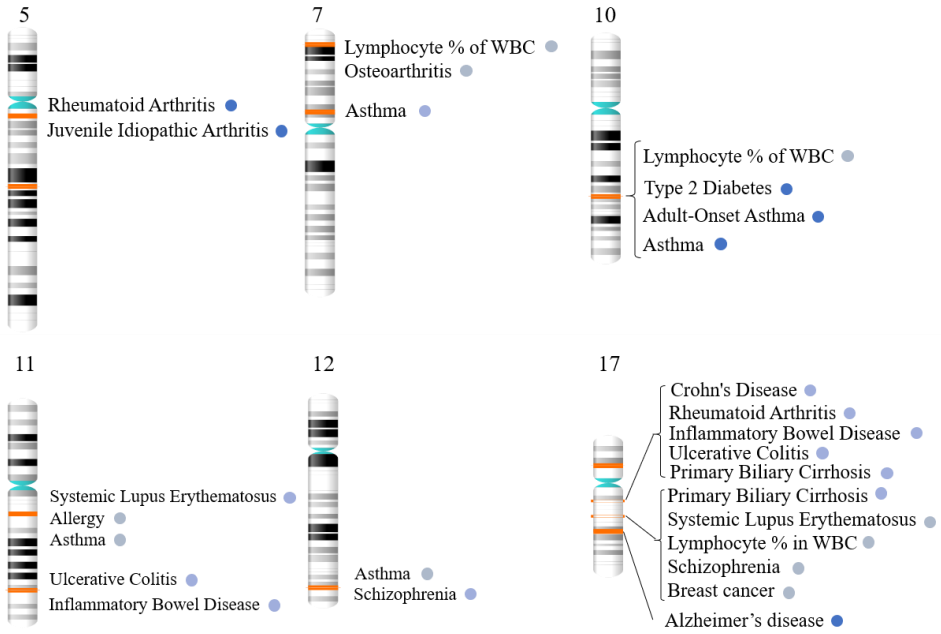


Figure 27**:** Chromosomal positions of positive colocalizations (PP4 > 50%) with IgG N-glycan traits where SNPs were significantly associated in both IgG N-glycome and disease GWAS (1×10^{-5}). Orange lines on chromosomes denote chromosomal region; dots indicate the level of probability for colocalization: gray 50% < PP4 < 75%, light blue 75% < PP4 < 95% and dark blue PP4 > 95%.

4.15 Enrichment in cell-type-specific regulatory regions

FORGE2 was used to investigate cell type-specific enrichment of top SNPs for overlap in DNase I hypersensitivity sites as derived by ENCODE and Roadmap Epigenomics Projects. Additionally, enrichment in 15 cell-type-specific chromatin states from Roadmap Epigenomics Project was assessed. Using ENCODE data, the strongest enrichment (q>0.01) was found across 16 immune cell types originating from blood tissue including: GM12865 (B lymphocytes; q= 5.7 x 10^{-8}), GM12864 (B lymphocytes; q=3.47 x 10^{-6}), Th1 (T helper type 1; q=1.73 x 10^{-5}), CD14+ cells (q=2.01 x 10^{-4}), GM06990 (B lymphocytes; q=2.01 x 10^{-4}) and GM12878 (B lymphocytes; q= 2.01 x 10^{-4}). These results were concordant with the results obtained using Roadmap Epigenomics Project data, where the significant enrichment (q < 0.05) was found for Primary T cells from cord blood, primary B cells, monocytes and natural killer cells from peripheral blood, while the strongest enrichment was found in fetal thymus cells (q < 0.006) (Figure 29). In addition, significant enrichment for 15 chromatin states in a wide range of cell types was identified, including strong enrichment for enhancers in B cells from the cord and peripheral blood, as well as weak transcription in T helper cells and natural killer cells. The snapshot of blood cell type-specific enrichment for chromatin states is shown in Figure 28.

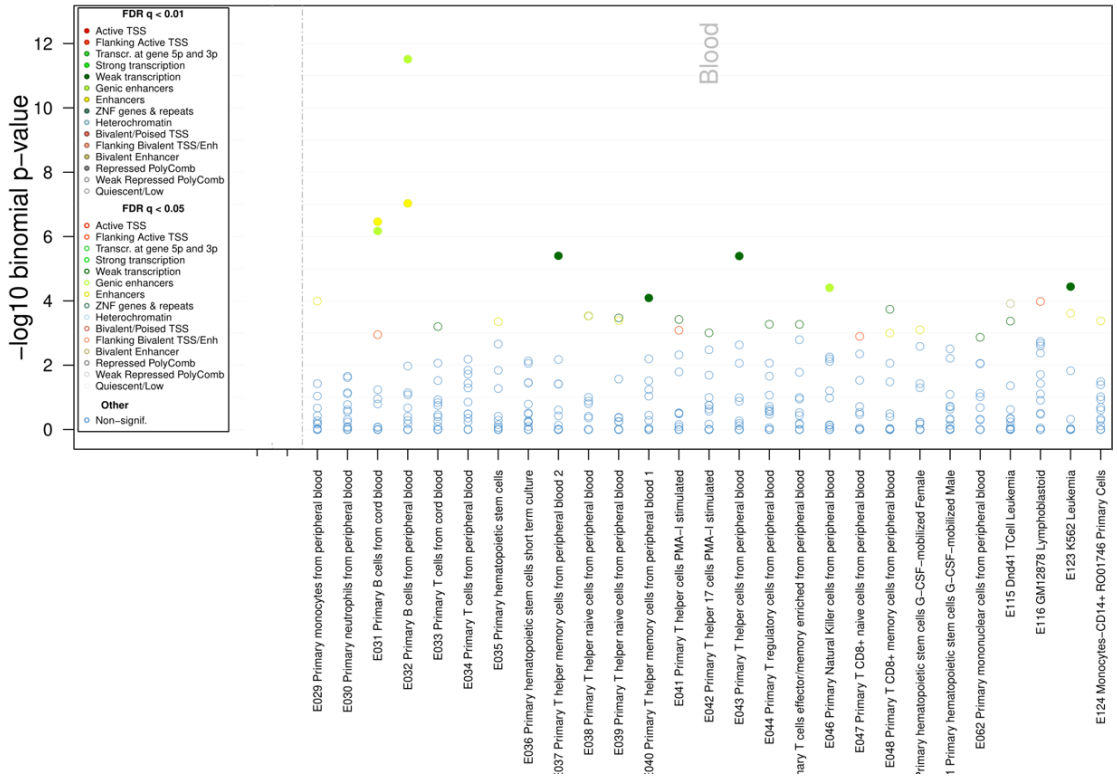
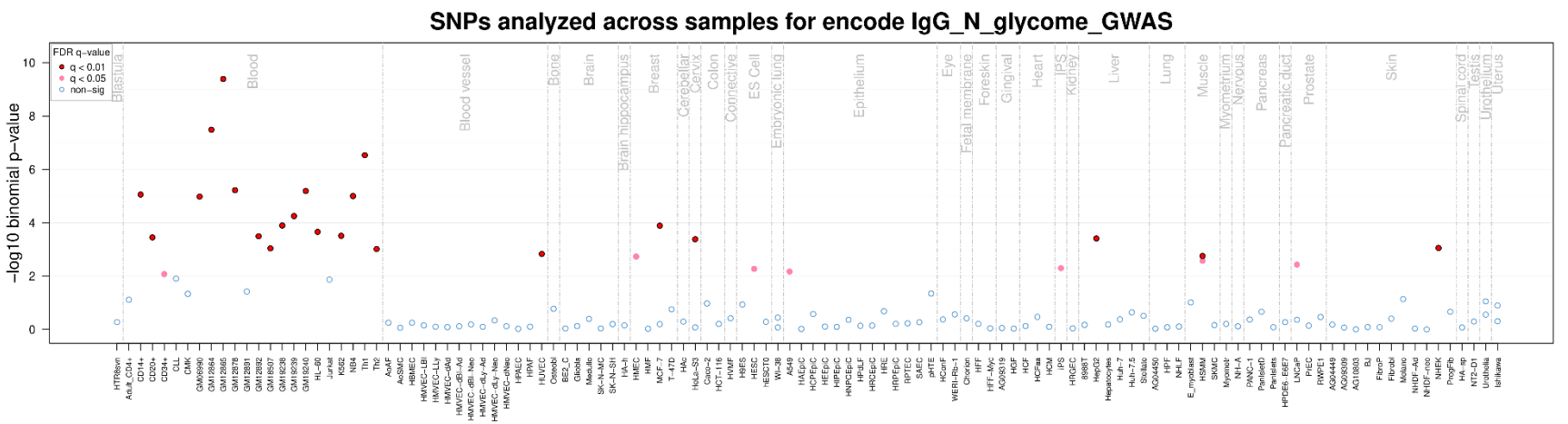


Figure 28: Blood cell-specific enrichment for regulatory elements (15 chromatin states)

A



B

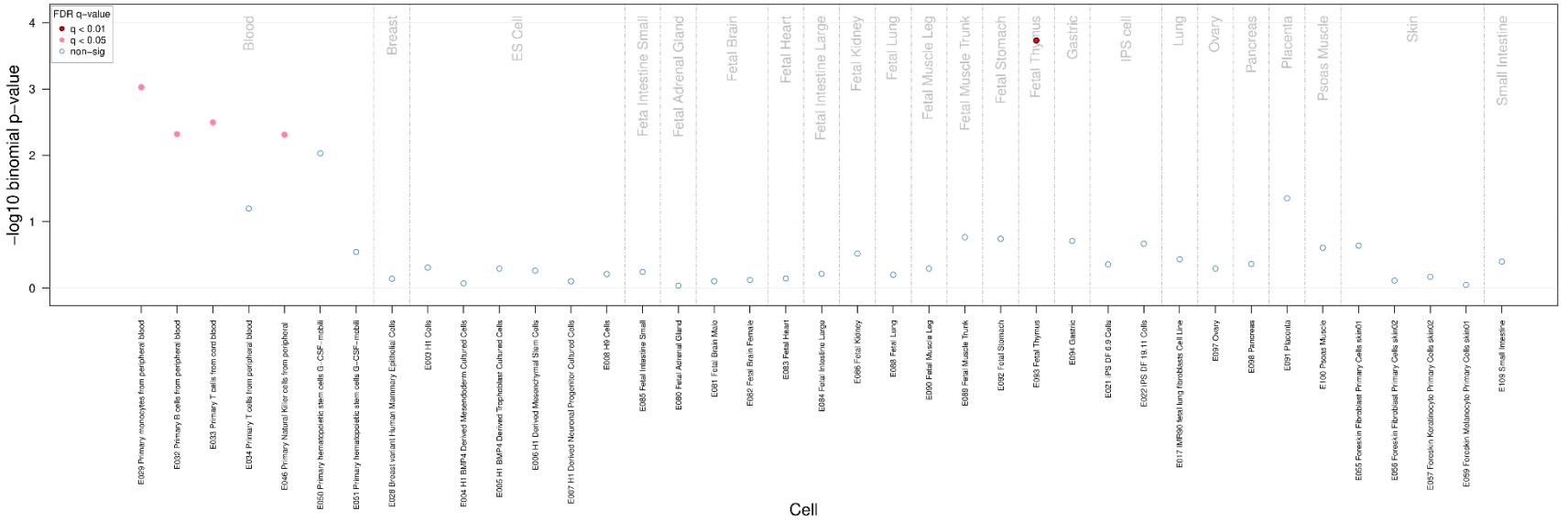


Figure 29**:** Enrichment of IgG N-glycome GWAS SNPs in DNase I hypersensitivity sites (DHS) across different cell types. A) ENCODE derived DHS data is used as background B) Roadmap Epigenomics Project derived DHS data is used as background

5. DISCUSSION

Here we conduct the largest GWAS of IgG N-glycome focusing on the glycan phenotypes describing the percentage of specific sugar units in the total IgG N-glycome. The main motivation lies in a more straightforward interpretation of the findings as it enables a link between genes and variable levels of four sugars, galactose, sialic acid, fucose and bisecting GlcNAc, all of which specifically alter the affinity of IgG for Fc receptors found on innate immune cells[154].

Meta-analysis of summary statistics from GWAS of IgG N-glycome in seven cohorts of European descent resulted in 42 genome-wide significant loci ($p < 1 \times 10^{-8}$) associated with the tested IgG N-glycome phenotypes.

The detected heterogeneity among effect estimates in the studied cohorts ranged from none to moderate. The reasons for heterogeneity can stem from the differences in genetic architecture among the studied cohorts or differences in the phenotypes due to quantification discrepancies between the cohorts. Here we also meta-analyse summary statistics from cohorts where glycan quantification was carried out using UPLC or LC-MS, which also introduces heterogeneity due to differences in trait derivation. The correlation values of traits derived from UPLC- and LC-MS-measured glycan values indicate differences in phenotypes, especially in fucosylation traits where the Pearson's correlation value has the maximum value of 0.45. However, the heterogeneity remained at a moderate level allowing the meta-analysis of GWA summary statistics for these cohorts. The heterogeneity sourced in the genetic structure might also be present since cohorts such as ORCADES[82], CROATIA-Vis[81], CROATIA-Korcula[81] and VIKING[86] are considered genetic isolates, thus causing inflated frequency for some rare alleles.

Thirteen of the observed associations are novel, meaning that they were not previously discovered in GWAS of IgG N-glycome. The traits which were tested in the previous GWAS by Klarić *et al.*[10] were either directly measured glycan traits or a set of 54 derived traits as defined in the function for calculation of derived traits *igg.uplc.derived.traits.2014*() in glycanr package[101]. These traits differ from the traits in this GWAS as they describe various characteristics of IgG N-glycome but do not describe the overall percentages of structures with certain sugar units as the traits in the current GWAS. Now, we can associate genomic loci with

the addition reaction of the specific sugar unit to the glycan chain in a more straightforward manner.

In addition, the derived trait definition allowed for meta-analysis of the higher number of cohorts, thereby increasing the sample number and power to detect novel loci. The derived traits were comparable between platforms used for measurement of IgG N-glycans, even if the original measurements are not. The used genome-wide significance threshold of $1x10^{-8}$ is the threshold corrected for five principal components which describe 99% of the variation in the tested traits. This allowed for less stringent multiple testing correction given that the tested traits are highly correlated and correction directly on the number of traits would result in an overly conservative significance threshold and a possibility to omit true positives.

The previous GWAS of IgG N-glycome[7,8,71,76] were all performed on the subset of samples from the current GWAS, thus cannot be considered as a true replication of the findings. However, the replication of genomic loci indicates the robustness of the GWAS findings given that the phenotype definition differs between studies.

Among 42 discovered loci, 25 can be associated with only one phenotype group describing the addition of a specific sugar unit to the glycan chain- galactosylation (n=4), fucosylation (n=6), monosialylation (n=10) and bisection (n=5). The remaining fourteen loci are associated with more than one trait while two loci were associated with at least one trait from each phenotype set indicating the overall effect on the glycosylation process. Monosialylation and galactosylation phenotypes are interconnected as galactosylation is the prerequisite for the sialylation to occur and 7 loci are associated with these two phenotype sets. The indication for association with the specific reaction in the process enables the setup of functional analysis as the hypothesis can be set to investigate the effect of the locus on specific glycosylation traits.

The locus containing *IKZF1* gene was associated with bisection and fucosylation phenotypes, which are known to be partially mutually excluding as N-glycan with a bisecting GlcNAc cannot be a substrate for some glycosyltransferases including FUT8[155], so the addition of bisecting GlcNAc suppresses the further processing and elongation of N-glycans. A positive effect of the effect allele is observed for both traits indicating that the gene might not be involved in the suppression of the fucosyltransferases activity based on the presence of bisecting GlcNAc but rather a regulation independent of the substrate.

In comparison with the discovered loci from Klarić *et al.*, the lack of replication is observed in two loci, one of which is *IRF1-IL3-SLC22A4* locus associated with IGP2, the percentage of A2 glycan in the total IgG N-glycome. This glycan property is not captured by any of the derived traits in the current study, thus providing a valid reason for the lack of replication. In *FUT6* locus, previously there was an association with GP20, a structure that remained uncharacterized at the time of the analysis, therefore was not included in the study. In the later study[156], the structure was characterized as an antennary fucosylated glycan, thus explaining the association with *FUT6*, a gene encoding a fucosyltransferase that catalyses the addition of antennary fucose to the IgG molecule[155]. Non-significant association in this locus is expected as we did not investigate the genome-wide associations with antennary fucosylation in IgG N-glycome.

We also fail to replicate three out of six novel loci from the latest multivariate GWAS of IgG N-glycome[76], which could be due to the lower power of the univariate analysis as opposed to the multivariate approach and differences in glycan structures which are included in each of the defined traits. Interestingly, we rediscover the association with *IL6ST-ANKRD55* locus from Lauc et al. study[7] which was not identified in the subsequent GWAS of IgG N-glycome potentially due to the increase of power as we include the highest number of samples in the discovery analysis.

Replication analysis was conducted on LLS, KORA F4, EGCUT and GCKD cohorts which represent an admixture of studies collected for various purposes, including population-based studies but also disease-specific cohorts such as GCKD. Replication analysis in such a sample might be challenging but also represents evidence for the robustness of the results. When looking into the same glycan-SNP association as in discovery analysis, 34 genome-wide loci were replicated at the significance threshold of 0.001. Given that the replication study was underpowered compared to the discovery analysis, we take into account SNPs that are in LD with top SNPs as they are all considered replication candidates[63]. When looking into these SNPs and SNPs with the same effect direction as in discovery analysis, additional 6 loci are considered as replicated. However, the remaining two loci are not replicated but given the restricted sample size and the nature of the cohorts used, as well as the potential heterogeneity introduced by differences in phenotype measures, we cannot argue whether they are false positives.

We estimate SNP-heritability based on more than 1 million SNPs using LDSC and observe the highest heritability estimates among sialylation traits, while the lowest is for galactosylation

traits. The SNP heritability estimate for fuc trait was 0.22 but also had a high SE of 0.1 indicating the low power for estimation of heritability using the current GWAS data. These findings indicate the need for additional samples for GWAS of IgG N-glycome to capture the SNP-based heritability.

Secondary associations were detected in five loci, among which three are loci harbouring glycosyltransferases, enzymes responsible for the addition of sugar units to the growing glycan chain. The multiple independent associations in three glycosyltransferases loci were observed in Klarić *et al.*[71] study with the same number of independent SNPs in *FUT8* locus (n=6). As expected, *FUT8* locus is associated solely with the fucosylation phenotype.

ST6GAL1 locus is associated exclusively with phenotypes describing monosialylation (s1, s1_g1, s1_g2, s1_no_bis, s1_gal_total) but with differing numbers of independent associations. The s1_g2 trait has the largest number of independent associations (n=5), where the strongest association is in high LD with a variant that is also associated with all other traits. The s1_g2 trait is the ratio of monosialylated and digalactosylated structures and as such mostly represents the trait that describes the monosialylation by itself rather than monosialylation with galactosylation as captured by other monosialylation traits.

B4GALT1 locus was significantly associated with all traits from galactosylation and monosialylation phenotype sets as expected since galactosylated glycans are the substrate for ST6GAL1 enzyme which adds sialic acid. The number of independent associations across the 9 traits ranged from one to five, where g1 and s1_g2 traits had one, and s1_g1 and s1_no_bis had 5 independent associations. *RUNX1* locus on chromosome 21 harbours two independently associated SNPs for bisection trait, while galactosylation (g0, g2 and gal_total) and monosialylation (s1 and s1_no_bis) phenotypes had a single associated SNP in the same region. The differing number of independent associations in the same region among traits could be either the lack of statistical power or that traits do have a different number of causal SNPs. The uncertainty in estimation might stem from the experimental error between different glycans. The glycans containing sialic acids carry a negative charge making them more prone to experimental error. Therefore, this might potentially explain the differences in the number of independent associations in galactosylation and sialylation phenotypes in *RUNX1* and *B4GALT1* loci.

On chromosome 14, s1_g2 trait was associated with two independent variants, rs11624007 and rs10444775. This region harbours genes coding for heavy chains of immunoglobulins but also

TMEM121 gene. Previous GWAS by Klarić *et al.*[71] also found an association with a variant in *TMEM121* gene, while no evidence was found for association with the *IGH* locus. However, multivariate GWAS of IgG N-glycome by Shen *et al.*[8] found the association with *IGH* locus. Previously, differences in glycosylation profiles of wild-type IgG and IgG proteins with single amino acid substitutions in heavy chains were observed with the assumption that interaction of glycan and protein might be reduced, thereby increasing the accessibility to glycosylation enzymes[157]. However, given the independent association with *TMEM121*, its potential role in IgG glycosylation cannot be excluded.

After conditional analysis, the percentage of variance explained could be derived to avoid over or underestimation of the variant effect in the single-SNP analysis. The variance explained increased due to multiple loci having secondary associations which contributed to more variance explained in all traits except g1. The variance explained for monosialylation was 22% which is also the percentage reported by Klarić *et al.* for IGP29, the degree of monosialylation of fucosylated digalactosylated structures without bisecting GlcNAc. The limitation of this study in discovering loci associated with sialylation phenotypes is the lack of traits describing disialylation which is not measured by LC-MS and therefore is omitted from the analysis. Given the rank transformation of glycan traits before the genome-wide association test, the variance explained must be interpreted accordingly.

Using positional, eQTL and 3D chromatin mapping, 336 genes in total were mapped across 42 genomic regions, making it difficult to discuss the possible roles of those genes in the process of IgG glycosylation. Therefore, multiple gene prioritization approaches were applied to narrow down the list of the candidate genes, such as identifying 1) genes coding for proteins that are functionally affected by the genetic variants 2) pleiotropy with gene expression in the whole blood 3) pleiotropy with gene expression in immune cells, 4) genes in gene analysis based on SNP-wise model (MAGMA) and ultimately, 5) positionally mapped genes. More than 90% of the candidate variants are located in the noncoding regions of the genome, intronic and intergenic regions, and the same is observed in the majority of the GWAS studies of complex diseases and traits[65,158,159]. A low percentage (<5%) of the variants is found in the gene coding regions where they could have a direct impact on the encoded protein by altering its structure. The human genome contains protein-altering mutations but only a small number are considered deleterious as they can introduce premature stop codons or disrupt normal splicing of

mRNA[160]. A missense variant changes the encoded amino acid and consequently affects protein stability[161], protein-protein interactions[162], activity[163] and protein folding[164], hence making it likely that the gene which is affected by the candidate SNPs resulting in missense mutation, is the causal gene. Given the extensive LD structure in the IgG N-glycome-associated loci, the number of considered candidate SNPs is relatively high. Several loci contain more than one gene that could be affected by candidate variants, making it impossible to prioritize one of the genes.

Gene expression measurements with RNA sequencing[165] or microarrays[166] have been utilized in the GWAS setting as outcome traits. Such studies are called expression quantitative trait locus or eQTL analysis. We used colocalization analysis as one of the ways to prioritize genes from the genomic regions because the evidence for shared association pattern between gene expression and IgG glycosylation could indicate the same underlying mechanism for pathway control and thus prioritize the given gene over other genes in the region. Colocalization also contributes to the understanding of the biological basis for association with IgG N-glycosylation as altered gene expression might be the intermediate phenotype[68]. eQTLgen[146] dataset represents the biggest resource for eQTL data from human whole blood samples. A total of 31,684 samples was used in the generation of eQTL associations, thereby increasing the power to detect them. The eQTLgen dataset offers high SNP coverage across the whole genome which makes it suitable for use in colocalization analysis by coloc which requires up to a few hundreds of SNPs as input. The individuals included in eQTL analysis in eQTLgen are of European descent making it appropriate for our analysis as the method used assumes that samples in two GWAS analyses where we are undertaking the colocalization test are drawn from the same population, meaning that the allele frequencies and pattern of linkage disequilibrium (LD) match in both populations.

The advantage of the approximate Bayes factor method used in colocalization analysis is that it enables quick computation of posterior probabilities by use of p-values and MAFs only, or estimated allelic effect and standard error. However, there is a requirement for the availability of the same SNPs in both primary and secondary GWAS, but many studies are conducted using different imputation panels or the quality of imputation is low, resulting in a low overlap between SNPs and detection of pleiotropy becomes less powered. So the posterior probabilities should be interpreted with caution as a low posterior probability for shared variant (PP4) might

not indicate the lack of pleiotropy when PP3 is low, but rather the limited power when PP0, PP1, and PP2 are high.

Our knowledge of the regulation of IgG N-glycosylation remains limited and so does the involvement of various cell types and gene regulation in the specific cells and its effect on the N-glycosylation process. Also, there can be several eQTLs showing the matching pattern with a trait of interest, thus colocalization analysis generates high PP4 for multiple genes, but only one of the genes is biologically relevant, thereby illustrating the principle of PP4 being the measure of correlation and not causality. Additionally, in the case of multiple independent associations in the region of interest, the drawback of the algorithm is that it considers the strongest of the associations. But it is important to note that high PP4 indicates "at least one causal variant" and that low PP4 means that the two phenotypes do not share all of the causal variants and not that the two phenotypes cannot share one.

Immunoglobulin G is produced by mature B cells or plasma cells where the addition of N-glycan to IgG occurs as it travels through the ER and GA where glycosyltransferases and glycosidases act to form the N-glycan. It is reasonable to hypothesize that the levels of N-glycans that contain a certain sugar unit are mainly controlled by the levels of glycosyltransferase in a plasma cell. But also, IgG glycome is controlled by the T cell and B cell activation and IgG production[167,168]. Multiple molecules which are involved in immune activation can stimulate the B cells to produce different glycoforms of IgG[169] in addition to environmental signals which can modulate gene expression via epigenetic mechanisms in B cells, as well as other relevant cell lines[92,170]. Therefore, restricting the analysis to the gene regulation in B cells where IgG N-glycosylation occurs is not optimal as the external signals by other cells of the immune system could be driving the changes in glycosylation pattern. Moreover, in the case of sialylation phenotype, there is a study suggesting that IgG sialylation occurs in the circulation and not within B-cells[171]. The available single-cell eQTL datasets are not well powered and the overlap between SNPs in these studies with SNPs in the IgG N-glycome GWAS is low to allow for reliable and well-powered colocalization test and therefore we concentrate our colocalization efforts on to whole blood eQTL dataset, eQTLgen. However, we do look into the overlap of immune cell type-specific eQTLs and IgG N-glycome associated variants via the FUMA platform but we do not perform colocalization analysis.

Among four glycosyltransferases genes, only variants in *MGAT3* locus show the same pattern of association with IgG glycosylation and gene expression in whole blood. *MGAT3* gene encodes the *N*-acetylglucosaminyltransferase III enzyme which catalyses the addition of GlcNAc to the core β-mannose unit of N-glycans[72]. Since the association pattern in the *MGAT3* locus matches the pattern of eQTL associations for this gene, it implies that the variants regulate the expression of *MGAT3* gene and subsequently influence the levels of structures with bisecting GlcNAc in the IgG N-glycome.

Since glycosyltransferases are functioning in the constricted area in the cell, Golgi apparatus, the levels of enzymes are low so in turn, their gene expression might be hard to detect in the expression level measurements in whole blood samples. For *ST6GAL1*, *B4GALT1* and *FUT8*, the colocalization analysis showed high posterior probability for support for hypothesis 3 (PP3=1; both traits have associations in the region, but different causal variants), however, since conditional analysis has shown evidence for multiple independent associations in these loci, this might disrupt the colocalization test as it considers only the strongest association.

B4GALT1 gene is located on chromosome 9 and it codes for *beta-1,4-galactosyltransferase 1*, an enzyme catalysing the transfer of galactose unit to the GlcNAc residue in the non-reducing end of N-linked glycans[155]. In the study by Klarić *et al.*, it was shown that the variants in *B4GALT1* were pleiotropic with the expression of this gene in B cells, thus indicating the regulation of the *B4GALT1* expression as the mechanisms for the control of IgG galactosylation[10]. However, the pleiotropy was implicated for expression of *B4GALT1* in CD4 and CD19 cells and IGP8 glycan, while there was a lack of evidence for pleiotropy for other glycans due to multiple associations in the region, further confirming the restrictions of the colocalization tests in case of multiple independent associations in the locus.

It is important to note that in colocalization analysis, we take into account genes from the defined region around the top SNP, thus making it possible that we fail to see the effect of the variant on a gene that is positioned further away on the chromosome.

Given the prior knowledge and direct involvement of glycosyltransferases enzymes in the IgG glycosylation process, we prioritize these genes in their corresponding loci even if the additional genes are mapped in the same region. Locus on chromosome 3 (chr3:186607935-186819448) harbours *ST6GAL1*, a gene encoding β-galactoside-α-2,6-sialyltransferase 1, an enzyme responsible for the addition of α2,6-linked sialic acid units to terminal galactose structures in N-linked glycans[72]. The associated region on chromosome 14 contains *FUT8*

gene, gene encoding fucosyltransferase 8, an enzyme involved in the transfer of fucose residue to the inner GlcNAc residue of N-linked glycans, referred to as core fucose[72].

The chromatin interaction mapping resulted in the highest number of mapped genes; however, the chromatin interaction mapping was not primarily considered in the prioritization of the genes. We restrict the chromatin interaction mapping to GM12878, a B cell-derived cell line, but also must consider that chromatin interactions are time and cell population dependent[172], thus representing a limitation in this approach. The gene prioritization by chromatin interaction data was applied for chr8:103542538-103550211 locus where *UBR5, RRM2B* and *ODF1* genes were not positionally mapped in the defined region but shown to interact with the variants in the region with the 3D conformation of the chromatin in GM12878 cells. In ten of the remaining 38 loci, we prioritize the same genes as Klarić *et al.*, while in nine overlapping loci we widen the list of the candidate genes and in one locus we change the prioritized gene.

Chromosome 1 harbours 3 genomic regions associated with IgG N-glycome, one of which contains previously identified[71,74,76] and prioritized *runt-related transcription factor 3 (RUNX3)* gene which encodes a transcription factor with a role in the maturation of B cells[173] and differentiation of T cells[174,175]**.** Wahl *et al.*[74] suggested that *RUNX3* could affect the glycosylation of IgG through mechanisms of T cell differentiation which was shown to stimulate B cells and thus influence IgG1 glycosylation. In addition, RUNX3 is known to interact with RUNX1[176], a runt-related transcription factor 1, encoded by a gene located on IgG glycome-associated locus on chromosome 21. *RUNX1* has a role in tumorigenesis, embryonic development, haematopoiesis and inflammatory response[177,178]. Given the strong association of these two loci in the functional network, we can speculate that the mutual action of these two genes is involved in the regulation of gene expression in IgG glycosylation.

Another locus on chromosome 1 is associated with fucosylation phenotype and we prioritize two genes, *MYC binding protein* (*MYCBP*) and *Ras-related GTP binding C* (*RRAGC*). The association pattern in this locus matches the eQTL pattern for *MYCBP* gene which has a role in transcriptional activation by MYC and its aberrant expression was observed in multiple cancers[179,180]. *RRAGC* gene encodes a RagC GTPase, an activator of mTORC1 upon sensing of cellular nutrients, which is shown to be mutated in follicular lymphoma where it enhances B cell activation and accelerates lymphomagenesis[181]. As such, none of the two genes were

associated with N-glycosylation processes, however, we should consider the role of *RRAGC* in B cell activation and proliferation which might have an indirect impact on IgG glycoprofile.

Locus on chromosome 4 associated with galactosylation and monosialylation phenotypes, harbours two genes, *NFKB1* and *MANBA*. *NFKB1* gene encodes a subunit of the nuclear factor of kappa light polypeptide gene enhancer in B-cells (NF-κB) TF family, known for its critical role in cell survival and inflammation[182,183]. The *Nfkb1*$^{-/-}$ mice display increased levels of inflammation and DNA damage which could lead to cancer, as well as rapid ageing phenotype[184–186]. On the other hand, *MANBA* encodes beta mannosidase which is an exoglycosidase enzyme cleaving beta-mannose residues from the non-reducing end of N-linked glycans[187]. How this function might be linked to IgG galactosylation and monosialylation phenotypes is still unknown, hence both *NFKB1* and *MANBA* represent credible candidate genes in the region.

In locus on chromosome 7, we prioritize both *diacylglycerol lipase beta (DAGLB)* and *KDEL endoplasmic reticulum protein retention receptor 2 (KDELR2)* genes, while Klarić *et al.* prioritize only *DAGLB*, a gene which encodes a serine hydrolase with a role in proinflammatory signalling in neuroinflammation[188]. *KDELR2* gene codes for a member of the KDEL receptor family which has a function in recycling ER-resident proteins from the GA back to ER[189], while also being essential for Golgi-to-plasma protein trafficking[190]. Both genes contain eQTLs in this associated region influencing their expression in both B and T cells and their functions indicate importance in immune response and protein transport, however, their functions have not been described in the context of glycosylation, which makes it hard to speculate on prioritization of one of the genes.

Kinesin Family Member 11 (KIF11) and *Hematopoietically Expressed Homeobox (HHEX)* genes are found on chromosome 10 in a region associated with IgG fucosylation phenotype. *KIF11* encodes a member of the kinesin protein family with a role in the formation of a bipolar spindle during mitosis[191], but also protein secretions from the Golgi to the cell surface[192]. As such, the encoded protein could potentially play a role in the secretory pathway which upon addition of glycan chains to IgG in Golgi. On the other hand, *HHEX* encodes a transcription factor involved in the regulation of memory B cell differentiation[193] which differentiate and proliferate as antigen-secreting cells upon antigen encounter[194]. Although distinct, the functions of both proteins might be relevant in the process of IgG glycosylation.

A locus harboring *TNF receptor superfamily member 13B* (*TNFRSF13B*) gene was newly identified in the recent multivariate GWAS of IgG N-glycome[76] and in the current work, we find the association with monogalactosylation phenotype. *TNFRSF13B* encodes a transmembrane activator calcium modulator and cyclophilin ligand interactor (TACI) protein, a lymphocyte-specific member of the tumor necrosis factor (TNF) receptor superfamily, which is involved in the signalling pathway leading to B cell differentiation and antibody production[195,196]. The described roles might indicate its importance in the production and secretion of antibodies with specific glycoprofile.

It is important to mention one locus on chromosome 20 which we replicate from the previous study[71] but we prioritize a different gene. Initially, *MGME1* gene was prioritized based on its proximity to the associated variants, however, we prioritize *RRBP1, ribosome binding protein 1*, but solely based on the presence of B cell eQTLs in the bisection-associated locus even though the gene is not physically positioned in the defined region. *RRBP1* encodes a ribosome-binding protein found on the ER membrane and as such is implicated in the transport and secretion of intracellular proteins in the mammalian cells[197].

A newly discovered locus on chromosome 20 harbours *nuclear factor of activated T cells (NFATC2)* gene, one of the calcium-regulated members of the NFAT family of TFs with a role in gene expression regulation in the immune response to antigen[198]. Together with other members of NFAT family, *NFATC2* plays a crucial role in T cell activation, differentiation and proliferation, and cytokine balance maintanance[199] but also B cell homeostasis, as the *NFATC2*-deficient cells were shown to exhibit hyperactivation and increased immunoglobulin secretion. The described function indicates a potential role of *NFATC2* in IgG glycosylation via control of B cell activation and IgG secretion.

We omit the MHC region (chr6:30798697-32879471) from the prioritization efforts due to the complexity of the region and extensive LD structure[200] which makes it challenging to pinpoint single or only a few genes which might be causal.

Besides the expected enrichment in glycosylation-related gene sets, results of the gene-set enrichment test indicate a potential role of transport of substrates and enzymes and their availability for the N-glycosylation process to occur. Additionally, activation of B and T cells also points to mechanisms of glycosylation control via control of cell proliferation and development rather than solely control of expression of glycosyltransferases genes which are directly involved in the process. The question remains whether the specific changes in IgG N-

glycome are due to the proliferation of specific B cell clones which contain IgG with certain glycosylation as the gene-sets we observe do point to the importance of T cells and potential interplay with B cells to define IgG glycoprofile.

The idea of functional network construction was based on the work by Klarić *et al.*[10] where SNP glycome-wide effects were used to detect correlations between effects of different genomic regions on glycans and their potential involvement in the same biological pathway. This approach allows for hypothesis generation as it can indicate the effect of one gene on another, such as transcription factors influencing the expression of a gene.

In the current work, we replicate edges between *RUNX1* and *RUNX3* genes which were previously associated with bisection trait; however, we observe positive associations with all trait sets except fucosylation. Klarić *et al.*[10] hypothesize that RUNX1 and RUNX3 together with the chromatin remodelling protein SMARCB1, regulate the expression of *MGAT3*, resulting in an increased incidence of IGP40 (bisecting GlcNAc in all fucosylated disialylated structures of IgG). *RUNX1* and *RUNX3* were shown to be expressed in B cells, where RUNX3 binds near the transcription start site to inhibit *RUNX1* transcription, thereby decreasing the proliferation ability of the lymphoblastoid cells[201]. Furthermore, RUNX proteins regulate multiple B-cell-specific genes throughput various developmental stages of B cells[202]. In addition, components of the network cluster such as *NFKB1, TBX21* and *RUNX* genes were previously shown to be involved in T cell development[186,203,204], indicating the potential role of B and T cell interaction or clonal selection on immunoglobulin glycosylation patterns.

In the subnetwork containing *ST6GAL1* gene, we observe correlations between *ST6GAL1* and *NXPE1/NXPE4* and between *NXPE1/NXPE4* and *MEF2B/MAU2* locus. This is the replication of the subnetwork generated by Klarić et al., but the difference lies in the prioritized gene in the locus on chromosome 19, where Klarić et al. prioritized *RFXANK* and we prioritize *MEF2B* and *MAU2* genes. Given the fact that the functions of *NXPE1/NXPE4* and *MEF2B/MAU2* genes were not described in the immune system context and the lack of literature evidence of their connection to *ST6GAL1*, further speculations are currently not possible.

The variants in *TXLNB* and *TCF3* loci exhibit glycome-wide effect correlation as well as variants in *TXLNB* and *TMEM121/IGH* loci. *TXLNB* encodes beta taxilin, a component of intracellular vesicle transport. *TCF3* is a transcription factor associated with expression of genes during B lymphocyte development[205,206], but also shown to bind enhancer in *IGH* locus encoding heavy chain portion of immunoglobulin in humans[207], thus their mutual correlation

could be explained by the differences in immunoglobulin heavy chains which might affect glycosylation of IgG as we discussed earlier.

B4GALT1, *COG7/GGA2* and *SLC38A10* clustered together previously[71] but now we observe a correlation of *COG7/GGA2* and *SLC38A10* with the novel *SPPL3* gene. *SPPL3* is known to affect N-glycosylation by the release of active site-containing ectodomains of glycosyltransferases within the Golgi compartment, thereby reducing the levels of active enzymes including *B4GALT1*[208]. Further, COG complex has a function in the recycling of glycosyltransferases localized in Golgi apparatus such as B4GALT1 and ST6GAL1[209,210]. Recently, it was suggested that besides its role in amino acid transport, *SLC38A10* has a signalling role in the Golgi membrane for sustaining of protein synthesis and modifications[211]. The observed SNP glycome-wide correlations between these loci implicate the importance of levels of active glycosyltransferases enzymes in Golgi apparatus for the creation of specific glycosylation patterns especially for galactosylation and subsequently sialylation of IgG.

FUT8 and *MGAT3* were not observed among the network nodes due to the lack of significant edges with other loci since most of the fucosylation and bisection-related genomic regions are trait-specific, hence resulting in lower correlation values of SNP effects across traits. The intermediate phenotypes for glycosylation patterns are needed to increase the connectivity of trait-specific loci to other loci just as shown in the network constructed by Klarić *et al.*

StringPPI relies on existing knowledge about protein-protein interactions mainly relying on public resources to construct the patterns of interactions between proteins[153]. The main drawback of this approach is the overrepresentation of well-studied pathways and less evidence for understudied processes, including glycosylation. We deliberately omit the edge construction based on text mining as this is considered less supportive of the interaction than other approaches which include knowledge of interaction from the curated databases, experimental evidence, protein homology and co-expression of the two proteins. However, these are also biased towards more studied traits and therefore, we observe a cluster of glycosyltransferases with only one additional protein (MANBA) as further interactions have not been explored yet.

We fail to replicate the majority of edges from the inference-based network mainly due to the lack of existing knowledge about glycosylation-specific protein interactions in StringPPI but also the lack of correlations among loci in the inference-based network where many of the loci are isolated and lack significant connections with other genes due to phenotype definition in

GWAS. However, both networks can be used for different purposes, such as setting hypotheses about transcriptional regulation or co-dependence of genes in the same subnetwork when interpreting the network based on SNP glycome-wide effects. On the other hand, the StringPPI network can give us an overview of the involvement of the genes in the same pathways or having similar molecular functions without necessarily exploiting their direct interaction.

Along with the first genome-wide association study of IgG N-glycome[7], it was shown that the discovered regions overlap with risk loci for numerous autoimmune and inflammatory conditions including SLE, RA, UC, CD, T1D, MS, celiac disease, Graves' disease, nodular sclerosis and haematological cancers. But even long before that, the differences in IgG N-glycome profile were observed in RA patients[28] where they occur years before the disease manifests, with further studies[212–214] implying that IgG glycans can reflect a predisposition or act as effectors in the disease pathogenesis.

In an overview of existing associations with IgG N-glycome loci, we observe significant associations in 21 genomic loci with at least one trait or disease, among which novel loci were associated with immune-system related diseases and traits such as allergic disease, WBC count, T2D and PBC. However, we cannot make any inference about changes in IgG N-glycosylation and diseases sharing the same underlying genetic mechanism as the pattern of association and causal genes or variants can differ between phenotypes.

Klarić *et al.*[71] investigated regional association patterns to see whether the same causal variants rather than just genes are shared between IgG glycosylation and diseases. Given the increased number of novel genomic loci, we also applied the Approximate Bayesian method for colocalization test to investigate the sharing of the same causal variants between IgG N-glycosylation and range of autoimmune and inflammatory diseases. We retrieve positive tests for colocalization for fifteen different diseases and traits across ten genomic regions and six glycan traits. In the region on chromosome 17 (chr17:37579383-38215117) which is a fucosylation-specific locus, we observe pleiotropy with UC, CD, IBD, PBC and RA, all of which were previously observed in Klarić *et al.* study. In the current study pleiotropic effects with asthma, allergy, HDL and SLE were not observed as all the tests for these traits have shown higher posterior probability for hypothesis 3 which states that both traits are associated with the genomic region, but the causal variant is not shared. The difference could lie in the glycan traits for which the pleiotropy was tested as the previous study observed significant associations with traits describing agalactosylated and monogalactosylated structures which

could potentially be associated with the different causal variants as opposed to differential levels of fucosylation. This locus contains at least 3 genes with a potential role in glycosylation: *IKZF3*, *ORMDL3* and *GSDMB*, all of which are also indicated as candidate genes in higher risk for asthma[215–217].

Another locus on chromosome 17 (chr17:43463493-44865603) displays pleiotropic effect for IgG monosialylation and SLE, PBC, schizophrenia, breast cancer, as well as a trait describing the percentage of WBC. Unravelling the exact mechanism through which this locus affects IgG glycans and the mentioned complex diseases might be a laborious task as this region spans at least eight candidate genes.

One of the loci on chromosome 5 which was initially discovered in the first GWAS of IgG N-glycome[7] but not replicated in GWAS by Klarić *et al.* was shown to be pleiotropic for monosialylation (s1_g2) and JIA and RA. Previously, a study of changes in IgG glycosylation in JIA cases and controls demonstrated that JIA cases exhibit lower levels of IgG galactosylation and sialylation[218]. Furthermore, in rheumatoid arthritis, the levels of galactosylation and sialylation were lower in healthy cases and controls[219], and these changes appear to be relatively stable and present years before the RA diagnosis[213]. The locus exhibits pleiotropy for the expression of interleukin 6 cytokine family signal transducer (*IL6ST*) and ankyrin repeat domain 55 (A*NKRD55*) in whole blood. *ANKRD55* gene is located in a region on chromosome 5 associated with one of the monosialylation phenotypes, s1_g2. The variants in *ANKRD55* gene were associated with a range of diseases including multiple sclerosis[220], rheumatoid arthritis[221], diabetes[222,223], celiac disease[224], Crohn's disease[225] and Alzheimer's disease[226]. Except containing the ankyrin repeats which are important for protein-protein interactions, the exact function of the encoded protein is not yet clear. Ankyrin repeat proteins are known to be involved in the correct placement and orientation of membrane proteins to compartments in the ER and plasma membrane[227]. However, eQTL colocalization analysis has shown that the variants in this locus are associated with the expression of *IL6ST* thus it cannot be excluded as a candidate gene. *IL6ST* encodes a signalling receptor subunit which is shared by several cytokines such as interleukin 6 (IL6), ciliary neurotrophic factor (CNTF), leukaemia inhibitory factor (LIF) and oncostatin M (OSM) and variants in *IL6ST* locus were previously associated with risk for rheumatoid arthritis and multiple myeloma[228]. Lower levels of sialylation are constantly observed changes in autoimmune diseases including RA and JIA and

the colocalization with IgG glycosylation loci could bring new evidence and potential role of the genes in this locus to understand the underlying mechanism of these changes.
The colocalization of schizophrenia and monosialylation signals was observed in locus harbouring *SPPL3* gene. *SPPL3* gene encodes an enzyme that sheds the activated domains from the glycosyltransferases to control the levels of active enzymes in GA and its role has been recognized in overall glycosylation[208]. The pleiotropy of the locus for schizophrenia and monosialylation can indicate the inflammatory context of the disease[229] or just the common function of *SPPL3* in both IgG glycosylation and post-translational modifications of the proteins in schizophrenia[230] through the same underlying genetic mechanism.

In *OVOL1/AP5B1* locus we observe a potential pleiotropic effect on the risk for asthma, SLE and allergy. Pleiotropy for asthma was also previously shown in multivariate GWAS[76] where *OVOL1/AP5B1* locus was initially discovered. *HHEX-KIF11* locus on chromosome 10 colocalizes with risk loci for T2D, AOA, asthma and percentage of lymphocytes. The locus is fucosylation-specific, potentially indicating the importance of core fucosylation levels for the mentioned diseases. *KIF11* gene encodes a transporter protein which is part of the complex required for transport of protein from Golgi complex to the cell surface[192], while *HHEX* encodes a hematopoietic transcription factor important for lymphopoiesis and pancreatic development[231]. A study of IgG N-glycan patterns in T2D has shown that T2D is associated with decreased fucosylation in glycan structures without bisecting GlcNAc and increased fucosylation in glycan structures with bisecting GlcNAc, thus indicating a higher ADCC potential of IgG in T2D[232].

Locus on chromosome 17 (chr17:43463493-44865603) associated with monosialylation and total galactosylation was shown to be potentially pleiotropic for IgG glycosylation and breast cancer. Previous studies of IgG N-glycans across different cancers have shown that there is no unique pattern of change in IgG glycans in cancer patients, although, a substantial number of cancers display a decrease in the level of IgG galactosylation when compared to healthy controls, such as non-small cell lung cancer, gastric cancer, breast cancer, ovarian cancer, prostate cancer, lung cancer, colorectal cancer and multiple myeloma[233]. The decrease in galactosylation of IgG might reflect the host's defensive immune response, but also potential activation of acute-phase response pathways in cancer progression or decreased binding of IgG

Fc portion to complement component subsequently affecting the depletion of complement-dependent cytotoxicity and cancer cell escape[4].

The enrichment analysis of cell-type-specific regulatory regions indicated the enrichment across immune cell types in blood, such as B and T cells, monocytes, and NK cells. These findings are an indication of the functional relevance of discovered variants for the regulation of gene expression in cells relevant for the immune system. This also provides evidence for the notion that IgG glycosylation patterns are greatly defined by the change in expression of glycosyltransferases and other genes involved in this process and is not mainly dependent on changes in protein structures to diminish or lessen their function.

6. CONCLUSIONS

This is the largest genome-wide association study of the human IgG N-glycome to date, involving a total of 13,705 samples in the discovery analysis and 7,775 in the replication analysis. The increase in the number of samples was enabled through the use of derived IgG N-glycan traits which were harmonized across UPLC and LC-MS-measured samples. The increase in sample size led to the discovery of 13 novel loci associated with IgG N-glycosylation.

Approximate conditional analysis for multiple independent associations resulted in multiple independently associated SNPs in the glycosyltransferases loci, *FUT8, ST6GAL1, B4GALT1*, just as in the previous work by Klarić et al. Additional locus on chromosome 14 also contained two associations in *TMEM121* gene and *IGH* genes, pointing to different causal variants and genes which might independently affect monosialylation levels of IgG.

We narrow down the mapped genes to a set of 83 candidate genes either through the pleiotropy with their expression in whole blood, variant association with gene expression in immune cells (especially B-cells), damaging effect of the variants on the protein structure, and in cases where we lack the mentioned evidence, the genes were suggested by gene-based analysis or positional mapping.

The gene-set enrichment analysis has shown that besides the enrichment in the N-glycosylation pathway, there was a significant enrichment in B- and T-cell development pointing to the potential mechanisms of regulation via clone development and proliferation. Additionally, candidate genes were enriched in cellular transport gene-sets indicating the importance of substrate availability and transport across the cellular membrane, as well as intracellular transport in the regulation of IgG N-glycosylation.

With functional network which was constructed based on the approach described by Klarić et al., we were able to form a hypothesis of how genes in the genomic loci could be regulating the process of IgG N-glycosylation. We found significant edges between *TXLNB, TCF3* and *TMEM121/IGH* loci which might reflect the *IGH* gene expression regulation by *TCF3* and, hence, the glycosylation regulation based on the difference in immunoglobulin heavy chain composition.

Pleiotropic effects of the variants for IgG N-glycosylation and a range of autoimmune and inflammatory diseases exist. This is indicative of previous assumptions that the IgG N-glycosylation is either changing with the course of the disease and appears as the consequence or the changes in IgG N-glycans can be indicative of disease development in the future.

The function of the prioritized genes in IgG N-glycome GWAS can be confirmed in the suitable model system which secretes IgG in high concentrations to allow for glycoprofiling, such as the Freestyle HEK293 cell line transfected with a vector coding for IgG light and heavy chains from the Zoldoš group at the University of Zagreb.

7. REFERENCES

1. Apweiler, R., Hermjakob, H. & Sharon, N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta - Gen. Subj.* **1473**, 4–8 (1999).
2. Vidarsson, G., Dekkers, G. & Rispens, T. IgG subclasses and allotypes: From structure to effector functions. *Front. Immunol.* **5**, 520 (2014).
3. Shade, K.-T. C. & Anthony, R. M. Antibody Glycosylation and Inflammation. *Antibodies 2013, Vol. 2, Pages 392-414* **2**, 392–414 (2013).
4. Martinić Kavur, M., Lauc, G. & Pezer, M. *Systems Glycobiology: Immunoglobulin G Glycans as Biomarkers and Functional Effectors in Aging and Diseases*. *Comprehensive Glycoscience* (2021). doi:10.1016/b978-0-12-819475-1.00086-9
5. 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 (2011).
6. Gudelj, I., Lauc, G. & Pezer, M. Immunoglobulin G glycosylation in aging and diseases. *Cell. Immunol.* **333**, 65–79 (2018).
7. Lauc, G. *et al.* Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers. *PLoS Genet.* **9**, (2013).
8. Shen, X. *et al.* Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation. *Nat. Commun.* **8**, 1–10 (2017).
9. Wahl, A. *et al.* Genome-wide association study on immunoglobulin G glycosylation patterns. *Front. Immunol.* **9**, (2018).
10. Klarić, L. *et al.* Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. *Sci. Adv.* **6**, (2020).
11. Varki, A. Essentials of Glycobiology [Internet]. *Cold Spring Harb.* 823 (2015).
12. Nairn, A. V. *et al.* Regulation of glycan structures in animal tissues: Transcript profiling of glycan-related genes. *J. Biol. Chem.* **283**, 17298–17313 (2008).
13. Aebi, M. N-linked protein glycosylation in the ER. *Biochimica et Biophysica Acta - Molecular Cell Research* **1833**, 2430–2437 (2013).

14. Zoldoš, V., Grgurević, S. & Lauc, G. Epigenetic regulation of protein glycosylation. *Biomolecular Concepts* **1**, 253–261 (2010).
15. Alberts, B. *et al. Molecular Biology of the Cell* . (Garland Science, 2002).
16. Schur, P. H. IgG subclasses. A historical perspective. *Monographs in allergy* **23**, 1–11 (1988).
17. Vidarsson, G., Dekkers, G. & Rispens, T. IgG subclasses and allotypes: From structure to effector functions. *Front. Immunol.* **5**, (2014).
18. van de Bovenkamp, F. S., Hafkenscheid, L., Rispens, T. & Rombouts, Y. The Emerging Importance of IgG Fab Glycosylation in Immunity. *J. Immunol.* **196**, 1435–1441 (2016).
19. Niwa, R. *et al.* Enhancement of the antibody-dependent cellular cytotoxicity of low-fucose IgG1 is independent of FcγRIIIa functional polymorphism. *Clin. Cancer Res.* **10**, 6248–6255 (2004).
20. Shields, R. L. *et al.* Lack of fucose on human IgG1 N-linked oligosaccharide improves binding to human FcγRIII and antibody-dependent cellular toxicity. *J. Biol. Chem.* **277**, 26733–26740 (2002).
21. Iida, S. *et al.* Nonfucosylated therapeutic IgG1 antibody can evade the inhibitory effect of serum immunoglobulin G on antibody-dependent cellular cytotoxicity through its high binding to FcγRIIIa. *Clin. Cancer Res.* **12**, 2879–2887 (2006).
22. Scanlan, C. N., Burton, D. R. & Dwek, R. A. Making autoantibodies safe. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 4081–4082 (2008).
23. Kapur, R. *et al.* A prominent lack of IgG1-Fc fucosylation of platelet alloantibodies in pregnancy. *Blood* **123**, 471–480 (2014).
24. Kapur, R. *et al.* Low anti-RhD IgG-Fc-fucosylation in pregnancy: A new variable predicting severity in haemolytic disease of the fetus and newborn. *Br. J. Haematol.* **166**, 936–945 (2014).
25. Zou, G. *et al.* Chemoenzymatic synthesis and Fcγ receptor binding of homogeneous glycoforms of antibody Fc domain. Presence of a bisecting sugar moiety enhances the affinity of Fc to FcγIIIa receptor. *J. Am. Chem. Soc.* **133**, 18975–18991 (2011).
26. Ferrara, C. *et al.* Modulation of therapeutic antibody effector functions by glycosylation engineering: Influence of golgi enzyme localization domain and co-

expression of heterologous β1, 4-N-acetylglucosaminyltransferase III and Golgi α-mannosidase II. *Biotechnol. Bioeng.* **93**, 851–861 (2006).

27. Gornik, O. & Lauc, G. *Glycosylation of serum proteins in inflammatory diseases*. *Disease Markers* **25**, (IOS Press, 2008).
28. Parekh, R. B. *et al.* Association of rheumatoid arthritis and primary osteoarthritis with changes in the glycosylation pattern of total serum IgG. *Nature* **316**, 452–457 (1985).
29. Malhotra, R. *et al.* Glycosylation changes of IgG associated with rheumatooid arthritis can activate complement via the mannose-binding protein. *Nat. Med.* **1**, 237–243 (1995).
30. Štambuk, J. *et al.* Global variability of the human IgG glycome. *Aging (Albany. NY).* **12**, 1–13 (2020).
31. Ahmed, A. A. *et al.* Structural characterization of anti-inflammatory immunoglobulin G Fc proteins. *J. Mol. Biol.* **426**, 3166–3179 (2014).
32. Sondermann, P., Pincetic, A., Maamary, J., Lammens, K. & Ravetch, J. V. General mechanism for modulating immunoglobulin effector function. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 9868–9872 (2013).
33. Kaneko, Y., Nimmerjahn, F. & Ravetch, J. V. Anti-inflammatory activity of immunoglobulin G resulting from Fc sialylation. *Science (80-.).* **313**, 670–673 (2006).
34. Selman, M. H. J. *et al.* Fc specific IgG glycosylation profiling by robust nano-reverse phase HPLC-MS using a sheath-flow ESI sprayer interface. *J. Proteomics* **75**, 1318–1329 (2012).
35. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).
36. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
37. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
38. Griffith, O. L. *et al.* ORegAnno: An open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* **36**, 107–113 (2008).
39. Affymetrix. Affymetrix Genome-Wide Human SNP Array 6.0 Data Sheet. (2007).
40. Genome-Wide DNA. Analysis BeadChips Data Sheet. *San Diego, Calif. Illumina* (2009).

41. Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics* **6**, 109–118 (2005).
42. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.* **38**, 663–667 (2006).
43. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499–511 (2010).
44. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
45. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* **5**, e1000529 (2009).
46. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
47. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
48. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
49. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
50. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
51. Zhang, B. *et al.* Practical consideration of genotype imputation: Sample size, window size, reference choice, and untyped rate. *Stat. Interface* **4**, 339–352 (2011).
52. Lettre, G., Lange, C. & Hirschhorn, J. N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet. Epidemiol.* **31**, 358–362 (2007).
53. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).

54. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
55. Zeggini, E. & Ioannidis, J. P. A. Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10**, 191–201 (2009).
56. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
57. Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F. & Botella, J. Assessing heterogeneity in meta-analysis: Q statistic or I 2 Index? *Psychol. Methods* **11**, 193–206 (2006).
58. Higgins, J. P. T. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology* **37**, 1158–1160 (2008).
59. Pfeiffer, R. M., Gail, M. H. & Pee, D. On combining data from genome-wide association studies to discover disease-associated SNPs. *Stat. Sci.* **24**, 547–560 (2009).
60. Kavvoura, F. K., John, A. E. & Ioannidis, P. A. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. doi:10.1007/s00439-007-0445-9
61. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
62. Zöllner, S. & Pritchard, J. K. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* **80**, 605–615 (2007).
63. Chanock, S. J. *et al.* Replicating genotype-phenotype associations. *Nature* **447**, 655–660 (2007).
64. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
65. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin 2015 81* **8**, 1–18 (2015).
66. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
67. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense

mutations. *Nature Methods* **7**, 248–249 (2010).

68. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014).
69. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1–11 (2017).
70. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
71. Taniguchi, N., Honke, K. & Fukuda, M. *Handbook of glycosyltransferases and related genes*. (Springer Science & Business Media, 2011).
72. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
73. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, (2015).
74. Shadrina, A. S. *et al.* Multivariate genome-wide analysis of immunoglobulin G N-glycosylation identifies new loci pleiotropic with immune function. *Hum. Mol. Genet.* **00**, (2021).
75. Pezer, M. *Antibody Glycosylation*. (Springer Nature, 2021).
76. Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.* **16**, 144–149 (2013).
77. Boeing, H., Korfmann, A. & Bergmann, M. M. Recruitment Procedures of EPIC-Germany. *Ann. Nutr. Metab.* **43**, 205–215 (1999).
78. Bergmann, M. M., Bussas, U. & Boeing, H. Follow-up procedures in EPIC-Germany - Data quality aspects. *Annals of Nutrition and Metabolism* **43**, 225–234 (1999).
79. Campbell, H. *et al.* Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. *Hum. Mol. Genet.* **16**, 233–241 (2007).
80. McQuillan, R. *et al.* Runs of Homozygosity in European Populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
81. Schoenmaker, M. *et al.* Evidence of genetic enrichment for exceptional survival using a family approach: The Leiden Longevity Study. *Eur. J. Hum. Genet.* **14**, 79–84 (2006).
82. Holle, R., Happich, M., Löwel, H. & Wichmann, H. E. KORA - A research platform

for population based health research. *Gesundheitswesen* **67**, 19–25 (2005).

83. Rathmann, W. *et al.* Hemoglobin A1c and glucose criteria identify different subjects as having type 2 diabetes in middle-aged and older populations: The KORA S4/F4 Study. *Ann. Med.* **44**, 170–177 (2012).
84. Halachev, M. *et al.* Increased ultra-rare variant load in an isolated Scottish population impacts exonic and regulatory regions. *PLoS Genet.* **15**, e1008480–e1008480 (2019).
85. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
86. Titze, S. *et al.* Disease burden and risk profile in referred patients with moderate chronic kidney disease: composition of the German Chronic Kidney Disease (GCKD) cohort. *Nephrol. Dial. Transplant.* **30**, 441–451 (2015).
87. Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
88. Langenberg, C. *et al.* Gene-Lifestyle Interaction and Type 2 Diabetes: The EPIC InterAct Case-Cohort Study. *PLOS Med.* **11**, e1001647 (2014).
89. Jäger, S., Cuadrat, R., Hoffmann, P., Wittenbecher, C. & Schulze, M. B. Desaturase Activity and the Risk of Type 2 Diabetes and Coronary Artery Disease: A Mendelian Randomization Study. *Nutr. 2020, Vol. 12, Page 2261* **12**, 2261 (2020).
90. Menni, C. *et al.* Glycosylation of immunoglobulin g: Role of genetic and epigenetic influences. *PLoS One* **8**, e82558 (2013).
91. Landini, A. *et al.* Genetic regulation of post-translational modification of two distinct proteins. *Nat. Commun. 2022 131* **13**, 1–13 (2022).
92. Trbojević-Akmačić, I., Ugrina, I. & Lauc, G. Comparative Analysis and Validation of Different Steps in Glycomics Studies. in *Methods in Enzymology* **586**, 37–55 (Academic Press Inc., 2017).
93. Krištić, 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).
94. Agakova, A., Vučković, F., Klarić, L., Lauc, G. & Agakov, F. Automated Integration of a UPLC Glycomic Profile. *Methods Mol. Biol.* **1503**, 217–233 (2017).
95. Selman, M. H. J. *et al.* Fc specific IgG glycosylation profiling by robust nano-reverse phase HPLC-MS using a sheath-flow ESI sprayer interface. *J. Proteomics* **75**, 1318–1329 (2012).

96. 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).
97. Karaman, I. Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis. in 145–161 (Springer, Cham, 2017). doi:10.1007/978-3-319-47656-8_6
98. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B* **44**, 139–160 (1982).
99. Ugrina, I., Klaric, L., Vuckovic, F. & Russell, A. glycanr: Tools for Analysing N-Glycan Data. (2018).
100. Dieterle, F., Ross, A. & Senn, H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures . Application in 1 H NMR Metabonomics. **78**, 4281–4290 (2006).
101. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
102. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
103. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161–e161 (2014).
104. R Core Team. R: A language and environment for statistical computing. (2017).
105. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
106. Haller, T., Kals, M., Esko, T., Magi, R. & Fischer, K. RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Brief. Bioinform.* **16**, 39–44 (2015).
107. Hirschhorn, J. N. & Gajdos, Z. K. Z. Genome-wide association studies: Results from the first few years and potential implications for clinical medicine. *Annu. Rev. Med.* **62**, 11–24 (2011).
108. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–212 (2014).
109. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

110. Lee, C. H., Cook, S., Lee, J. S. & Han, B. Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. *Genomics Inform.* **14**, 173 (2016).
111. Dahiru, T. P-Value, a true test of statistical significance? a cautionary note. *Ann. Ibadan Postgrad. Med.* **6**, (2011).
112. Panagiotou, O. A. *et al.* What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int. J. Epidemiol.* **41**, 273–286 (2012).
113. Judson, R., Salisbury, B., Schneider, J., Windemuth, A. & Stephens, J. C. How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics* **3**, 379–391 (2002).
114. Gauderman, J., Ph, D. & Morrison, J. QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies,. *http://hydra.usc.edu/gxe* 35–50 (2006).
115. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
116. Sanna, S. *et al.* Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLoS Genet.* **7**, e1002198 (2011).
117. Allen, H. L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
118. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–978 (2011).
119. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
120. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
121. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
122. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

123. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
124. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).
125. 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).
126. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics* **39**, 1181–1186 (2007).
127. Leslie, R., O'Donnell, C. J. & Johnson, A. D. GRASP: Analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**, (2014).
128. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
129. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715.e16 (2018).
130. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
131. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science (80-.).* **343**, (2014).
132. Momozawa, Y. *et al.* IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, (2018).
133. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
134. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* **17**, 2042–2059 (2016).
135. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
136. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl

API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).

137. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
138. Ng, P. C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* **12**, 436–446 (2002).
139. Dunham, I., Kulesha, E., Iotchkova, V., Morganella, S. & Birney, E. FORGE: A tool to discover cell specific enrichments of GWAS associated SNPs in regulatory regions. *F1000Research* **4**, 18 (2015).
140. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* **10**, 184–194 (2009).
141. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
142. Nica, A. C. *et al.* Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLoS Genet.* **6**, e1000895 (2010).
143. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P* -values. *Genet. Epidemiol.* **33**, 79–86 (2009).
144. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* **18**, 10 (2018).
145. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
146. Massy, W. F. Principal Components Regression in Exploratory Statistical Research. *J. Am. Stat. Assoc.* **60**, 234–256 (1965).
147. Hou, C. D. A simple approximation for the distribution of the weighted combination of non-independent or independent probabilities. *Stat. Probab. Lett.* **73**, 179–187 (2005).
148. Brown, M. B. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics* **31**, 987 (1975).
149. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).
150. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
151. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with

increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

152. Mahan, A. E. *et al.* Antigen-Specific Antibody Glycosylation Is Regulated via Vaccination. *PLOS Pathog.* **12**, e1005456 (2016).
153. Rini, J. M. & Esko, J. D. *Glycosyltransferases and Glycan-Processing Enzymes*. *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, 2015).
154. Russell, A. C. *et al.* The N-glycosylation of immunoglobulin G as a novel biomarker of Parkinson's disease. *Glycobiology* **27**, 501–510 (2017).
155. Lund, J., Takahashi, N., Pound, J. D., Goodall, M. & Jefferis, R. Multiple interactions of IgG with its core oligosaccharide can modulate recognition by complement and human Fc gamma receptor I and influence the synthesis of its oligosaccharide chains. *J. Immunol.* **157**, 4963–9 (1996).
156. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102 (2015).
157. Giral, H., Landmesser, U. & Kratzer, A. Into the Wild: GWAS Exploration of Non-coding RNAs. *Front. Cardiovasc. Med.* **5**, 181 (2018).
158. Miosge, L. A. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proc. Natl. Acad. Sci.* **112**, E5189–E5198 (2015).
159. Zhang, Z., Miteva, M. A., Wang, L. & Alexov, E. Analyzing Effects of Naturally Occurring Missense Mutations. *Comput. Math. Methods Med.* **2012**, 15 (2012).
160. Teng, S., Madej, T., Panchenko, A. & Alexov, E. Modeling Effects of Human Single Nucleotide Polymorphisms on Protein-Protein Interactions. *Biophys. J.* **96**, 2178 (2009).
161. S, W., K, T., C, S. & E, A. A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics. *Proteins* **79**, 2444–2454 (2011).
162. CM, D. Protein folding and misfolding. *Nature* **426**, 884–890 (2003).
163. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
164. Nica, A. C. & Dermitzakis, E. T. Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet.* **17**, R129–R134 (2008).
165. Pfeifle, R. *et al.* Regulation of autoantibody activity by the IL-23-T H 17 axis

determines the onset of autoimmune disease. *Nat. Immunol.* **18**, 104–113 (2017).

166. Hess, C. *et al.* T cell-independent B cell activation induces immunosuppressive sialylated IgG antibodies. *J. Clin. Invest.* **123**, 3788–3796 (2013).
167. Wang, J. *et al.* Fc-glycosylation of IgG1 is modulated by B-cell stimuli. *Mol. Cell. Proteomics* **10**, M110.004655 (2011).
168. Štambuk, T., Klasić, M., Zoldoš, V. & Lauc, G. N-glycans as functional effectors of genetic and epigenetic disease risk. *Molecular Aspects of Medicine* **79**, 100891 (2021).
169. Jones, M. B. *et al.* B-cell-independent sialylation of IgG. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 7207–7212 (2016).
170. Dekker, J. & Misteli, T. Long-Range Chromatin Interactions. *Cold Spring Harb. Perspect. Biol.* **7**, (2015).
171. Whiteman, H. J. & Farrell, P. J. RUNX Expression and Function in Human B Cells. *Crit. Rev. Eukaryot. Gene Expr.* **16**, 31–44 (2006).
172. Overgaard, N. H., Jung, J.-W., Steptoe, R. J. & Wells, J. W. CD4 + /CD8 + double-positive T cells: more than just a developmental stage? . *J. Leukoc. Biol.* **97**, 31–38 (2015).
173. Steinke, F. C. *et al.* TCF-1 and LEF-1 act upstream of Th-POK to promote the CD4 + T cell fate and interact with Runx3 to silence Cd4 in CD8 + T cells. *Nat. Immunol.* **15**, 646–656 (2014).
174. Spender, L. C., Whiteman, H. J., Karstegl, C. E. & Farrell, P. J. Transcriptional cross-regulation of RUNX1 by RUNX3 in human B cells. *Oncogene* **24**, 1873–1881 (2005).
175. Scheitz, C. J. F. & Tumbar, T. New insights into the role of Runx1 in epithelial stem cell biology and pathology. *J. Cell. Biochem.* **114**, 985–993 (2013).
176. Okuda, T., Van Deursen, J., Hiebert, S. W., Grosveld, G. & Downing, J. R. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell* **84**, 321–330 (1996).
177. S, M. & JF, M. c-Myc-induced genomic instability. *J. Environ. Pathol. Toxicol. Oncol.* **22**, 179–199 (2003).
178. Dang, C. V. MYC on the Path to Cancer. *Cell* **149**, 22 (2012).
179. Ortega-Molina, A. *et al.* Oncogenic Rag GTPase signaling enhances B cell activation and drives follicular lymphoma sensitive to pharmacological inhibition of mTOR. *Nat. Metab.* **1**, 775 (2019).

180. Pereira, S. G. & Oakley, F. Nuclear factor-κB1: Regulation and function. *Int. J. Biochem. Cell Biol.* **40**, 1425–1430 (2008).
181. Cartwright, T., Perkins, N. D. & Wilson, C. L. NFKB1: a suppressor of inflammation, ageing and cancer. *FEBS J.* **283**, 1812–1822 (2016).
182. CL, W. *et al.* NFκB1 is a suppressor of neutrophil-driven hepatocellular carcinoma. *Nat. Commun.* **6**, (2015).
183. GM, B. *et al.* Loss of Nfkb1 leads to early onset aging. *Aging (Albany. NY).* **6**, 931–943 (2014).
184. Oh, H. & Ghosh, S. NF-κB: Roles and regulation in different CD4+ T-cell subsets. *Immunol. Rev.* **252**, 41–51 (2013).
185. Zhu, M. *et al.* β-Mannosidosis mice: a model for the human lysosomal storage disease. *Hum. Mol. Genet.* **15**, 493–500 (2006).
186. Viader, A. *et al.* A chemical proteomic atlas of brain serine hydrolases identifies cell type-specific pathways regulating neuroinflammation. *Elife* **5**, (2016).
187. Capitani, M. & Sallese, M. The KDEL receptor: New functions for an old protein. *FEBS Lett.* **583**, 3863–3871 (2009).
188. M, M., ME, M., J, H. & A, V. Protein kinase A activity is required for the budding of constitutive transport vesicles from the trans-Golgi network. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 14461–14466 (1997).
189. Vale, R. D. The Molecular Motor Toolbox for Intracellular Transport. *Cell* **112**, 467–480 (2003).
190. Wakana, Y. *et al.* Kinesin-5/Eg5 is important for transport of CARTS from the trans-Golgi network to the cell surface. *J. Cell Biol.* **202**, 241–250 (2013).
191. Laidlaw, B. J., Duan, L., Xu, Y., Vazquez, S. E. & Cyster, J. G. The transcription factor Hhex cooperates with the corepressor Tle3 to promote memory B cell development. *Nat. Immunol. 2020 219* **21**, 1082–1093 (2020).
192. WE, P., TF, T., S, J., D, B. & MS, D. Memory B cells, but not long-lived plasma cells, possess antigen specificities for viral escape mutants. *J. Exp. Med.* **208**, 2599–2606 (2011).
193. Ozcan, E. *et al.* Transmembrane activator, calcium modulator, and cyclophilin ligand interactor drives plasma cell differentiation in LPS-activated B cells. *J. Allergy Clin. Immunol.* **123**, 1277-1286.e5 (2009).

194. Castigli, E. *et al.* Transmembrane activator and calcium modulator and cyclophilin ligand interactor enhances CD40-driven plasma cell differentiation. *J. Allergy Clin. Immunol.* **120**, 885–891 (2007).
195. Savitz, A. J. & Meyer, D. I. Identification of a ribosome receptor in the rough endoplasmic reticulum. *Nat. 1990 3466284* **346**, 540–544 (1990).
196. Rao, A., Luo, C. & Hogan, P. G. TRANSCRIPTION FACTORS OF THE NFAT FAMILY:Regulation and Function. *http://dx.doi.org/10.1146/annurev.immunol.15.1.707* **15**, 707–747 (2003).
197. Rengarajan, J., Tang, B. & Glimcher, L. H. NFATc2 and NFATc3 regulate T H 2 differentiation and modulate TCR-responsiveness of naïve T H cells. *Nat. Immunol. 2001 31* **3**, 48–54 (2001).
198. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol. 2017 181* **18**, 1–21 (2017).
199. Spender, L. C., Whiteman, H. J., Karstegl, C. E. & Farrell, P. J. Transcriptional cross-regulation of RUNX1 by RUNX3 in human B cells. *Oncogene* **24**, 1873–1881 (2005).
200. Voon, D. C. C., Hor, Y. T. & Ito, Y. The RUNX complex: Reaching beyond haematopoiesis into immunity. *Immunology* **146**, 523–536 (2015).
201. Li, Y. *et al.* Downregulation of Runx3 is closely related to the decreased Th1-associated factors in patients with gastric carcinoma. *Tumor Biol.* **35**, 12235–12244 (2014).
202. Mikami, Y. & Kanno, Y. GoldiRunx and Remembering Cytotoxic Memory. *Immunity* **48**, 614–615 (2018).
203. Jacobs, Y., Vierra, C. & Nelson, C. E2A expression, nuclear localization, and in vivo formation of DNA- and non-DNA-binding species during B-cell development. *Mol. Cell. Biol.* **13**, 7321–7333 (1993).
204. Quong, M. W., Harris, D. P., Swain, S. L. & Murre, C. E2A activity is induced during B-cell activation to promote immunoglobulin class switch recombination. *EMBO J.* **18**, 6307–6318 (1999).
205. Greenbaum, S. & Zhuang, Y. Identification of E2A target genes in B lymphocyte development by using a gene tagging-based chromatin immunoprecipitation system. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 15030–15035 (2002).

206. Voss, M. *et al.* Shedding of glycan-modifying enzymes by signal peptide peptidase-like 3 (<scp>SPPL</scp> 3) regulates cellular N-glycosylation. *EMBO J.* **33**, 2890–2905 (2014).
207. Cottam, N. P. *et al.* Dissecting Functions of the Conserved Oligomeric Golgi Tethering Complex Using a Cell-Free Assay. *Traffic* **15**, 12–21 (2014).
208. Pokrovskaya, I. D. *et al.* Conserved oligomeric Golgi complex specifically regulates the maintenance of Golgi glycosylation machinery. *Glycobiology* **21**, 1554–1569 (2011).
209. Tripathi, R., Hosseini, K., Arapi, V., Fredriksson, R. & Bagchi, S. SLC38A10 (SNAT10) is located in ER and Golgi compartments and has a role in regulating nascent protein synthesis. *Int. J. Mol. Sci.* **20**, (2019).
210. Scherer, H. U. *et al.* Glycan profiling of anti–citrullinated protein antibodies isolated from human serum and synovial fluid. *Arthritis Rheum.* **62**, 1620–1629 (2010).
211. 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).
212. Y, R. *et al.* Anti-citrullinated protein antibodies acquire a pro-inflammatory Fc glycosylation phenotype prior to the onset of rheumatoid arthritis. *Ann. Rheum. Dis.* **74**, 234–241 (2015).
213. Halapi, E., Gudbjartsson, D., … G. J.-E. journal of & 2010, undefined. A sequence variant on 17q21 is associated with age at onset and severity of asthma. *nature.com*
214. Li, X., Hastie, A., Hawkins, G., Allergy, W. M.- & 2015, undefined. eQTL of bronchial epithelial cells and bronchial alveolar lavage deciphers GWAS-identified asthma genes. *Wiley Online Libr.* **70**, 1309–1318 (2015).
215. Murphy, A., Chu, J., Xu, M., … V. C.-H. molecular & 2010, undefined. Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *academic.oup.com*
216. Cheng, H. D. *et al.* High-throughput characterization of the functional impact of IgG Fc glycan aberrancy in juvenile idiopathic arthritis. *Glycobiology* **27**, 1099–1108 (2017).
217. Gińdzieńska-Sieśkiewicz, E. *et al.* Changes of glycosylation of IgG in rheumatoid arthritis patients treated with methotrexate. *Adv. Med. Sci.* **61**, 193–197 (2016).

218. Alloza, I. *et al.* ANKRD55 and DHCR7 are novel multiple sclerosis risk loci. *Genes Immun. 2012 133* **13**, 253–257 (2011).
219. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
220. Fortune, M. D. *et al.* Statistical Colocalization of Genetic Risk Variants for Related Autoimmune Diseases in the Context of Common Controls. *Nat. Genet.* **47**, 839 (2015).
221. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981 (2012).
222. A, Z. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, (2011).
223. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
224. Sherva, R. *et al.* Genome Wide Association Study of the Rate of Cognitive Decline in Alzheimer's Disease. *Alzheimers. Dement.* **10**, 45–52 (2014).
225. V, B. & AJ, B. Spectrin and ankyrin-based pathways: metazoan inventions for integrating cells into tissues. *Physiol. Rev.* **81**, 1353–1392 (2001).
226. Rose-John, S. Interleukin-6 family cytokines. *Cold Spring Harb. Perspect. Biol.* **10**, a028415 (2018).
227. Fond, G., Lançon, C., Korchia, T., Auquier, P. & Boyer, L. The Role of Inflammation in the Treatment of Schizophrenia. *Front. Psychiatry* **11**, 160 (2020).
228. Mueller, T. M. & Meador-Woodruff, J. H. Post-translational protein modifications in schizophrenia. *NPJ Schizophr.* **6**, (2020).
229. Bort, R., Martinez-Barbera, J. P., Beddington, R. S. P. & Zaret, K. S. Hex homeobox gene-dependent tissue positioning is required for organogenesis of the ventral pancreas. *Development* **131**, 797–806 (2004).
230. Lemmers, R. F. H. *et al.* IgG glycan patterns are associated with type 2 diabetes in independent European populations. *Biochim. Biophys. Acta - Gen. Subj.* **1861**, 2240–2249 (2017).
231. Gudelj, I., Lauc, G. & Pezer, M. Immunoglobulin G glycosylation in aging and diseases. *Cell. Immunol.* **333**, 65–79 (2018).

8. SUPPLEMENTARY MATERIAL

8.1 Abbreviations

ABF - approximate Bayes factor

ADCC - antibody-dependent cell-mediated cytotoxicity

AOA - adult-onset asthma

Asn - Asparagine

AUC - Area Under the Curve

bp - base pair

CADD - Combined Annotation Dependent Depletion

ACN - acetonitrile

CD - Crohn's disease

CDG - congenital disorder of glycosylation

COJO - conditional and joint analysis

DHS - DNase I hypersensitive sites

DICE - database od immune cell eQTLs

EA - effect allele

EAF - effect allele frequency

EGCUT - Estonian's Genome Center of the University of Tartu

ENCODE - Encylopedia of DNA elements

EPIC - European Prospective Investigation into Cancer and Nutrition

eQTL - expression quantitative trait loci

ER - endoplasmic reticulum

F - core fucose

Fab - fragment antigen-binding

Fc - fragment crystallizable

FDR - False Discovery Rate

G - galactose

GA - Golgi Apparatus

GC - genomic control

GlcNAc - *N*-acetylglucosamine

GWAS - genome-wide association study

HDL - high-density lipoprotein

HILIC-SPE - hydrophilic interaction liquid chromatography solid phase extraction

HRC - Haplotype Reference Consortium

HWE - Hardy-Weinberg equilibrium

IBD - inflammatory bowel disease

IgG - Immunoglobulin G

IVIG - Intravenous immunoglobulin

JIA - juvenile idiopathic arthritis

KORA - Kooperative Gesundheitsforschung in der Region Augsburg

LD - linkage disequilibrium

LDSC – linkage disequilibrium score regression

LLS - Leiden Longevity Study

MAC - minor allele count

MAF - minor allele frequency

MAGMA - Multi-marker Analysis of GenoMic Annotation

MHC - major histocompatibility complex

MS - mass spectrometry

N - bisecting *N*-acetylglucosamine

NK - natural killer cells

ORCADES - The Orkney Complex Disease study

PBC - primary biliary cirrhosis

PPI - protein-protein interaction

Pro - Proline

QC - quality control

RA - rheumatoid arthritis

RF - response factor

SE - standard error

SE - standard error

SLE - systemic lupus erythematosus

SNP - single nucleotide polymorphism

T1DM - Type 1 diabetes mellitus

T2D - type 2 diabetes

TF - transcription factor

Thr - Threonine

UC - Ulcerative colitis

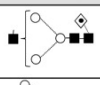
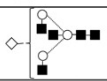
UPLC - Ultra-performance liquid chromatography

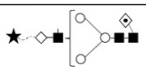
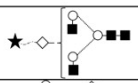
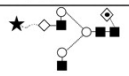
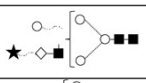
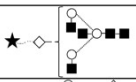
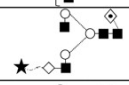
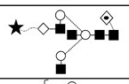
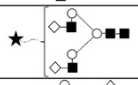
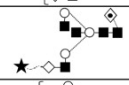
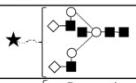
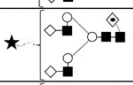
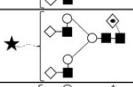
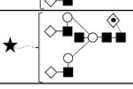
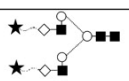
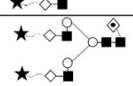
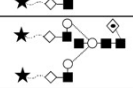
VEP - Variant Effect Predictor

WBC - white blood cell

8.2 Supplementary tables

Supplementary Table 1: IgG N-glycome composition measured by Ultra performance liquid chromatography. Table from Pučić *et al*[5]

Glycan peak	Peak composition	Structure	%
GP1	F(6)A1		100
GP2	A2		100
GP3	A2B		100
GP4	F(6)A2		100
GP5	M5		63
	F(6)A2		37
GP6	F(6)A2B		97
	A2[6]G1		3
GP7	A2[3]G1		75
	F(6)A2B		25
GP8a	A2BG1		93
	F(6)A2[6]G1		7
GP8b	F(6)A2[6]G1		100
GP9	F(6)A2[3]G1		100
GP10	F(6)A2[6]BG1		100
GP11	F(6)A2[3]BG1		100
GP12	A2G2		91
	F(6)A2[3]BG1		9
GP13	A2BG2		87
	F(6)A2G2		13
GP14	F(6)A2G2		100

Glycan peak	Peak composition	Structure	%
GP15	F(6)A2BG2		83
	F(6)A1G1S1		8
	A2G1S1		5
	F(6)A2G2		4
GP16a	F(6)A2[6]G1S1		63
	M4A1G1S1		25
	A2BG1S1		13
GP16b	F(6)A2[3]G1S1		91
	F(6)A2[6]BG1S1		9
GP17	A2G2S1		89
	F(6)A2[3]BG1S1		11
GP18a	A2BG2S1		91
	F(6)A2G2S1		9
GP18b	F(6)A2G2S1		100
GP19	F(6)A2BG2S1		100
GP20	n.d.		/
GP21	A2G2S2		100
GP22	A2BG2S2		100
GP23	F(6)A2G2S2		100
GP24	F(6)A2BG2S2		100

Supplementary Table 2: Glycan names and description of the most abundant glycan structure in each peak.

Glycan name	Description
GP1	FA1 glycan
GP2	A2 glycan
GP4	FA2 glycan
GP5	M5 glycan
GP6	FA2B glycan
GP7	A2G1 glycan
GP8	FA2[6]G1 glycan
GP9	FA2[3]G1 glycan
GP10	FA2[6]BG1 glycan
GP11	FA2[3]BG1 glycan
GP12	A2G2 glycan
GP13	A2BG2 glycan
GP14	FA2G2 glycan
GP15	FA2BG2 glycan
GP16	FA2G1S1 glycan
GP17	A2G2S1 glycan
GP18	FA2G2S1 glycan
GP19	FA2BG2S1 glycan
GP20	Structure not determined
GP21	A2G2S2 glycan
GP22	A2BG2S2 glycan
GP23	FA2G2S2 glycan
GP24	FA2BG2S2 glycan

Supplementary Table 3: List of glycan structures quantified with LC-MS.

	Glycan name	Glycan trait	Description
IgG1 subclass glycans	LC_IGP1	IgG1_G0F	FA2 glycan
	LC_IGP2	IgG1_G1F	FA2G1 glycan
	LC_IGP3	IgG1_G2F	FA2G2 glycan
	LC_IGP4	IgG1_G0FN	FA2B glycan
	LC_IGP5	IgG1_G1FN	FA2BG1 glycan
	LC_IGP6	IgG1_G2FN	FA2BG2
	LC_IGP7	IgG1_G1FS1	FA2G1S1 glycan
	LC_IGP8	IgG1_G2FS1	FA2G2S1 glycan
	LC_IGP9	IgG1_G1FNS1	FA2BG1S1 glycan
	LC_IGP10	IgG1_G2FNS1	FA2BG2S1 glycan
	LC_IGP11	IgG1_G0	A2 glycan
	LC_IGP12	IgG1_G1	A2G1 glycan
	LC_IGP13	IgG1_G2	A2G2 glycan

	LC_IGP14	IgG1_G0N	A2B glycan
	LC_IGP15	IgG1_G1N	A2BG1 glycan
	LC_IGP16	IgG1_G2N	A2BG2 glycan
	LC_IGP17	IgG1_G1S1	A2G1S1 glycan
	LC_IGP18	IgG1_G2S1	A2G2S1 glycan
	LC_IGP19	IgG1_G1NS1	A2BG1S1 glycan
	LC_IGP20	IgG1_G2NS1	A2BG2S1 glycan
IgG2/3 subclass glycans	LC_IGP87	IgG23_G0F	FA2 glycan
	LC_IGP88	IgG23_G1F	FA2G1 glycan
	LC_IGP89	IgG23_G2F	FA2G2 glycan
	LC_IGP90	IgG23_G0FN	FA2B glycan
	LC_IGP91	IgG23_G1FN	FA2BG1 glycan
	LC_IGP92	IgG23_G2FN	FA2BG2 glycan
	LC_IGP93	IgG23_G1FS1	FA2G1S1 glycan
	LC_IGP94	IgG23_G2FS1	FA2G2S1 glycan
	LC_IGP95	IgG23_G1FNS1	FA2BG1S1 glycan
	LC_IGP96	IgG23_G2FNS1	FA2GG2S1 glycan
	LC_IGP97	IgG23_G0	A2 glycan
	LC_IGP98	IgG23_G1	A2G1 glycan
	LC_IGP99	IgG23_G2	A2G2 glycan
	LC_IGP100	IgG23_G0N	A2B glycan
	LC_IGP101	IgG23_G1N	A2BG1 glycan
	LC_IGP102	IgG23_G2N	A2BG2 glycan
	LC_IGP103	IgG23_G1S1	A2G1S1 glycan
	LC_IGP104	IgG23_G2S1	A2G2S1 glycan
	LC_IGP105	IgG23_G1NS1	A2BG1S1 glycan
	LC_IGP106	IgG23_G2NS1	A2BG2S1 glycan
IgG4 subclass glycans	LC_IGP173	IgG4_G0F	FA2 glycan
	LC_IGP174	IgG4_G1F	FA2G1 glycan
	LC_IGP175	IgG4_G2F	FA2G2 glycan
	LC_IGP176	IgG4_G0FN	FA2B glycan
	LC_IGP177	IgG4_G1FN	FA2BG1 glycan
	LC_IGP178	IgG4_G2FN	FA2BG2 glycan
	LC_IGP179	IgG4_G1FS1	FA2G1S1 glycan
	LC_IGP180	IgG4_G2FS1	FA2G2S1 glycan
	LC_IGP181	IgG4_G1FNS1	FA2BG1S1 glycan
	LC_IGP182	IgG4_G2FNS1	FA2BG2S1 glycan

Supplementary Table 4: Trait list and formulas used for calculation of derived glycan traits

IgG N-glycan trait	Description	LCMS formula	UPLC formula
g0	Percentage of agalactosylated structures in total IgG N-glycome	(LC_IGP1+LC_IGP4+LC_IGP11+LC_IGP14+LC_IGP87+LC_IGP90+LC_IGP97+LC_IGP100+LC_IGP173+LC_IGP176)/SUM(ALL)*100	SUM(GP1:GP6)/SUM(GP1:GP24)*100
g1	Percentage of monogalactosylated structures in total IgG N-glycome	(LC_IGP2+LC_IGP5+LC_IGP7+LC_IGP9+LC_IGP12 + LC_IGP15 + LC_IGP17 + LC_IGP19+LC_IGP88 + LC_IGP91 + LC_IGP93+LC_IGP95 + LC_IGP98+ LC_IGP101+LC_IGP103+LC_IGP105+LC_IGP174+LC_IGP177+LC_IGP179+LC_IGP181)/SUM(ALL)*100	(GP7+GP8+GP9+GP10+GP11+GP16)/SUM(GP1:GP24)*100
g2	Percentage of digalactosylated structures in total IgG N-glycome	(LC_IGP3+LC_IGP6+LC_IGP8+LC_IGP10+LC_IGP13+LC_IGP16+LC_IGP18+LC_IGP20+LC_IGP89+LC_IGP92+LC_IGP94+LC_IGP99+LC_IGP102+LC_IGP104+LC_IGP106+LC_IGP175+LC_IGP178+LC_IGP180+LC_IGP182)/SUM(ALL)*100	(GP12+GP13+GP14+GP15+GP17+GP18+GP19+GP21+GP22+GP23+GP24)/SUM(GP1:GP24)*100
gal_total	Percentage of mono- and digalactosylated structures in total IgG N-glycome	(LC_IGP2+LC_IGP3+LC_IGP5:LC_IGP10+LC_IGP12+LC_IGP13+LC_IGP15:LC_IGP20)+SUM(LC_IGP88+LC_IGP89+LC_IGP91:LC_IGP96+LC_IGP98+LC_IGP99+LC_IGP101+LC_IGP106)+SUM(LC_IGP174+LC_IGP175+LC_IGP177:LC_IGP182)/SUM(ALL)*100	(SUM(GP7:GP19)+ SUM(GP21:24))/SUM(GP1:GP24)*100
s1	Percentage of monosialylated structures in total IgG N-glycome	(SUM(LC_IGP7:LC_IGP10) + SUM(LC_IGP17:LC_IGP20)+SUM(LC_IGP93:LC_IGP96)+SUM(LC_IGP103:LC_IGP106)+SUM(LC_IGP179:LC_IGP182))/SUM(ALL)*100	SUM(GP16:GP19)/SUM(GP1:GP24)*100
s1_no_bis	Percentage of monosialylated structures without bisecting GlcNAc in total IgG N-glycome	(LC_IGP7+LC_IGP8+LC_IGP17+LC_IGP18+LC_IGP93+LC_IGP94+LC_IGP103+LC_IGP104+LC_IGP179+LC_IGP180)/SUM(ALL)*100	SUM(GP16:GP18)/SUM(GP1:GP24)*100
s1_g1	Ratio of monosialylated and monogalactosylated structures in IgG N-glycans	(SUM(LC_IGP7:LC_IGP10) + SUM(LC_IGP17:LC_IGP20)+SUM(LC_IGP93:LC_IGP96)+SUM(LC_IGP103:LC_IGP106)+SUM(LC_IGP179:LC_IGP182))/ (LC_IGP2+LC_IGP5+LC_IGP7+LC_IGP9+LC_IGP12 + LC_IGP15 + LC_IGP17 + LC_IGP19+LC_IGP88 + LC_IGP91 + LC_IGP93+LC_IGP95 + LC_IGP98+ LC_IGP101+LC_IGP103+LC_IGP105+LC_IGP174+LC_IGP177+LC_IGP179+LC_IGP181)*100	(SUM(GP16:GP19)/(GP7+GP8+GP9+GP10+GP11+GP16)*100
s1_g2	Ratio of monosialylated and digalactosylated structures in IgG N-glycans	(SUM(LC_IGP7:LC_IGP10) + SUM(LC_IGP17:LC_IGP20)+SUM(LC_IGP93:LC_IGP96)+SUM(LC_IGP103:LC_IGP106)+SUM(LC_IGP179:LC_IGP182))/(LC_IGP3+LC_IGP6+LC_IGP8+LC_IGP10+LC_IGP13+LC_IGP16+LC_IGP18+LC_IGP20+LC_IGP89+LC_IGP92+LC_IGP94+LC_IGP99+LC_IGP102+LC_IGP104+LC_IGP106+LC_IGP175+LC_IGP178+LC_IGP180+LC_IGP182) *100	SUM(GP16:GP19)/(GP12+GP13+GP14+GP15+GP17+GP18+GP19+GP21+GP22+GP23+GP24) *100

s1_gal_total	Ratio of monosialylated structures and galactosylated IgG N-glycans	(SUM(LC_IGP7:LC_IGP10) + SUM(LC_IGP17:LC_IGP20)+SUM(LC_IGP93:LC_IGP96)+SUM(LC_IGP103:LC_IGP106)+SUM(LC_IGP179:LC_IGP182))/(LC_IGP2+LC_IGP3+LC_IGP5:LC_IGP10+LC_IGP12+LC_IGP13+LC_IGP15:LC_IGP20)+SUM(LC_IGP88+LC_IGP89+LC_IGP91:LC_IGP96+LC_IGP98+LC_IGP99+LC_IGP101+LC_IGP106)+SUM(LC_IGP174+LC_IGP175+LC_IGP177:LC_IGP182)*100	SUM(GP16:GP19)/(SUM(GP7:GP19)+ SUM(GP21:GP24))*100
bisecting	Percentage of structures with bisecting GlcNAc in IgG N-glycome	(LC_IGP4+LC_IGP5+LC_IGP6+LC_IGP9+LC_IGP10+LC_IGP14+LC_IGP15+LC_IGP16+LC_IGP19+LC_IGP20+LC_IGP90+LC_IGP91+LC_IGP92+LC_IGP95+LC_IGP96+LC_IGP100+LC_IGP101+LC_IGP102+LC_IGP105+LC_IGP106+LC_IGP176+LC_IGP177+LC_IGP178+LC_IGP181+LC_IGP182)/SUM(ALL)*100	(GP3+GP6+GP10+GP11+GP13+GP15+GP19+GP22+GP24)/SUM(GP1:GP24)*100
fuc	Percentage of fucosylated structures in total IgG N-glycome	(SUM(LC_IGP1:LC_IGP10)+SUM(LC_IGP87:LC_IGP96)+SUM(LC_IGP173:LC_IGP182))/SUM(ALL)*100	(GP1+GP4+GP6+GP8+GP9+GP10+GP11+GP14+GP15+GP16+GP18+GP19+GP23+GP24)/SUM(GP1:GP24)*100

Supplementary Table 5: Descriptive statistics for derived glycans traits; median (minimum value-maximum value)

	EPIC	ORCADES	TwinsUK 1&2)	TwinsUK 3&4	CROATIA-Korcula 3	VIKING	EGCUT	CROATIA-Korcula 1	CROATIA-Korcula 2	CROATIA-Split	KORA F4	CROATIA-Vis
fuc	96.01 (85.11-98.17)	95.57 (81.89-97.74)	94.63 (85.72-97)	95.25 (82.95-97.77)	95.81 (90.11-98.05)	96.58 (84.13-98.33)	96.06 (90.23-98.88)	95.04 (81.57-98.27)	95.06 (82.17-98.3)	95.24 (86.26-98.41)	93.39 (80.91-97.79)	95.53 (85.64-98.07)
g0	27.99 (9.36-60.9)	23.74 (8.73-60.23)	26.4 (8.73-53.9)	24.63 (7.14-61.22)	30.05 (9.58-60.8)	27.26 (9.55-59.82)	36.47 (11.27-71.72)	35.21 (13.03-60.88)	35 (17.31-58.18)	32.9 (16.4-60.2)	41.83 (19.22-65.82)	35.87 (15.32-65.72)
g1	38.19 (25.73-45.72)	38.78 (25.98-51.96)	40.09 (28.1-49.61)	38.9 (27.11-48.21)	37.54 (25.24-47.89)	38.97 (25.71-46.22)	37.32 (19.79-46.12)	40.02 (29.4-49.4)	40.56 (29.06-46.44)	41.25 (29.18-46.15)	40.5 (24.27-47.56)	39.95 (23.09-47.43)
g2	32.88 (13.22-58.1)	36.17 (13.58-58.51)	32.25 (14.36-57.26)	34.91 (7.29-63.08)	31.18 (13.65-56.42)	32.53 (14.28-59.39)	25.85 (8.31-46.62)	23.97 (8.7-51.73)	23.75 (10.33-45.54)	25.22 (10.5-45.09)	17.35 (7.3-38.93)	23.78 (8.86-48.59)
gal_total	71.78 (38.96-90.37)	75.9 (39.56-90.83)	73.08 (45.81-90.66)	75 (38.7-91.7)	69.71 (38.98-89.84)	72.54 (40.06-90.07)	63.24 (28.09-88.54)	64.79 (39.12-86.97)	65 (41.82-82.69)	67.1 (39.8-83.6)	58.17 (34.18-80.78)	64.13 (34.28-84.68)
s1	14.38 (7.22-27.36)	17.2 (7.72-30.1)	15.24 (8.37-26.98)	15.88 (2.89-43.79)	14.35 (5.88-29.47)	14.9 (6.09-25.97)	12.68 (4.68-22.2)	17.24 (7.4-33.54)	14.51 (7.22-27.13)	14.9 (7.63-25.24)	10.6 (4.76-22.21)	17.24 (6.66-34.91)

s1_g1	37.3 (22.13-89.12)	43.64 (23.5-91.77)	37.56 (19.3-80.77)	40.2 (9.2-82.98)	37.95 (12.29-73.68)	37.75 (14.28-84.62)	34.11 (12.69-70.67)	42.89 (21.84-69.56)	35.64 (18.39-73.01)	35.98 (20.36-65.39)	26.33 (12.22-52.73)	43.28 (23.01-88.23)
s1_g2	44.02 (33.13-62.98)	48.06 (29.28-68.44)	47.9 (24.14-62.81)	45.67 (20.09-74.17)	46.23 (29.55-92.78)	46.1 (31.64-65.18)	49.14 (29.73-83.32)	71.13 (44.28-94.78)	61.48 (46.32-82.65)	59.12 (44.15-78.99)	61.58 (32.06-91.9)	72.68 (50.25-94.45)
s1_gal_tot	20.29 (14.72-32.63)	22.93 (14.32-33.96)	21.07 (10.92-30.06)	21.4 (7.46-48.18)	20.91 (9.08-39.48)	20.82 (9.84-29.39)	20.18 (8.89-30.25)	26.73 (16.22-40.67)	22.54 (13.83-32.8)	22.27 (14.6-31.56)	18.41 (8.95-29.2)	27.12 (16.69-42.79)
s1_no_bis	12.41 (5.87-25.64)	15.06 (6.61-28.3)	13.32 (7.02-25.05)	14.08 (2.54-39.14)	12.56 (5.62-27.22)	13.17 (3.4-24.28)	10.92 (4.53-21.45)	16.37 (6.8-32.21)	13.5 (6.61-25.81)	13.95 (6.96-24.14)	9.15 (3.48-19.82)	16.37 (6.1-33.45)
bisecting	16.87 (10.13-30.66)	16.47 (8.65-28.28)	18.97 (9.57-27.59)	16.65 (8.6-30.27)	17.65 (5.17-44.15)	15.24 (8.23-29.31)	18.22 (6.11-29.89)	17.78 (10.16-29.48)	20.69 (11.33-32.77)	19.73 (11.09-32.14)	18.18 (9.9-31.72)	17.65 (10.51-33.42)

Supplementary Table 6: List of participating studies with names of PIs and analysts who performed GWAS

Cohort	**Principal investigator**	**Analyst**
TwinsUK	Tim Spector	Massimo Mangino
EPIC	Mathias Schulze	Rafael Cuadrat
CROATIA-Korcula	Caroline Hayward, Veronique Vitart	Azra Frkatović
CROATIA-Split	Caroline Hayward, Veronique Vitart	Azra Frkatović
CROATIA-Vis	Caroline Hayward, Veronique Vitart	Azra Frkatović
VIKING	Jim Wilson	Azra Frkatović
ORCADES	Jim Wilson	Azra Frkatović
LLS	Eline Slagboom	Erik van der Akker
KORA F4	Christian Gieger	Sapna Sharma
EGCUT	Andres Metspalu	Toomas Haller

Supplementary Table 7: Summary of file-level QC in participating studies. SNPs In- number of SNPs prior to QC; SNPs out- number of SNPs after QC; Invalid SE- number of SNPs excluded due to invalid standard error value; Invalid BETA- number of SNPs excluded due to invalid effect value; Monomorph SNPs- number of SNPs excluded due to EAF=0 or EAF=1, MAC ≤ 6- number of SNPs excluded because minor allele count ≤ 6; Low Info- number of SNPs excluded due to low imputation quality; AF outliers- number of SNPs excluded due to outlying SNPs in comparison to the reference dataset; Λ GC- genomic control inflation factor

Cohort	SNPs In	SNPs Out	Invalid SE	Invalid BETA	Monomorph SNPs	MAC ≤ 6	Low Info	Allele mismatch	AF outliers	λ GC
TwinsUK (batch1&2)	34483137	9411938	17251726	161060	1986173	5592342	35703	20326	11	1.00
LLS	39117105	4835454	6465465	0	0	1726588	0	12441	539834	1.01

TwinsUK (batch3&4)	38446170	14255310	13355020	324888	10246305	0	183555	26697	8	1.01
EGCUT	10538731	7154563	7249	2985	0	2954049	417176	16570	79	1.03
CROATIA -Korcula 1	11615984	9420199	4464	659	0	2185707	4615	19199	174	1.00
CROATIA -Korcula 2	11603767	8881758	123791	11649	0	2571515	14812	18220	154	1.00
CROATIA -Korcula 3	11744080	9253093	100460	13712	0	2350131	26308	18911	411	1.01
ORCADES	12455327	10443725	51891	5211	17929	1934850	590	21016	90	1.00
CROATIA -Split	11347869	9233198	4757	401	0	2101184	882	20151	38	1.00
VIKING	13502001	9904718	194658	23634	44	3369300	9069	18845	129	1.00
CROATIA -Vis	12345343	9130580	44956	6218	0	3144107	19187	18299	19	0.98
EPIC (subset 1)	39131578	7993934	0	0	0	2475531	34805	12353	26	1.01
EPIC (subset 2)	39127678	10250017	0	0	0	3566700	82760	15349	66	1.01
EPIC (subset 3)	39127678	8372849	0	0	0	2476641	79048	12774	4	0.99
EPIC (subset 4)	39127678	7021984	0	0	0	2062457	42103	10953	2	0.99
KORA F4	20023742	9739164	0	0	0	3352462	240487	0	1058805	1.00

Supplementary Table 8: Gene table with evidence for prioritization listed. No- order of genomic locus; Top SNP position- position of the SNP with the lowest p-value for the association in the defined region; PosMap- yes or no for positional mapping; eQTL mapping based on CEDAR and Fairfax datasets; Coloc PP4- posterior probability for shared causal variant between glycan trait and expression in whole blood eQTL dataset eQTLgen; VEP- score for functional consequences as obtained by SIFT and Polyphen algorithms; CADD score- score for deleteriousness by CADD; MAGMA- p-value obtained by genomewide gene-based association test; other- any other evidence taken into account when prioritizing genes. Prioritized genes are depicted in bold.

no	Top SNP position	Gene	Pos Map	eQTL mapping (FDR< 0.05)	Coloc PP4	VEP (chr:pos:A1:A2 (aa change); SIFT score; Polyphen score)	CADD score	MAGMA (p-value)	other
1	1:25291697	**RUNX3**	yes			1:25291010:A:T (N/I)	1:25291010:A:T (25.2)	5.90E-14	previous _studies
2	1:39302020	**RRAGC**	yes					2.26E-09	
		MYCBP	yes		0.994			4.22E-07	
		GJA9	yes						
		RHBDL2	yes						
3	1:233723112	**KCNK1**	yes						
4	2:26139430	ASXL2	yes						
		KIF3C	yes		0.900				
5	2:101991907	**RNF149**	yes	Fairfax_B_cells; Fairfax_monocytes; Fairfax_naive_monocytes;					
		CREG2	yes						
		RFX8	no					2.10E-07	
6	2:158469050	**ACVR1C**	yes		0.928		2:158415564:A:G (15.29)	8.41E-11	
7	3:186727170	**ST6GAL1**	yes	CEDAR_T_cells;Fairfax_B_cells; Fairfax_monocytes; Fairfax_naive_monocytes				5.45E-158	
8	4:103519487	**NFKB1**	yes	CEDAR_neutrophils; Fairfax_monocytes	0.930	4:103423326:T:G (splice donor)	4:103422504:C:G (19.81)	1.43E-13	
		MANBA	yes	CEDAR_B_cells; CEDAR_monocytes; CEDAR_T_cells; Fairfax_monocytes				2.31E-07	
		CISD2	no	Fairfax_B_cells					
9	5:55438851	**ANKRD55**	yes	CEDAR_T_cells	0.975				
		IL6ST	no		0.957				
10	5:95240996	**ELL2**	yes		0.932	5:95236459:T:C (A/T) (SIFT=0.66; Polyphen=0.017)	5:95245384:C:T (18.39)	3.20E-12	
11	6:22053674	**CASC15**	yes						
12	6:31351764	**HLA region**							
13	6:74230859	MB21D1	yes	Fairfax_naive_monocytes					
		MTO1	yes	CEDAR_monocytes;CEDAR_T_cells; Fairfax_monocytes; Fairfax_naive_monocytes	0.960			5.68E-09	

		EEF1A1	yes	CEDAR_B_cells;Fairfax_B_cells CEDAR_monocytes;CEDAR_T_cells; Fairfax_monocytes; Fairfax_naive_monocytes	0.966		6:74230859:C:T (15.74)	1.18E-10	
14	6:139629524	**TXLNB**	yes	DICE_B_cell_naive				2.65E-08	
15	6:143169723	**HIVEP2**	yes				6:143193344:C:T (19.67)	1.30E-19	
16	7:6531268	**DAGLB**	yes	CEDAR_T_cells; Fairfax_B_cells				1.67E-07	
		KDELR2	yes	CEDAR_B_cells;CEDAR_neutrophils; CEDAR_monocytes;CEDAR_T_cells; Fairfax_B_cells;Fairfax_monocytes; Fairfax_naive_monocytes	0.970			1.38E-11	
		FLJ20306	yes					5.68E-12	
		GRID2IP	yes					7.61E-08	
17	7:50352695	**IKZF1**	yes				7:50333960:C:T (21.1)	2.43E-27	
18	7:150942349	**ABCF2**	yes					4.59E-13	
		CHPF2	yes					8.03E-11	
		SMARCD3	yes					1.86E-16	
19	8:103545983	KB-1980E6.3	yes					6.35E-12	
		UBR5	no						3D chromatin mapping
		RRM2B	no						3D chromatin mapping
		ODF1	no						3D chromatin mapping; previous studies
20	9:33124872	**B4GALT1**	yes	DICE_B_cell_naive			9:33164527:A:T (15.41)	1.36E-81	
		DNAJA1	yes				9:33039024:A:T (19.57)		
21	10:94446635	IDE	yes						
		KIF11	yes		0.990				
		HHEX	yes	Fairfax_monocytes					
22	11:65555524	**AP5B1**	yes						
		BANF1	no	Fairfax_naive					
		OVOL1	yes						
23	11:114381448	**REXO2**	yes	Fairfax_B_cells				1.17E-09	
		NXPE1	yes			11:114401611:A:G (S/L)		1.08E-13	
		NXPE4	yes			11:114442103:A:G (H/Y)	11:114442103:A:G (15.1)	1.10E-14	
24	12:121202664	**SPPL3**	yes				12:121202362:C:T (18.48)	3.06E-12	

25	14:65775695	**FUT8**	yes	CEDAR_monocytes; Fairfax_monocytes		14:66082793:A:C (Q/K); 14:66136163:A:C (T/K) (SIFT=0.02; Polyphen=0.517)	14:66082793:A:C (22.3)	8.33E-98	
		PTBP1	no		0.985				
		ESR2	yes						
26	14:106113281	TEX22	yes					2.14E-08	
		MTA1	yes					1.26E-09	
		CRIP2	yes						
		CRIP1	yes			14:105954705:T:C (A/V)			
		TMEM121	yes				14:105996049:T:TGCC (18.72)	2.06E-14	
		IGHG2	no		0.983	14:106110914:T:G (P/T) (SIFT=0.06;Polyphen=0.577); 14:106110137:T:C (V/M) (SIFT=0.08;Polyphen=0.329)			
		IGHA1	no			14:106174261:C:G (D/E) (Polyphen=0.508)			
		IGHG1	no			14:106208086:A:C (E/D); 14:106208082:G:T (M/L); 14:106209119:T:C (R/K)			
		IGHG3	no			14:106235767:T:C (S/N) (SIFT=1; Polyphen=0); 14:106236128:T:A (F/Y); 14:106236143:A:G (P/L) (SIFT=0.12; Polyphen=0.031)			
27	16:23412310	SCNN1B	yes						
		COG7	yes		0.789		16:23422672:C:G (17.17)	1.74E-14	
		GGA2	yes	CEDAR_T_cells;Fairfax_monocytes; DICE_B_cell_naive		16:23489711:C:G (P/A)	16:23521643:C:G (15.33)	4.36E-12	
		EARS2	yes			16:23536684:T:C (G/S)		3.93E-10	
		UBFD1	yes	Fairfax_monocytes					
		NDUFAB1	yes					3.66E-08	
		PALB2	yes						
		DCTN5	no	CEDAR_B_cells;Fairfax_B_cells; CEDAR_neutrophils;CEDAR_T_cells; Fairfax_monocytes; Fairfax_naive_monocytes	0.845				
28	17:16842991	**TNFRSF13B**	yes			17:16842991:A:G(P/L)(SIFT=0.23;PolyPhen=0.476)		1.30E-12	
29	17:38072727	**GSDMB**	yes	CEDAR_B_cells; Fairfax_B_cells		17:38062217:T:C(G/R)(SIFT=0; PolyPhen=0.999);17:38062196:A		8.09E-26	previous _studies

						:G (P/S); 17:38064469:T:C (splice variant)			
		ORMDL3	yes	CEDAR_B_cells; CEDAR_T_cells	0.787		17:38082807:C:T (20.7)	6.10E-18	previous _studies
		IKZF3	yes	CEDAR_T_cells; Fairfax_B_cells			17:37922259:A:G (19.78)		previous _studies
		ZPBP2	yes	Fairfax_B_cells		17:38028634:T:G (S/I)(SIFT=0.08;PolyPhen=0.26)	17:38029120:C:G (18.2)	2.84E-24	previous _studies
		PGAP3	yes	CEDAR_B_cells;CEDAR_T_cells; Fairfax_B_cells			17:37831297:C:CCCCA (17.97)	9.46E-13	
		STARD3	yes	Fairfax_monocytes		17:37814080:A:G (R/Q) (SIFT=0.11;PolyPhen=0.637)	17:37814080:A:G (23.1)		
		LRRC3C	yes						
		GSDMA	yes	CEDAR_T_cells		17:38121993:A:G (R/Q) (SIFT=0.14; Polyphen=0.01); 17:38122686:A:G (E/K) (SIFT=0.64; Polyphen=0.062); 17:38131187:A:C (T/N) (SIFT=0.1; Polyphen=0.637)	17:38121993:A:G (21.9)	1.62E-17	
		MED24	yes				17:38179492:A:G (18.11)		
		PPP1R1B	yes						
		CDK12	yes	CEDAR_monocyte; CEDAR_T_cells					
		ERBB2	yes			17:37884037:C:G (A/P)	17:37884037:C:G (23.5)	3.62E-15	
30	17:44331214	**ARHGAP27**	yes	Fairfax_naive_monocytes		17:43507297:T:C (A/T) (SIFT=0.51; Polyphen=0.167	17:43507649:A:G (18.95)	1.67E-13	
		PLEKHM1	yes	Fairfax_monocytes				4.28E-14	
		CRHR1	yes			17:43902861:A:C (P/T) (SIFT=0.17;PolyPhen=0); 17:43910507:T:C (A/V) (SIFT=0); 17:43912159:C:G (E/Q) (Polyphen=0.003)	17:43902505:C:T (19.39)	7.46E-16	
		SPPL2C	yes			17:43923654:C:G (R/P) (SIFT=0.01;PolyPhen=0.442); 17:43923266:A:G (A/T) (SIFT=0.54; Polyphen=0.036)		4.03E-14	
		MAPT	yes			17:44061278:T:C (R/W) (SIFT=0;PolyPhen=0.903); 17:44055647:A:T(stop_codon_lost) */K	17:44061278:C:T (26.8)	5.37E-15	
		STH	yes			17:44076665:A:G (R/Q)		6.08E-17	
		KANSL1	yes	Fairfax_monocytes		17:44248837:T:C (D/N); 17:44117119:A:G (P/S)	17:44249199:G:T (26.4)	1.51E-16	
		ARL17B	yes	Fairfax_B_cells				6.41E-17	

		LRRC37A	yes	Fairfax_B_cells; CEDAR_neutrophils	17:44408004:C:A (S/R) (SIFT=0.04;PolyPhen=0.019)		
		LRRC37A2	yes	DICE_B_cell_naive	17:44625866:C:A (S/R) (SIFT=0.04;PolyPhen=0.018)		
		ARL17A	yes	CEDAR_monocytes; Fairfax_B_cells			
		ARL17B	no	Fairfax_B_cells			
		NSF	yes	CEDAR_B_cells; CEDAR_monocytes; CEDAR_neutrophils; CEDAR_T_cells; Fairfax_monocytes; Fairfax_naive_monocytes		17:44793503:A:G (17.18)	1.66E-17
		WNT3	yes	Fairfax_B_cells		17:44856641:C:G (16.08)	8.33E-26
31	17:45809822	KPNB1	yes				
		TBKBP1	yes	Fairfax_B_cells;CEDAR_monocyte; CEDAR_neutrophil;CEDAR_T_cells; Fairfax_monocytes; Fairfax_naive_monocytes		17:45772447:A:G (18.46)	
		TBX21	yes			17:45809822:A:G (16.69)	1.87E-12
		ITGB3	no				6.06E-09
		ITGB3	no				6.06E-09
32	17:56410041	**BZRAP1**	yes				2.90E-07
		SUPT4H1	yes	Fairfax_B_cells; Fairfax_naive_monocytes			
		MPO	no	Fairfax_monocytes; Fairfax_naive_monocytes			
		hsa-mir-142	no	0.999			
		RAD51C	no	CEDAR_monocyte;Fairfax_B_cells; Fairfax_monocytes; Fairfax_naive_monocytes			
33	17:79218714	AATK	yes				3.52E-08
		AZI1	yes			17:79170576:C:T (15.28)	5.81E-30
		ENTHD2	yes	CEDAR_neutrophil; Fairfax_monocytes; Fairfax_naive_monocytes			1.18E-29
		C17orf89	yes				2.69E-21
		SLC38A10	yes	CEDAR_monocyte; CEDAR_neutrophils; Fairfax_monocytes; Fairfax_naive_monocytes	17:79220224:C:G (A/G) (SIFT=0.06;PolyPhen=0.825)		7.24E-27
		TMEM105	yes				
		GPS1	no	CEDAR_T_cells			

34	19:1657741	MEX3D	yes					
		MBD3	yes					
		UQCR11	yes					
		UQCR11	yes					
		TCF3	yes		0.962			1.42E-14
35	19:19294091	**MEF2B**	yes		0.973			6.98E-13
		MEF2BNB	yes					9.87E-13
		RFXANK	yes					
		MAU2	no	Fairfax_B_cells; Fairfax_monocytes				
36	20:4115720	**SMOX**	yes					
37	20:17831618	**RRBP1**	no	Fairfax_B_cells				
38	20:50077482	**NFATC2**	yes				20:50064221:A:G (16.53)	6.59E-06
39	20:61598731	DIDO1	yes					
		GID8	yes					
		SLC17A9	yes			20:61598731:C:T(T/M) (SIFT=0.01;PolyPhen=0.788)	20:61598731:C:T (24.2)	
		BHLHE23	yes					
		TCFL5	no	Fairfax_naive_monocytes; Fairfax_monocytes				
40	21:36564553	**RUNX1**	yes				21:36561598:G:T (20.7)	6.58E-26
41	22:24179922	**DERL3**	yes	Fairfax_B_cells		22:24179922:C:G (F/L) (SIFT=0.04;PolyPhen=0.246)	22:24179922:C:G (24.9)	3.26E-14
		RGL4	yes					
		ZNF70	yes					
		VPREB3	yes	Fairfax_B_cells				5.36E-16
		SLC2A11	yes					
		CHCHD10	yes	Fairfax_naive_monocytes				1.11E-23
		MMP11	yes					
		SMARCB1	yes	CEDAR_monocytes;Fairfax_naive_monocytes;Fairfax_monocytes				1.16E-15
42	22:39845898	**MGAT3**	yes	CEDAR_B_cells; Fairfax_B_cells	0.960			2.48E-89
		SYNGR1	yes	Fairfax_B_cells;Fairfax_monocytes; Fairfax_naive_monocytes; CEDAR_monocytes		22:39770597:A:G (A/T)	22:39777822:C:CCAA (16.86)	2.05E-64
		ATF4	yes	CEDAR_T_cells; Fairfax_monocytes		22:39917515:A:C (Q/P) (SIFT=0.25;PolyPhen=0.001)	22:39916626:T:TC (21.2)	
		RPL3	yes				22:39712981:A:G (22)	1.12E-12

Supplementary Table 9: List of traits and diseases used in colocalization analysis. Trait- tested trait or disease; #1 author- first author of the publication describing the GWAS of the trait or disease; Accession No- accession number in GWAS Catalog; N total- total number of subjects in the study; cases- number of case subjects in the study; controls- number of control subjects in the study; Download link- link for download of the GWAS summary statistics for the given trait or disease

Trait	#1 author	Accession No	N total	Cases	Controls	Download link
Adult-onset asthma	Ferreira MAR et al	GCST007799	327253	26582	300671	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/FerreiraMAR_30929738_GCST007799
Primary billiary cirrhosis	Cordell HJ	GCST003129	13239	2764	10475	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/CordellHJ_26394269_GCST003129/harmonised/26394269-GCST003129-EFO_1001486-Build37.f.tsv.gz
Asthma	Han Y	GCST010042	303859	64538	239321	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST010001-GCST011000/GCST010042/
Systemic Lupus Erythematosus	Julia A	GCST005831	16966	4943	8483	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/JuliaA_29848360_GCST005831
Type I diabetes	Forgetta V	GCST010681	24840	9266	15574	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/ForgettaV_32005708_GCST010681
IgG level	Scepanovic P	GCST006357	1000	-	-	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/ScepanovicP_30053915_GCST006357
Schizophrenia	Pardinas AF	GCST006803	105318	40675	64643	https://walters.psycm.cf.ac.uk/clozuk_pgc2.meta.sumstats.txt.gz
Rheumatoid arthritis	Eyre S	GCST005569	47580	13838	33742	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/EyreS_23143596_GCST005569
Type II diabetes	Mahajan A	GCST007518	298957	48286	250671	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MahajanA_29632382_GCST007518/T2D_European.BMIadjusted.txt
Allergic Disease	Ferreira	GCST005038	360838	180129	180709	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/FerreiraMA_29083406_GCST005038
Alzheimer's disease	Kunkle	NA	63926	21982	41944	https://www.niagads.org/datasets/ng00075
Lymphocyte percentage of WBC	Astle WJ	GCST004632	171748	-	-	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/AstleWJ_27863252_GCST004632
Total cholesterol	Willer CJ	GCST002221	94595	-	-	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST002001-GCST003000/GCST002221/harmonised/24097068-GCST002221-EFO_0004574-build37.f.tsv.gz
Crohn's disease	Liu JZ	GCST003044	20883	5956	14927	http://www.ibdgenetics.org/downloads.html
Ulcerative colitis	Liu JZ	GCST003045	27432	6968	20464	http://www.ibdgenetics.org/downloads.html
Inflammatory bowel disease	Liu JZ	GCST003043	34652	12882	21770	http://www.ibdgenetics.org/downloads.html
HDL cholesterol	Willer CJ	GCST002223	94595	-	-	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST002001-GCST003000/GCST002223/harmonised/24097068-GCST002223-EFO_0004612-build37.f.tsv.gz
LDL cholesterol	Willer CJ	GCST002222	94595	-	-	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST002001-GCST003000/GCST002222/harmonised/24097068-GCST002222-EFO_0004611-build37.f.tsv.gz

Juvenile idiopathic arthritis	Hinks A	GCST005528	15872	2816	13056	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST005001-GCST006000/GCST005528/harmonised/23603761-GCST005528-EFO_1001999-Build37.f.tsv.gz
Osteoarthritis	Tachmazidou I	GCST007092	417596	39427	378169	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST007001-GCST008000/GCST007092/harmonised/30664745-GCST007092-EFO_0002506-build37.f.tsv.gz
Chronic kidney disease	Wuttke M	GCST008065	625219	64164	561055	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST008001-GCST009000/GCST008065/CKD_overall_EA_JW_20180223_nstud23.dbgap.txt.gz
Hypertension	Zhu Z	GCST007610	458554	144793	313761	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST007001-GCST008000/GCST007610/ZhuZ_30940143_ukbb.bolt_460K_selfRepWhite.doctor_highbloodpressure.assoc.gz
Thyroid cancer	Zhou W	GCST008371	407757	358	407399	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST008001-GCST009000/GCST008371/PheCode_193_SAIGE_MACge20.txt.vcf.gz
Lung cancer	Rashkin SR	GCST90011812	412835	2485	410350	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90011001-GCST90012000/GCST90011812/harmonised/32887889-GCST90011812-EFO_0001071-Build37.f.tsv.gz
Ovarian cancer	Rashkin SR	GCST90011821	411609	1259	410350	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90011001-GCST90012000/GCST90011821/harmonised/32887889-GCST90011821-EFO_0001075-Build37.f.tsv.gz
Colorectal cancer	Zhou W	GCST008372	387318	4562	382756	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST008001-GCST009000/GCST008372/PheCode_153_SAIGE_MACge20.txt.vcf.gz
Breast cancer	Rashkin SR	GCST90011804	428231	17881	410350	http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90011001-GCST90012000/GCST90011804/harmonised/32887889-GCST90011804-EFO_0000305-Build37.f.tsv.gz

Supplementary Table 10: String PPI output network interactions. Node 1 and Node2- proteins involved in the interaction; Homology- score based on the homology evidence; Coexpression- score based on the evidence for the coexpression of the proteins; Experimentally determined- score based on the experimental evidence for the interaction of the two proteins; Database annotated- score for interaction based on the curated databases; combined score- final score for the interaction of two proteins as combined from all four evidence types

Node 1	**Node 2**	**Homology**	**Coexpression**	**Experimentally determined**	**Database annotated**	**Combined score**
B4GALT1	ST6GAL1	0	0	0	0.9	0.9
B4GALT1	MANBA	0	0.061	0	0.9	0.902
B4GALT1	FUT8	0	0	0	0.9	0.9
COG7	KDELR2	0	0	0	0.9	0.9
COG7	DCTN5	0	0.087	0	0.9	0.904
COG7	NSF	0	0.068	0	0.9	0.902
DCTN5	KDELR2	0	0.055	0	0.9	0.901
DCTN5	KIF11	0	0	0	0.9	0.9
DCTN5	KIF3C	0	0	0	0.9	0.9
EEF1A1	UBR5	0	0.047	0.421	0	0.425

ELL2	SUPT4H1	0	0	0	0.9	0.9
FUT8	MGAT3	0	0	0	0.9	0.9
IKZF1	IKZF3	0.941	0.31	0.725	0	0.802
KDELR2	KIF3C	0	0	0	0.9	0.9
KDELR2	NSF	0	0.061	0	0.9	0.902
KDELR2	KIF11	0	0.065	0	0.9	0.902
KIF11	KIF3C	0.736	0	0	0.9	0.9
MEF2B	NFATC2	0	0	0	0.6	0.6
NFATC2	RUNX1	0	0.065	0.135	0.9	0.912
NXPE1	NXPE4	0.969	0.555	0	0	0.555
RUNX1	TBX21	0	0	0.299	0.8	0.853
RUNX1	SMARCD3	0	0	0.109	0.9	0.907
RUNX1	TCF3	0	0	0.361	0.9	0.933
RUNX1	SMARCB1	0	0	0.308	0.9	0.927
RUNX1	RUNX3	0.956	0.065	0	0.9	0.902
RUNX3	TBX21	0	0.152	0.127	0.8	0.839
SMARCB1	SMARCD3	0	0.129	0.926	0.9	0.992

Supplementary Table 11: Phenoscanner output in novel genomic regions discovered in the GWAS. Ref_hg19_coord- chromosomal location of the reference SNP, ref_rsid- rsID for the reference SNP; trait- trait with the previously known association in the region; study- First author in the study describing the association.

ref_hg19_coord	**ref_rsid**	**trait**	**study**
chr11:65555524	rs10896045	Eosinophil count	Astle W
chr11:65555524	rs10896045	Sum eosinophil basophil counts	Astle W
chr11:65555524	rs10896045	Atopic dermatitis	EAGLE
chr11:65555524	rs10896045	Hayfever: allergic rhinitis or eczema	Neale B
chr11:65555524	rs10896045	No blood clot: bronchitis: emphysema: asthma: rhinitis: eczema or allergy diagnosed by the doctor	Neale B
chr11:65555524	rs10896045	Self-reported asthma	Neale B
chr10:94446635	rs10786052	Eosinophil count	Astle W
chr10:94446635	rs10786052	Eosinophil percentage of granulocytes	Astle W
chr10:94446635	rs10786052	Eosinophil percentage of white cells	Astle W
chr10:94446635	rs10786052	Neutrophil percentage of granulocytes	Astle W
chr10:94446635	rs10786052	Sum eosinophil basophil counts	Astle W
chr10:94446635	rs10786052	Type II diabetes	DIAGRAM
chr10:94446635	rs10786052	Type II diabetes adjusted for BMI	DIAGRAM
chr10:94446635	rs10786052	Birth weight	Neale B
chr10:94446635	rs10786052	Diabetes diagnosed by doctor	Neale B
chr10:94446635	rs10786052	Self-reported diabetes	Neale B
chr10:94446635	rs10786052	Inflammatory bowel disease	IBDGC
chr4:103403494	rs28882677	Lymphocyte count	Astle W
chr4:103403494	rs28882677	Lymphocyte percentage of white cells	Astle W
chr4:103403494	rs28882677	Monocyte percentage of white cells	Astle W

chr4:103403494	rs28882677	Hayfever: allergic rhinitis or eczema	Neale B
chr4:103407428	rs11097788	Lymphocyte count	Astle W
chr4:103407428	rs11097788	Lymphocyte percentage of white cells	Astle W
chr4:103407428	rs11097788	Neutrophil percentage of white cells	Astle W
chr4:103407428	rs11097788	Allergic disease	Ferreira M
chr4:103407428	rs11097788	Hayfever: allergic rhinitis or eczema	Neale B
chr4:103407428	rs11097788	log eGFR creatinine in non diabetics	CKDGen
chr4:103407428	rs11097788	log eGFR creatinine	CKDGen
chr4:103407428	rs11097788	Primary biliary cholangitis	Qiu F
chr4:103519487	rs3774964	Lymphocyte count	Astle W
chr4:103519487	rs3774964	Lymphocyte percentage of white cells	Astle W
chr4:103519487	rs3774964	Hayfever: allergic rhinitis or eczema	Neale B
chr4:103519487	rs3774964	Monocyte percentage of white cells	Astle W
chr4:103519487	rs3774964	Neutrophil percentage of white cells	Astle W
chr12:121202664	rs9431	Allergic disease	Ferreira M

Supplementary Table 12: Colocalization with diseases and traits. Genomic region- genomic locus which is significantly associated with at least one of the traits; N SNPs-number of SNPs in the region which is used in colocalization test; Trait-glycan trait with the strongest association in the genomic region; PP.H3 - posterior probability for H3- both traits are associated with the region but have different causal variants. PP.H4 posterior probability for H4- both traits are associated with the regions and have the same causal variant

Genomic region	**N SNPs**	**Glycan trait**	**Disease/Trait**	**PP.H3**	**PP.H4**
chr10:94336963-94495241	24	fuc	Type 2 Diabetes	0.001	0.998
chr17:56398006-56417002	1128	s1_no_bis	Alzheimers	0.004	0.996
chr10:94336963-94495241	1327	fuc	Adult-onset Asthma	0.005	0.995
chr17:56398006-56417002	938	s1_no_bis	Adult-onset Asthma	0.007	0.993
chr5:55436851-55444683	77	s1_g2	Juvenile Idiopathic Arthritis	0.001	0.993
chr10:94336963-94495241	1353	fuc	Asthma	0.016	0.984
chr21:36524140-36787961	34	bisecting	Juvenile Idiopathic Arthritis	0.002	0.958
chr5:55436851-55444683	1635	s1_g2	Rheumatoid Arthritis	0.043	0.957
chr17:37579383-38215117	1284	fuc	Ulcerative Colitis	0.052	0.948
chr17:37579383-38215117	1325	fuc	Inflammatory Bowel Disease	0.053	0.947
chr17:37579383-38215117	1243	fuc	Crohn's Disease	0.057	0.943
chr11:114298893-114450529	1561	s1_g2	Inflammatory Bowel Disease	0.053	0.943
chr17:37579383-38215117	1157	fuc	Rheumatoid Arthritis	0.059	0.941
chr7:50325563-50362999	1460	fuc	Asthma	0.048	0.929
chr11:114298893-114450529	1531	s1_g2	Ulcerative Colitis	0.094	0.906
chr17:43463493-44865603	225	s1_g2	Primary Biliary Cirrhosis	0.121	0.876
chr12:121188641-121351934	1526	s1_gal_total	Schizophrenia	0.119	0.862
chr17:37579383-38215117	272	fuc	Primary Biliary Cirrhosis	0.16	0.84
chr11:65555524-65555524	999	g1	Systemic Lupus Erythematosus	0.182	0.817
chr7:50325563-50362999	1624	fuc	Rheumatoid Arthritis	0.086	0.784
chr12:121188641-121351934	457	s1_gal_total	Primary Biliary Cirrhosis	0.021	0.776

chr21:36524140-36787961	1451	bisecting	Systemic Lupus Erythematosus	0.072	0.774
chr5:95217242-95324375	1322	s1_no_bis	Type 1 Diabetes	0.097	0.744
chr4:103390496-103554821	1381	g2	Ulcerative Colitis	0.129	0.725
chr14:105877057-106270813	358	s1_g2	Systemic Lupus Erythematosus	0.035	0.71
chr5:55436851-55444683	1278	s1_g2	Systemic Lupus Erythematosus	0.043	0.71
chr1:23526335-25903455	1520	bisecting	Crohn's Disease	0.033	0.701
chr1:39302020-39380385	2305	fuc	Type 1 Diabetes	0.06	0.665
chr5:95217242-95324375	1304	s1_no_bis	Inflammatory Bowel Disease	0.101	0.661
chr10:94336963-94495241	1593	fuc	Inflammatory Bowel Disease	0.139	0.657
chr12:121188641-121351934	1554	s1_gal_total	Adult-onset Asthma	0.355	0.645
chr17:16820099-16875636	1710	g1	IgG Level	0.1	0.632
chr14:105877057-106270813	313	s1_g2	IgG Level	0.077	0.62
chr17:43463493-44865603	814	s1_g2	Systemic Lupus Erythematosus	0.389	0.609
chr10:94336963-94495241	2168	fuc	Lymphocytes	0.382	0.584
chr17:43463493-44865603	908	s1_g2	Breast cancer	0.421	0.579
chr17:43463493-44865603	1265	s1_g2	Lymphocytes	0.428	0.572
chr5:55436851-55444683	1686	s1_g2	Crohn's Disease	0.053	0.571
chr7:6497501-6550403	2117	bisecting	Inflammatory Bowel Disease	0.15	0.57
chr12:121188641-121351934	1589	s1_gal_total	Asthma	0.438	0.562
chr11:65555524-65555524	1009	g1	Allergy	0.438	0.561
chr11:65555524-65555524	1029	g1	Asthma	0.438	0.56
chr17:43463493-44865603	669	s1_g2	Schizophrenia	0.374	0.552
chr7:6497501-6550403	1893	bisecting	Osteoarthritis	0.336	0.547
chr7:6497501-6550403	3169	bisecting	Lymphocytes	0.471	0.529
chr17:43463493-44865603	1044	s1_g2	Crohn's Disease	0.375	0.509

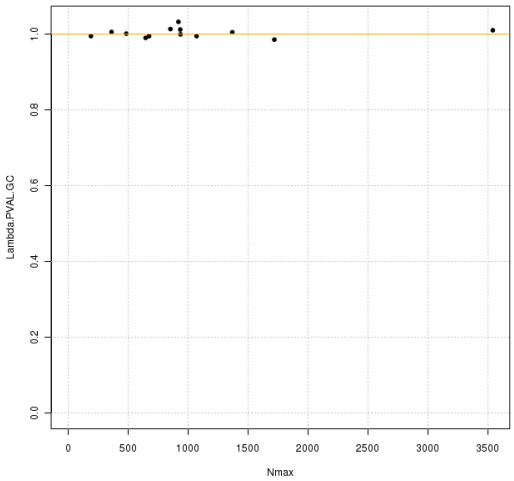
Supplementary Table 13: List of prioritized genes and their full names. Gene name- HGNC symbol of the gene, symbol in parentheses denote alias and formerly used name; Ensembl ID- gene ID used in Ensembl database; Entrez ID- gene ID used by NCBI database; Full name- full name for the gene

Gene name	**Ensembl ID**	**Entrez ID**	**Full name**
ABCF2	ENSG00000033050	10061	ATP binding cassette subfamily F member 2
ACVR1C	ENSG00000123612	130399	activin A receptor type 1C
ANKRD55	ENSG00000164512	79722	ankyrin repeat domain 55
AP5B1	ENSG00000254470	91056	adaptor related protein complex 5 subunit beta 1
ARHGAP27	ENSG00000159314	201176	Rho GTPase activating protein 27
CEP131(AZI1)	ENSG00000141577	22994	Centrosomal Protein 131
B4GALT1	ENSG00000086062	2683	beta-1 4-galactosyltransferase 1
TSPOAP1 (BZRAP1)	ENSG00000005379	9256	Benzodiazepine receptor (peripheral) associated protein 1
CASC15	ENSG00000272168	401237	cancer susceptibility 15
CHCHD10	ENSG00000250479	400916	coiled-coil-helix-coiled-coil-helix domain containing 10
CHPF2	ENSG00000033100	54480	chondroitin polymerizing factor 2
COG7	ENSG00000168434	91949	component of oligomeric golgi complex 7
CRHR1	ENSG00000120088	1394	corticotropin releasing hormone receptor 1

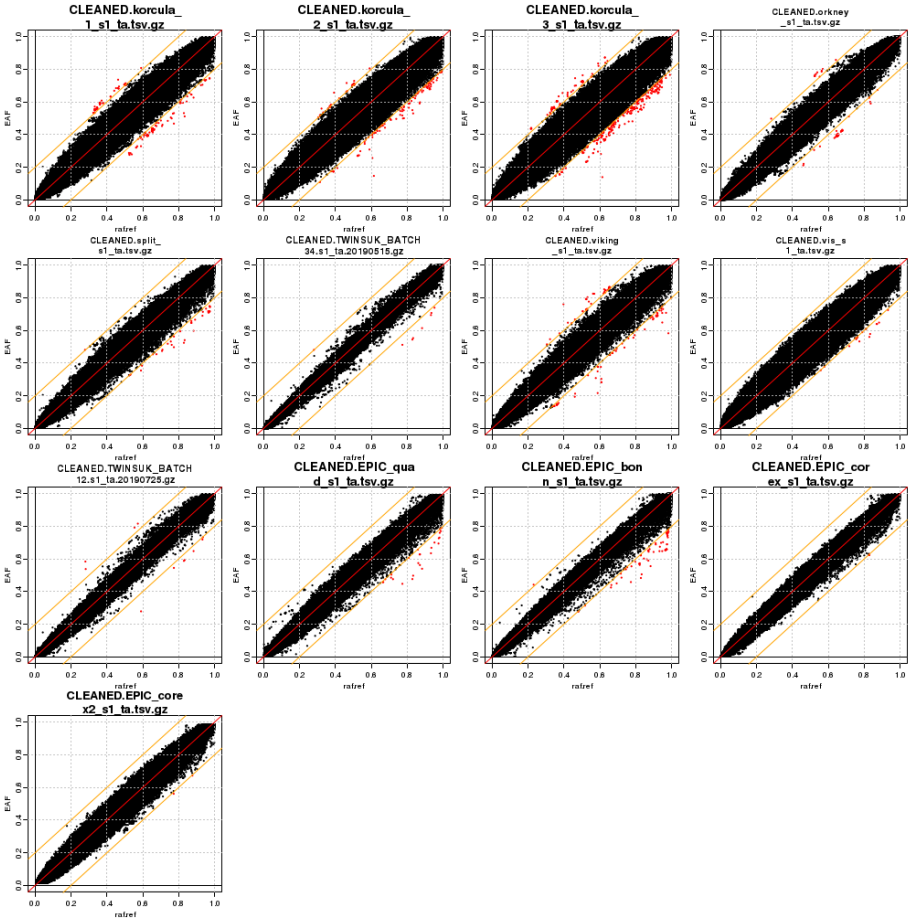
CRIP1	ENSG00000213145	1396	cysteine rich protein 1
DAGLB	ENSG00000164535	221955	diacylglycerol lipase beta
DCTN5	ENSG00000166847	84516	dynactin subunit 5
DERL3	ENSG00000099958	91319	derlin 3
EARS2	ENSG00000103356	124454	glutamyl-tRNA synthetase 2 mitochondrial
EEF1A1	ENSG00000156508	1915	eukaryotic translation elongation factor 1 alpha 1
ELL2	ENSG00000118985	22936	elongation factor for RNA polymerase II 2
ENTHD2	ENSG00000167302	146705	TEPSIN Adaptor Related Protein Complex 4 Accessory Protein
FUT8	ENSG00000033170	2530	fucosyltransferase 8
GGA2	ENSG00000103365	23062	Golgi associated gamma adaptin ear containing ARF binding protein 2
GSDMB	ENSG00000073605	55876	gasdermin B
HHEX	ENSG00000152804	3087	hematopoietically expressed homeobox
HIVEP2	ENSG00000010818	3097	HIVEP zinc finger 2
IGHA1	ENSG00000211895	3493	immunoglobulin heavy constant alpha 1
IGHG1	ENSG00000211896	3500	immunoglobulin heavy constant gamma 1 (G1m marker)
IGHG2	ENSG00000211893	3501	immunoglobulin heavy constant gamma 2 (G2m marker)
IGHG3	ENSG00000211897	3502	immunoglobulin heavy constant gamma 3 (G3m marker)
IKZF1	ENSG00000185811	10320	IKAROS family zinc finger 1
IKZF3	ENSG00000161405	22806	IKAROS family zinc finger 3
IL6ST	ENSG00000134352	3572	interleukin 6 signal transducer
KANSL1	ENSG00000120071	284058	KAT8 regulatory NSL complex subunit 1
KCNK1	ENSG00000135750	3775	potassium two pore domain channel subfamily K member 1
KDELR2	ENSG00000136240	11014	KDEL endoplasmic reticulum protein retention receptor 2
KIF11	ENSG00000138160	3832	kinesin family member 11
KIF3C	ENSG00000084731	3797	kinesin family member 3C
LRRC37A	ENSG00000176681	9884	leucine rich repeat containing 37A
MANBA	ENSG00000109323	4126	mannosidase beta
MAPT	ENSG00000186868	4137	microtubule-associated protein tau
MAU2	ENSG00000129933	23383	MAU2 sister chromatid cohesion factor
MEF2B	ENSG00000213999	1E+08	myocyte enhancer factor 2B
MGAT3	ENSG00000128268	4248	beta-1 4-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase
MTO1	ENSG00000135297	25821	mitochondrial tRNA translation optimization 1
MYCBP	ENSG00000214114	26292	MYC binding protein
NDUFAB1	ENSG00000004779	4706	NADH:ubiquinone oxidoreductase subunit AB1
NFATC2	ENSG00000101096	4773	nuclear factor of activated T cells 2
NFKB1	ENSG00000109320	4790	nuclear factor kappa B subunit 1
NSF	ENSG00000073969	4905	N-ethylmaleimide sensitive factor vesicle fusing ATPase
NXPE1	ENSG00000095110	120400	neurexophilin and PC-esterase domain family member 1
NXPE4	ENSG00000137634	54827	neurexophilin and PC-esterase domain family member 4
ODF1	ENSG00000155087	4956	outer dense fiber of sperm tails 1
ORMDL3	ENSG00000172057	94103	ORMDL sphingolipid biosynthesis regulator 3
OVOL1	ENSG00000172818	5017	ovo like transcriptional repressor 1
RAD51C	ENSG00000108384	5889	RAD51 paralog C

REXO2	ENSG00000076043	25996	RNA exonuclease 2
RNF149	ENSG00000163162	284996	ring finger protein 149
RRAGC	ENSG00000116954	64121	Ras related GTP binding C
RRBP1	ENSG00000125844	6238	ribosome binding protein 1
RRM2B	ENSG00000048392	50484	ribonucleotide reductase regulatory TP53 inducible subunit M2B
RUNX1	ENSG00000159216	861	RUNX family transcription factor 1
RUNX3	ENSG00000020633	864	RUNX family transcription factor 3
SLC17A9	ENSG00000101194	63910	solute carrier family 17 member 9
SLC38A10	ENSG00000157637	124565	solute carrier family 38 member 10
SMARCB1	ENSG00000099956	6598	SWI/SNF related matrix associated actin-dependent regulator of chromatin, subfamily b member 1
SMARCD3	ENSG00000082014	6604	SWI/SNF related matrix associated actin-dependent regulator of chromatin subfamily d member 3
SMOX	ENSG00000088826	54498	spermine oxidase
SPPL2C	ENSG00000185294	162540	signal peptide peptidase like 2C
SPPL3	ENSG00000157837	121665	signal peptide peptidase like 3
ST6GAL1	ENSG00000073849	6480	ST6 beta-galactoside alpha-2 6-sialyltransferase 1
SUPT4H1	ENSG00000213246	6827	SPT4 homolog DSIF elongation factor subunit
TBKBP1	ENSG00000198933	9755	TBK1 binding protein 1
TBX21	ENSG00000073861	30009	T-box transcription factor 21
TCF3	ENSG00000071564	6929	transcription factor 3
TMEM121	ENSG00000184986	80757	transmembrane protein 121
TNFRSF13B	ENSG00000240505	23495	TNF receptor superfamily member 13B
TXLNB	ENSG00000164440	167838	taxilin beta
UBFD1	ENSG00000103353	56061	ubiquitin family domain containing 1
UBR5	ENSG00000104517	51366	ubiquitin protein ligase E3 component n-recognin 5
VPREB3	ENSG00000128218	29802	V-set pre-B cell surrogate light chain 3
WNT3	ENSG00000108379	7473	Wnt family member 3
ZPBP2	ENSG00000186075	124626	zona pellucida binding protein 2

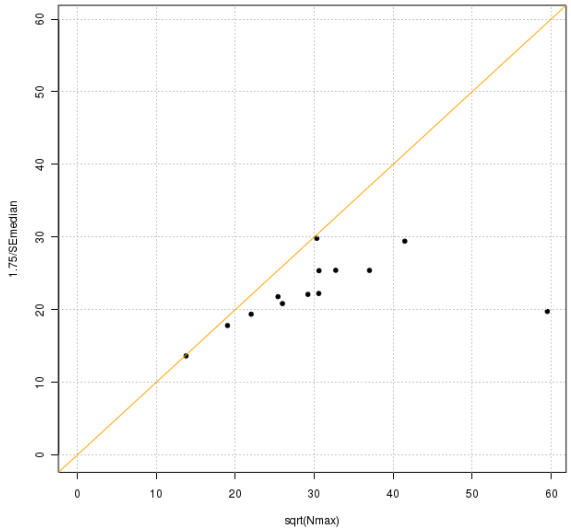
8.3 Supplementary figures



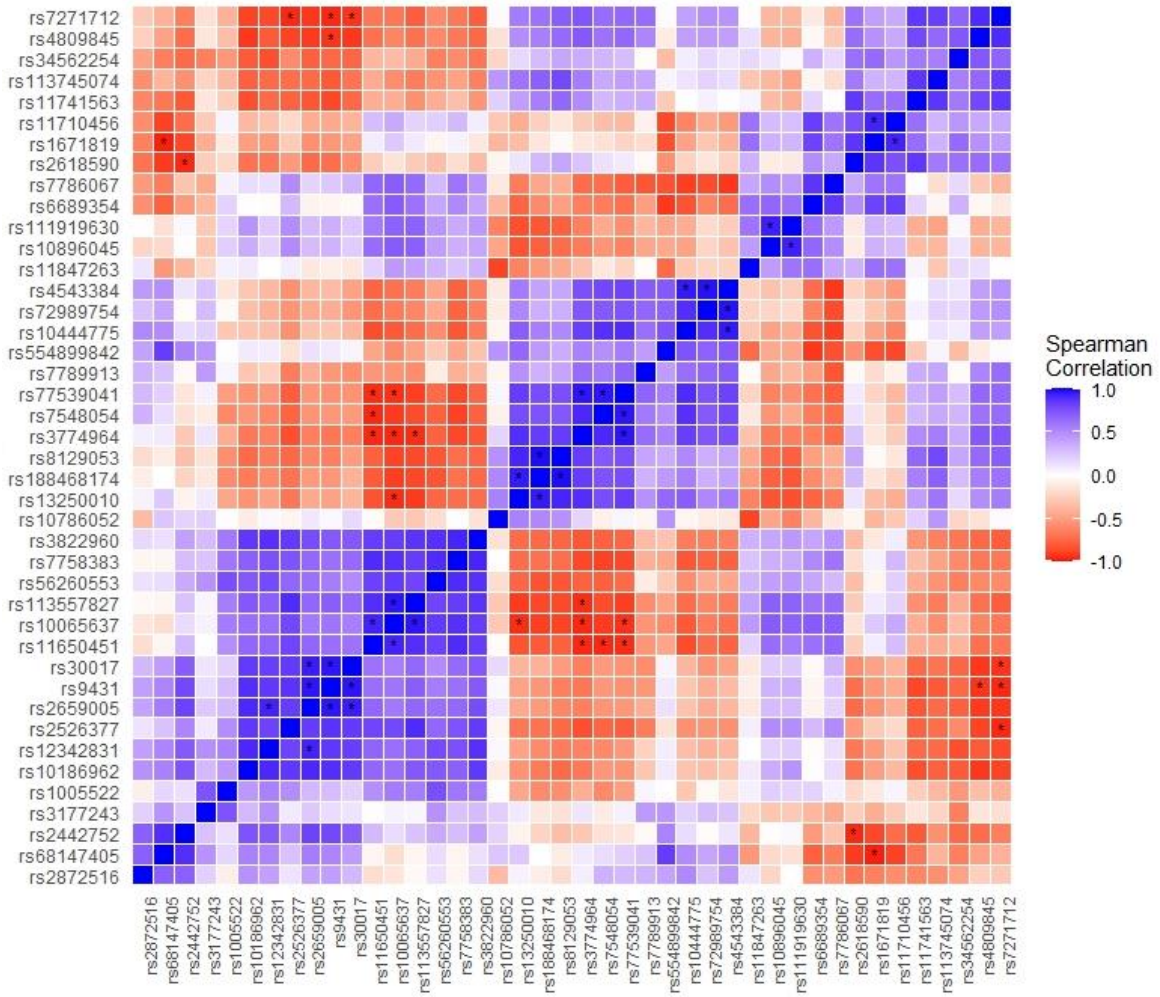
Supplementary Figure 1: Lambda-N plot to reveal issues with population stratification



Supplementary Figure 2: Plots showing allele frequencies in the cohort in comparison to 1000G reference panel in meta-level QC for cohorts that are included in the meta-analysis. Red dots denote outliers that are removed prior to meta-analysis.

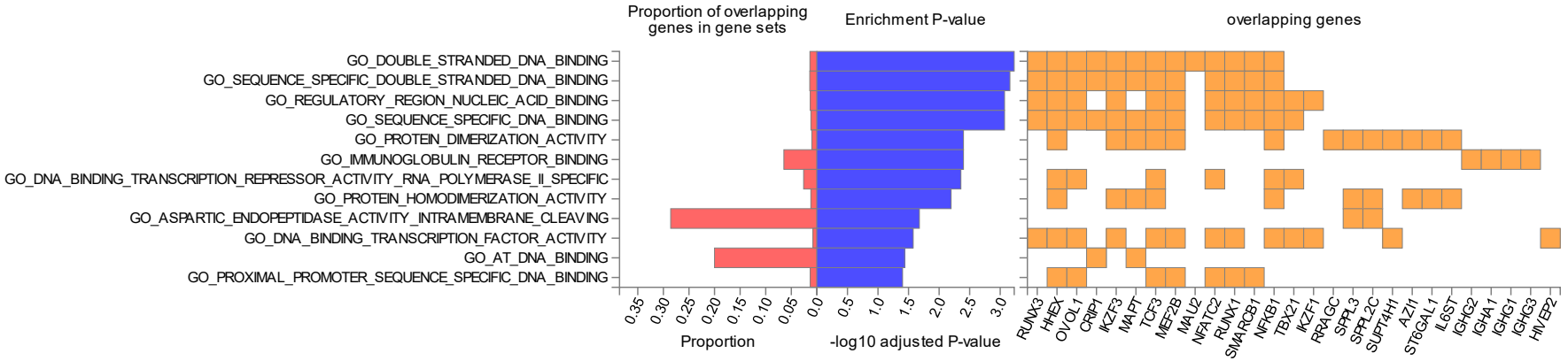


Supplementary Figure 3: SE-N plots to reveal issues with trait transformations. The data for the monosialylation trait is shown

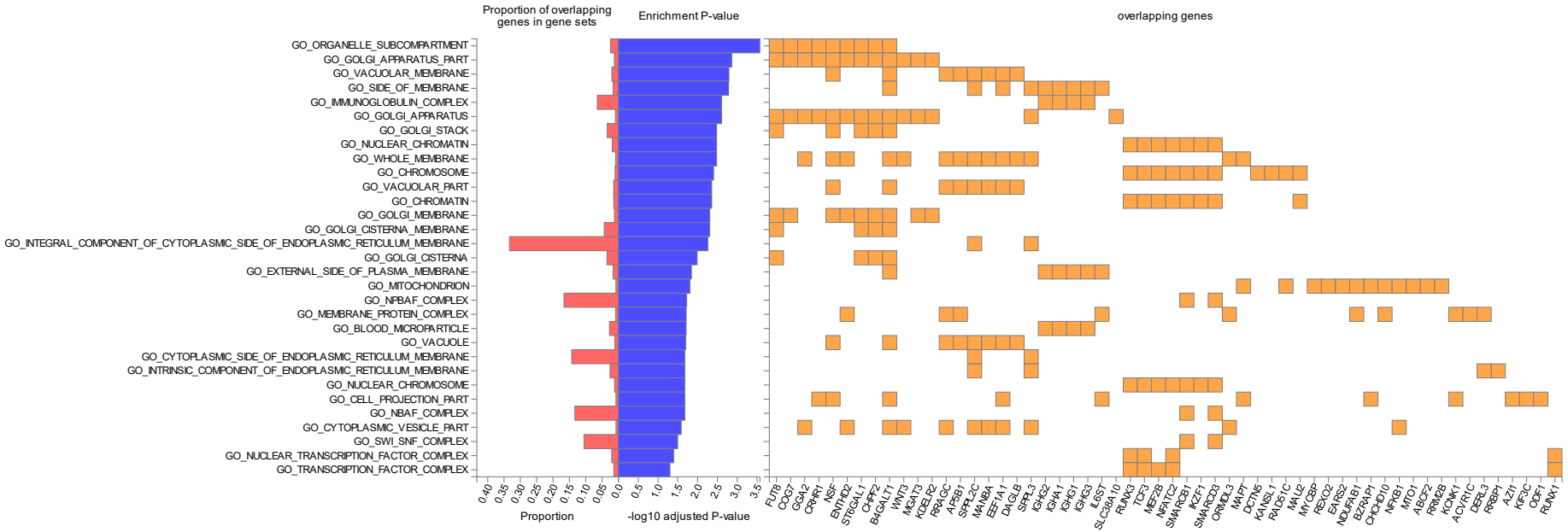


Supplementary Figure 4: Correlation values of the top SNP effects (Z scores)on glycan traits. Significant correlations ($p < 5.8x10^{-5}$)are denoted by asterisks sign *

A)

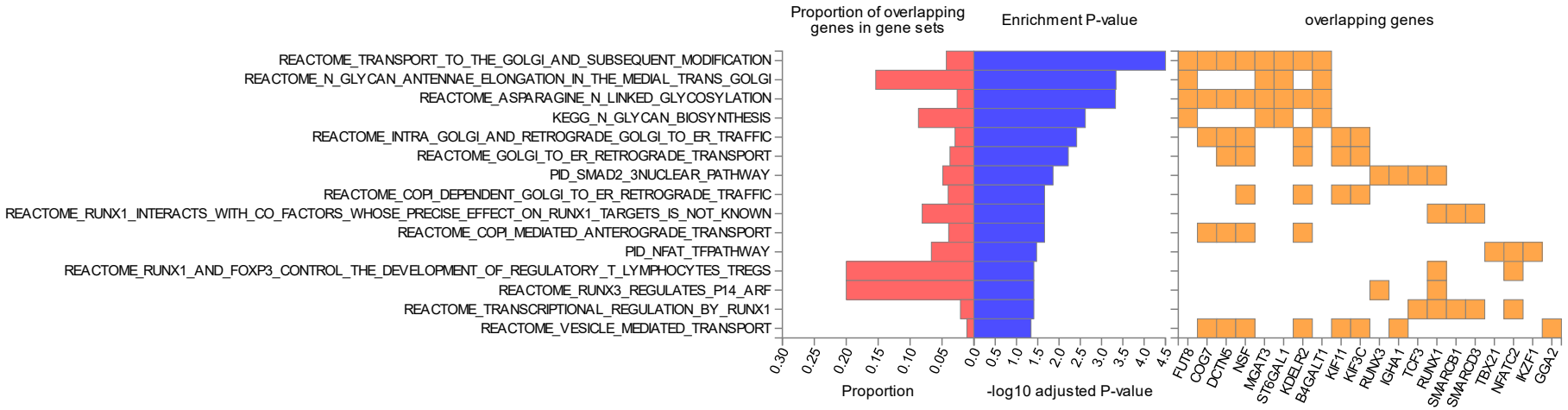

Proportion of overlapping genes in gene sets
Enrichment P-value
overlapping genes
GO_DOUBLE_STRANDED_DNA_BINDING
GO_SEQUENCE_SPECIFIC_DOUBLE_STRANDED_DNA_BINDING
GO_REGULATORY_REGION_NUCLEIC_ACID_BINDING
GO_SEQUENCE_SPECIFIC_DNA_BINDING
GO_PROTEIN_DIMERIZATION_ACTIVITY
GO_IMMUNOGLOBULIN_RECEPTOR_BINDING
GO_DNA_BINDING_TRANSCRIPTION_REPRESSOR_ACTIVITY_RNA_POLYMERASE_II_SPECIFIC
GO_PROTEIN_HOMODIMERIZATION_ACTIVITY
GO_ASPARTIC_ENDOPEPTIDASE_ACTIVITY_INTRAMEMBRANE_CLEAVING
GO_DNA_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY
GO_AT_DNA_BINDING
GO_PROXIMAL_PROMOTER_SEQUENCE_SPECIFIC_DNA_BINDING
0.35 0.30 0.25 0.20 0.15 0.10 0.05 0.0 0.5 1.0 1.5 2.0 2.5 3.0
Proportion
-log10 adjusted P-value
RUNX3 HHEX OVOL1 CRIP1 IKZF3 MAPT TCF3 MEF2B MAU2 NFATC2 RUNX1 SMARCB1 NFKB1 TBX21 IKZF1 RRAGC SPRL3 SPRL2C SUPT4H1 AZI1 ST6GAL1 IL6ST IGHG2 IGHA1 IGHG1 IGHG3 HIVEP2

B)


Proportion of overlapping genes in gene sets
Enrichment P-value
overlapping genes
GO_ORGANELLE_SUBCOMPARTMENT
GO_GOLGI_APPARATUS_PART
GO_VACUOLAR_MEMBRANE
GO_SIDE_OF_MEMBRANE
GO_IMMUNOGLOBULIN_COMPLEX
GO_GOLGI_APPARATUS
GO_GOLGI_STACK
GO_NUCLEAR_CHROMATIN
GO_WHOLE_MEMBRANE
GO_CHROMOSOME
GO_VACUOLAR_PART
GO_CHROMATIN
GO_GOLGI_MEMBRANE
GO_GOLGI_CISTERNA_MEMBRANE
GO_INTEGRAL_COMPONENT_OF_CYTOPLASMIC_SIDE_OF_ENDOPLASMIC_RETICULUM_MEMBRANE
GO_GOLGI_CISTERNA
GO_EXTERNAL_SIDE_OF_PLASMA_MEMBRANE
GO_MITOCHONDRION
GO_NPBAF_COMPLEX
GO_MEMBRANE_PROTEIN_COMPLEX
GO_BLOOD_MICROPARTICLE
GO_VACUOLE
GO_CYTOPLASMIC_SIDE_OF_ENDOPLASMIC_RETICULUM_MEMBRANE
GO_INTRINSIC_COMPONENT_OF_ENDOPLASMIC_RETICULUM_MEMBRANE
GO_NUCLEAR_CHROMOSOME
GO_CELL_PROJECTION_PART
GO_NBAF_COMPLEX
GO_CYTOPLASMIC_VESICLE_PART
GO_SWI_SNF_COMPLEX
GO_NUCLEAR_TRANSCRIPTION_FACTOR_COMPLEX
GO_TRANSCRIPTION_FACTOR_COMPLEX
0.40 0.35 0.30 0.25 0.20 0.15 0.10 0.05 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Proportion
-log10 adjusted P-value
FUT8 COG7 GGA2 CRHR1 NSF ENTPD2 ST6GAL1 CHPF2 B4GALT1 WNT3 MGAT3 KDELR2 RRAGC AP5B1 SPRL2C MAN5A EEF1A1 DAGLB SPRL3 IGHG2 IGHA1 IGHG1 IGHG3 IL6ST SLC39A10 RUNX3 TCF3 MEF2B NFATC2 SMARCB1 IKZF1 SMARCD3 ORMDL3 MAPT DCTN6 KANSL1 RAD51C MAU2 MYO5B REXO2 EARS2 NDUFAB1 BZRAP1 CHCHD10 NFKB1 MTO1 ABCF2 RRM2B KCNK1 ACVR1C DESI1 RRBP1 AZI1 KIF3C ODF1 RUNX1

C)

D)



Supplementary Figure 5: Gene-set enrichment test results. Gene sets with adjusted P-value < 0.05 in each category are shown. A) GO Molecular functions B) GO Cellular components C) GO Biological pathways D) Canonical Pathways

8.4 Other Supplementary Material

Analysis plan for IgG N-glycome GWAS as provided to every cohort analyst

AIM: The discovery GWAS has been performed looking at IgG N-glycosylation traits. The aim is to increase the sample size by combining cohorts analysed both by UPLC and LCMS, thereby increasing the power to detect loci involved in the process of IgG N-glycosylation.

SAMPLES: Men and women ≥ 18 years of age

TRAITS of INTEREST: IgG N-glycans are measured either by UPLC or LCMS, depending on the cohort. Analysed traits include derived IgG N-glycosylation traits representing the overall percentage of presence of certain sugar on the IgG N-glycan. See the end of the document for a complete list of the traits.

GENOTYPES: HRC imputed SNPs

DATA EXCHANGE & TIMELINE FOR ANALYSIS: We aim to be flexible but also need to know when to expect data. Please contact Azra also when you are ready to upload and she will pass on details for data upload.

CONTACT FOR QUESTIONS: Azra Frkatovic (afrkatovic@genos.hr)

GWAS ANALYSIS

Note: For all phenotypes, we use rank-based inverse normal transformation. Standardization and analysis are the same for all traits. **Please use the batch-corrected, normalised and transformed data received from Azra (Genos)**.

MODEL of ASSOCIATION: Additive model. Account for family relatedness and population substructure where needed.

COVARIATES: age (years), sex (0=females, 1=males) and cohort specific covariates (if applicable)

Adjust for covariates: trait ~ age + sex + other covariates (if applicable)

Model: residuals ~ SNP

Transformation: We already performed rank-based inverse normal transformation of the phenotype. No need to additionally transform the data, it is ready to be used as-is.

RESULTS FORMAT

Please provide association results in tab-delimited plain text files, including a single header line with all columns in the order listed below:

rsid – The RSID of the marker analyzed

snpid – SNP identifier in form of chr:pos

chr – The chromosome of the marker analyzed

pos – The position of the marker analyzed (hg19, build 37)

other_allele – a single upper-case character "A" "C" "G" or "T"- Indicating the other (non-effect) allele

effect_allele – a single upper-case character "A" "C" "G" or "T"- The allele associated with phenotypic traits (corresponding to change in betas)

n – The effective number of subjects analyzed

EAF – Effect allele frequency (range 0-1)

beta – Effect size of allele on phenotype for the marker analyzed

se – Standard error of the effect size of the marker analyzed

p – p-value of the effect size of the marker analyzed

strand – Strand on which the alleles are reported

info – A value (range 0-1) corresponding to the information content output from the association testing (according to the data type specified in the "info_type" column above);

info_type - Code indicating the type of data in the “info” column:

1 if the following column contains “r2_Hat” from MACH2DAT/MACH2QTL;

2 if the following column contains “proper_info” from SNPTEST;

3 if the following column contains “INFO” from PLINK.

NOTES:

- Please keep at least 6 digits after the decimal place for all statistics (the use of more precision is encouraged).
- No row indices column or any other extra columns should be provided.
- Columns should be in the order specified above. No specified column should be excluded.
- Please code missing values in any column as "NA"
- Please provide the results of each of the analyses in a separate file, named as described under 'File naming scheme' below. Following the requested format and naming scheme for your results will greatly assist us in collecting and processing the data from many different groups while minimizing errors.
- Please do not do any filtering of SNPs based on MAF, imputation quality, etc as this will be done centrally.

FILE NAMING SCHEME

Please use the following file naming scheme:

STUDY.PHENOTYPE.DATE.txt

STUDY is a short identifier for the population studied. If you have a case-control study and will be providing data separately for cases and controls, please use the suffix ".CASE" or ".CONTROL" after a short identifier for your case-control study name to identify case and control populations.

PHENOTYPE: „fuc_ta","g0_ta", „gal_total_mq", „g1_mq", etc.

DATE is the date on which the file was prepared, in the format "YYYYMMDD"

example: CROATIA.fuc.20190131.txt

List of traits:

fuc_mq	g2_mq
g0_mq	gal_total_mq
g1_mq	s1_mq

bisecting_mq

s1_no_bis_mq

s1_g1_mq

s1_g2_mq

s1_gal_total_m

9. BIOGRAPHY

Azra Frkatović was born on April 19, 1994 in Mannheim, Germany. She attended primary school in her hometown Zavidovići in Bosnia and Herzegovina before earning a scholarship to attend high school in Sarajevo. Azra completed her undergraduate and master studies in genetics at the Department of Genetics and Bioengineering at the Faculty of Engineering and Natural Sciences, Burch University, Sarajevo in 2017 with a master thesis on gene expression in human periapical inflammatory lesions.

Since February 2018, she has been employed as a doctoral researcher in the Glycoscience Research laboratory Genos where she was part of the EU-funded MSCA IMforFUTURE network as an early-stage researcher. She worked on elucidating the genetic and epigenetic background of immunoglobulin G N-glycosylation. During her PhD studies, she had an opportunity to be a visiting researcher at the Institute of Genetics and Cancer at the University of Edinburgh and the Department of Twin Research and Genetic Epidemiology at the King's College London, UK.

She has participated in several scientific conferences with poster and oral presentations, as well as the organization of IMforFUTURE workshop and Marie-Curie Alumni Association annual conferences. She is the co-author of five scientific papers published in international peer-reviewed journals.