Human Molecular Genetics, 2019, Vol. 28, No. 12 2062–2077

doi: 10.1093/hmg/ddz054
Advance Access Publication Date: 11 March 2019
Association Studies Article

ASSOCIATION STUDIES ARTICLE

Defining the genetic control of human blood plasma N-glycome using genome-wide association study

Sodbo Zh. Sharapov[1,2,†], Yakov A. Tsepilov[1,2,‡], Lucija Klaric[3,4], Massimo Mangino[5,6], Gaurav Thareja[7], Alexandra S. Shadrina[2,¶], Mirna Simurina[8], Concetta Dagostino[9], Julia Dmitrieva[10], Marija Vilaj[4], Frano Vuckovic[4], Tamara Pavic[8], Jerko Stambuk[4], Irena Trbojevic-Akmacic[4], Jasminka Kristic[4], Jelena Simunovic[4], Ana Momcilovic[4], Harry Campbell[11,12], Margaret Doherty[13,14], Malcolm G. Dunlop[12], Susan M. Farrington[12], Maja Pucic-Bakovic[4], Christian Gieger[15], Massimo Allegri[16], Edouard Louis[17], Michel Georges[10], Karsten Suhre[7], Tim Spector[5], Frances M.K. Williams[5], Gordan Lauc[4,8,§] and Yurii S. Aulchenko[1,2,18,§,*]

[1]Institute of Cytology and Genetics SB RAS, Prospekt Lavrentyeva 10, Novosibirsk, 630090, Russia, [2]Novosibirsk State University, 1, Pirogova str., Novosibirsk, 630090, Russia, [3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Crewe Road South, Edinburgh EH4 2XU, UK, [4]Genos Glycoscience Research Laboratory, Borongajska cesta 83h, 10000 Zagreb, Croatia, [5]Department of Twin Research and Genetic Epidemiology, School of Life Course Sciences, King's College London, St Thomas' Campus, Lambeth Palace Road, London, SE1 7EH, UK, [6]NIHR Biomedical Research Centre at Guy's and St Thomas' Foundation Trust, London SE1 9RT, UK, [7]Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, P.O. Box 24144 Doha, Qatar, [8]Faculty of Pharmacy and Biochemistry, University of Zagreb, Ante Kovacica 1, 10 000 Zagreb, Croatia, [9]Department of Medicine and Surgery, University of Parma, Via Gramsci 14, 43126 Parma, Italy, [10]Unit of Animal Genomics, WELBIO, GIGA-R and Faculty of Veterinary Medicine, University of Liège, (B34) 1 Avenue de l'Hôpital, Liège 4000, Belgium, [11]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh EH8 9AG, UK, [12]Colon Cancer Genetics Group, MRC Human Genetics Unit, MRC Institute of Genetics & Molecular Medicine, Western General Hospital, The University of Edinburgh, Edinburgh EH4 2XU, UK [13]Institute of Technology Sligo, Department of Life Sciences, Sligo, Ireland,

[†]Sodbo Zh. Sharapov, http://orcid.org/0000-0003-0279-4900
[‡]Yakov A. Tsepilov, http://orcid.org/0000-0002-4931-6052
[¶]Alexandra S. Shadrina, http://orcid.org/0000-0003-1384-3413
[§]These authors jointly supervised this work and contributed equally.
Received: August 3, 2018. **Revised:** March 1, 2019. **Accepted:** March 6, 2019

[14]National Institute for Bioprocessing Research & Training, Dublin, Ireland, [15]Institute of Epidemiology II, Research Unit of Molecular Epidemiology, Helmholtz Centre Munich, German Research Center for Environmental Health, Ingolstädter Landstr. 1, D-85764, Neuherberg, Germany, [16]Pain Therapy Department, Policlinico Monza Hospital, 20090 Monza, Italy, [17]CHU-Liège and Unit of Gastroenterology, GIGA-R and Faculty of Medicine, University of Liège, 1 Avenue de l'Hôpital, Liège 4000, Belgium and [18]PolyOmica, Het Vlaggeschip 61, 5237 PA 's-Hertogenbosch, The Netherlands

*To whom correspondence should be addressed. Tel: +7-383-3634963; Email: yurii@bionet.nsc.ru

Abstract

Glycosylation is a common post-translational modification of proteins. Glycosylation is associated with a number of human diseases. Defining genetic factors altering glycosylation may provide a basis for novel approaches to diagnostic and pharmaceutical applications. Here we report a genome-wide association study of the human blood plasma N-glycome composition in up to 3811 people measured by Ultra Performance Liquid Chromatography (UPLC) technology. Starting with the 36 original traits measured by UPLC, we computed an additional 77 derived traits leading to a total of 113 glycan traits. We studied associations between these traits and genetic polymorphisms located on human autosomes. We discovered and replicated 12 loci. This allowed us to demonstrate an overlap in genetic control between total plasma protein and IgG glycosylation. The majority of revealed loci contained genes that encode enzymes directly involved in glycosylation (*FUT3/FUT6*, *FUT8*, *B3GAT1*, *ST6GAL1*, *B4GALT1*, *ST3GAL4*, *MGAT3* and *MGAT5*) and a known regulator of plasma protein fucosylation (*HNF1A*). However, we also found loci that could possibly reflect other more complex aspects of glycosylation process. Functional genomic annotation suggested the role of several genes including *DERL3*, *CHCHD10*, *TMEM121*, *IGH* and *IKZF1*. The hypotheses we generated may serve as a starting point for further functional studies in this research area.

Introduction

Glycosylation—addition of carbohydrates to a substrate—is a common and structurally diverse cotranslational and posttranslational modification of proteins that can affect their physical properties (solubility, conformation, folding, stability, trafficking etc.) (1–4) as well as biological functions, including protein–protein, cell–cell, cell–matrix and host–pathogen interactions (2,3,5,6). Carbohydrate units can be attached to the protein by N- or O-glycosidic bonds with the rare exceptions of C-glycosidic attachment (7). N-glycosylation is the most abundant type of glycosylation (8,9) and, unlike other types, is specific to a consensus asparagine-containing sequence in the primary structure of the protein. Glycoproteins comprise various enzymes, hormones, cytokines, receptors, immunoglobulins, structural, adhesion and other protein molecules, and altered glycosylation is increasingly recognized to be implicated in human pathologies. In particular, association with changes in total plasma protein N-glycome composition or immunoglobulin G (IgG) glycosylation has been found for Parkinson's disease (10), low back pain (11), rheumatoid arthritis (12), ulcerative colitis, Crohn's disease (13) and type 2 diabetes (14). Beyond that, aberrant glycosylation is involved in key pathological steps of tumor development and is even considered a new hallmark of cancer (15–17). Glycans are considered as potential therapeutic targets (18) and biomarkers for early diagnosis and disease prognosis (19–22), which makes glycobiology a promising field for future clinical applications. An example of glycoprotein biomarker is AFP-L3—the fucosylated fraction of alpha-fetoprotein—that was approved by the U.S. Food and Drug Administration as a diagnostic marker of primary hepatocellular carcinoma (21).

Protein glycosylation is an extremely complex process depending on the interplay of multiple enzymes catalyzing glycan transfer, glycosidic linkage hydrolysis as well as glycan biosynthesis. Abundance of specific protein glycoforms can be influenced by a variety of parameters including activity of enzymes and availability of substrates, accessibility of a glycosylation site, protein synthesis and degradation. No surprise that, overall, protein glycosylation is a complex process that is controlled by genetic, epigenetic and environmental factors (23–25). Mechanisms of regulation of this process are only started to be understood. Genome-wide association studies (GWASs) can expand our knowledge on this topic by a hypothesis-free search of candidate genes involved in regulation of glycosylation. Their role can be clarified in subsequent functional follow-up studies.

Previous GWAS of total plasma protein N-glycome measured with high-performance liquid chromatography (HPLC) discovered six loci associated with protein glycosylation (26,27). Four of these loci contained genes that have well-characterized roles in glycosylation—the fucosyltransferases *FUT6* and *FUT8*, glucuronyltransferase *B3GAT1* and glucosaminyltransferase *MGAT5*. Other two loci—near the *SLC9A9* gene on chromosomes 3 and near the *HNF1A* gene on chromosome 12—did not contain any genes known to be involved in glycosylation processes. A functional *in vitro* follow-up study in HepG2 cells showed that the *HNF1A* gene product acts as a co-regulator of expression of most fucosyltransferase genes (*FUT3*, *FUT5*, *FUT6*, *FUT8*, *FUT10* and *FUT11*) (26). In addition, it co-regulates the expression of genes encoding key enzymes required for the synthesis of GDP-fucose, the substrate of these fucosyltransferases. It is noteworthy that identification of *HNF1A* as one of the master regulators of protein fucosylation enabled to propose a new diagnostic tool for discrimination between *HNF1A*-MODY monogenic diabetes and type 1 and type 2 diabetes based on the ratio of fucosylated to nonfucosylated triantennary glycans (19). The locus on chromosome 3 contains *SLC9A9* gene, which encodes a proton

Table 1. Replication of six previously reported loci [Huffman *et al.*(27)]

				Results of Huffman *et al.* (27) ($N = 3,533$)					This study, TwinsUK ($N = 2,763$)			
SNP	CHR:POS	Gene	Eff/Ref	EAF	N	Trait	BETA (SE)	P-value	EAF	Trait	BETA(SE)	P-value
rs1257220	2:135015347	*MGAT5*	A/G	0.26	3263	Tetra-antennary glycans	0.19 (0.03)	1.80E-10	0.25	FA3G3S[3,3,3]3	0.16 (0.032)	**3.98E-07**
rs4839604	3:142960273	*SLC9A9*	C/T	0.77	3320	Tetrasialylated	−0.22 (0.03)	3.50E-13	0.83	FBS2/(FS2 + FBS2)	−0.11 (0.038)	2.50E-03
rs7928758	11:134265967	*B3GAT1*	T/G	0.88	3233	A4F2G4 (DG13)	0.23 (0.04)	1.66E-08	0.84	A3G3S[3,6]2	0.24 (0.038)	**6.07E-10**
rs735396	12:121438844	*HNF1A*	T/C	0.61	3236	A2F1G2 (DG7)	0.18 (0.03)	7.81E-12	0.65	G4S3/G4S4	0.18 (0.031)	**6.06E-09**
rs11621121	14:65822493	*FUT8*	C/T	0.43	3234	A2 (DG1)	0.27 (0.03)	1.69E-23	0.40	A2[6]BG1n	0.21 (0.029)	**1.60E-12**
rs3760776	19:5839746	*FUT6*	G/A	0.87	3262	A3F1G3 (DG9)	0.44 (0.04)	3.18E-29	0.91	A4G4S[3,3,3]3	0.56 (0.050)	**3.71E-28**

Replicated loci are in bold. CHR:POS—chromosome and position of SNP according to GRCh37 human genome build; Eff/Ref—effective and reference alleles; gene—candidate gene for the locus reported in Huffman *et al.* (27); EAF—effective allele frequency; N—sample size; Trait—glycan trait with statistically strongest association with the SNP; DG (desialylated peak)—HPLC peak after sialidase treatment; BETA (SE)—effect (in SD units) and standard error of effect; P-value—P-value of association.

pump affecting pH in the endosomal compartment, reminiscent of recent findings that changes in Golgi pH can impair protein sialylation (27).

Since 2011, when the latest GWAS of plasma N-glycome was published, new technologies for glycome profiling have been developed (28). Ultra-performance liquid chromatography (UPLC) became a widely used technology for accurate analysis of plasma N-glycosylation due to its superior sensitivity, resolution, speed and capability to provide branch-specific information of glycan structures (29). Moreover, new imputation panels [such as 1000 Genomes (30) and HRC (31)] became available, increasing the resolution and power of genetic mapping.

In this work, we aimed to advance our understanding of the genetic control of the human plasma N-glycome and to establish a public resource that will facilitate future studies linking glycosylation and complex human diseases. For that, we performed and reported results of GWAS on 113 plasma glycome traits measured by UPLC and genotypes imputed to the 1000 Genomes reference panel in 2763 participants of TwinsUK. Further, we replicated our findings in 1048 samples from three independent and genetically diverse cohorts—PainOR, SOCCS and Qatar Metabolomics Study on Diabetes (QMDiab).

Results

Replication of previously reported loci

We started with replication of six loci that were reported previously. Huffman and colleagues (27) analyzed four independent cohorts with total sample size of 3533, using plasma N-glycome measured with HPLC. Because of technological differences, there is no one-to-one correspondence between HPLC and UPLC traits and exact replication is not possible. Therefore, we analyzed association of single-nucleotide polymorphisms (SNPs) reported by (27) with all 113 UPLC traits measured in this study and considered a locus replicated if we observed $P < 0.05/(6 \times 30) = 2.78 \times 10^{-4}$ (where 30 is a number of principal components, explaining 99% of the variation of the 113 studied traits) in the TwinsUK cohort ($N = 2763$). Using this procedure, we replicated five of the six previously reported SNPs (Table 1, Fig. 1). For more details, see Supplementary Material, Table 1.

These results not only confirm previous and establish five plasma glycome loci as replicated but also demonstrate that our study is well powered (among replicated loci, all P-values were less than 4×10^{-7}).

Discovery and replication of new loci

In the next step, we searched for new loci via genome-wide association scan with subsequent replication of the revealed association signals.

The discovery cohort comprised 2763 participants of the TwinsUK study with genotypes available for 8 557 543 SNPs. The genomic control inflation factor varied from 0.99 to 1.02, suggesting that influences of residual population stratification on the test statistics were small (see Supplementary Material, Table 2; QQ-plots in Supplementary Fig. 1). In total, 906 SNPs located in 14 loci were significantly associated ($P < 5 \times 10^{-8}/30 = 1.66 \times 10^{-9}$, where 30 is a number of principal components, explaining 99% of the variation of the 113 studied traits) with at least one of 113 glycan traits (in total 5052 SNP–trait associations, see Fig. 2, Table 1). Out of 113 traits, 68 were significantly associated with at least one of the 14 loci. For more details, see Supplementary Material, Table 3.

Among 14 loci, four were previously reported as associated with the plasma N-glycome (Fig. 1). Three loci—on chromosome 12 at 121 Mb (leading SNP: rs1169303, intronic variant of the *HNF1A* gene), on chromosome 14 at 105 Mb (leading SNP: rs7147636 located in the intron of the *FUT8* gene) and on chromosome 19 at 58 Mb (leading SNP: rs7255720, upstream variant of the *FUT6* gene)—were reported to be associated with the plasma N-glycome in two previous GWAS (26,27), while association of the locus on chromosome 11 at 126 Mb (leading SNP: rs1866767 located in the intron of *B3GAT1* gene) was reported only in the latest GWAS meta-analysis of plasma N-glycome (27).

Ten further loci that have not been reported before were found here. In order to replicate our findings, we have performed association analysis of these 10 SNPs in three independent cohorts—PainOR, SOCCS and QMDiab (total $N = 1048$)—and then meta-analyzed the results. Seven of ten novel loci were replicated at threshold $P < 0.05/10 = 0.005$ (see Table 2). The direction of association was concordant between discovery and replication for all 10 loci. The effects of loci between the replication cohorts were homogeneous (P-value of Cochran's Q-test varied from 0.07 to 0.96, see Supplementary Material, Table 3).

Given seven replicated novel loci found in this study and five loci found previously and replicated in this study, we now have 12 replicated loci in total (Fig. 1).

Functional annotation *in silico*

For 12 replicated loci, we performed an *in silico* functional annotation in order to prioritize potentially causal genes. Prioritization used multiple lines of evidence, such as presence of predicted damaging variants in a gene, pleiotropic effects of glycan-associated SNPs on gene expression and the results of a DEPICT (Data-driven Expression Prioritized Integration for Complex Traits) analysis, which employs predicted gene functions and reconstituted gene sets (32). Although most of the observed loci contain genes encoding proteins with a known role in glycosylation (*FUT3/FUT6*, *FUT8*, *B3GAT1*, *ST6GAL1*, *B4GALT1*, *ST3GAL4*, *MGAT3*, *MGAT5* and *HNF1A*), we still included them in the

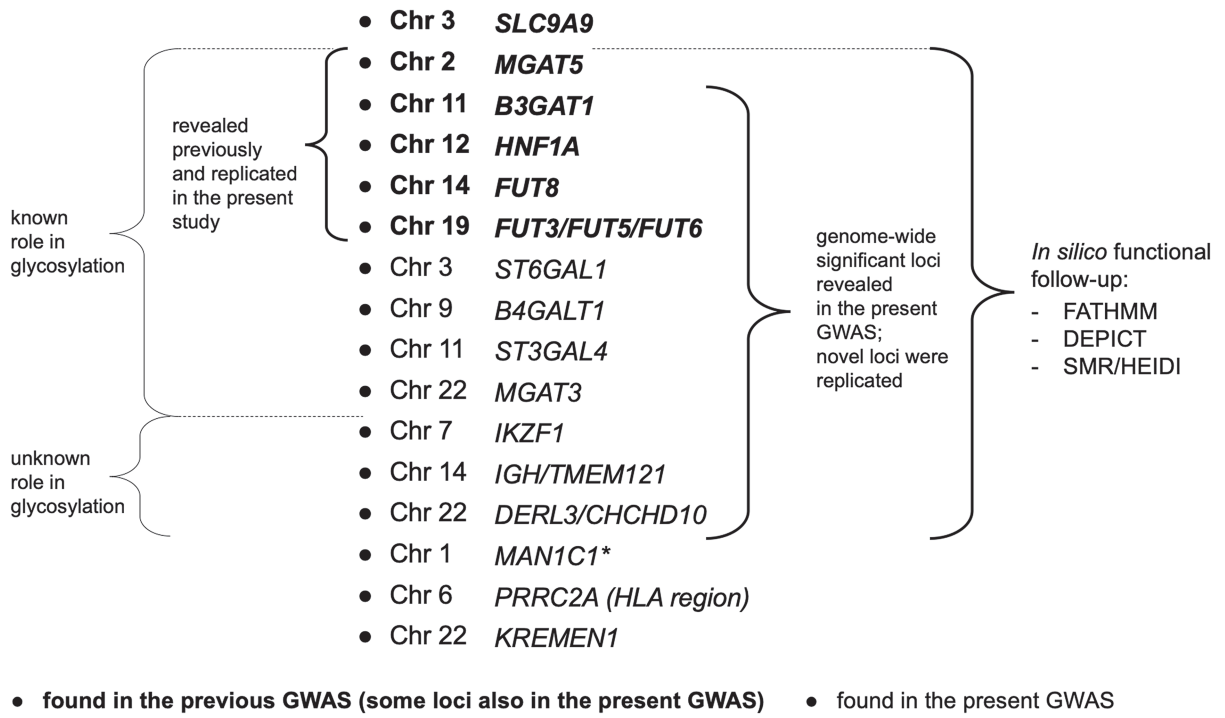

Figure 1. Schematic overview of the novel and known loci associated with plasma protein N-glycosylation. Chr—chromosome. Genes, prioritized in this study, represent corresponding loci. *The role of MAN1C1 (mannosidase alpha class 1C member 1) gene in glycosylation is known, but the association of MAN1C1 locus was not replicated in the present study.

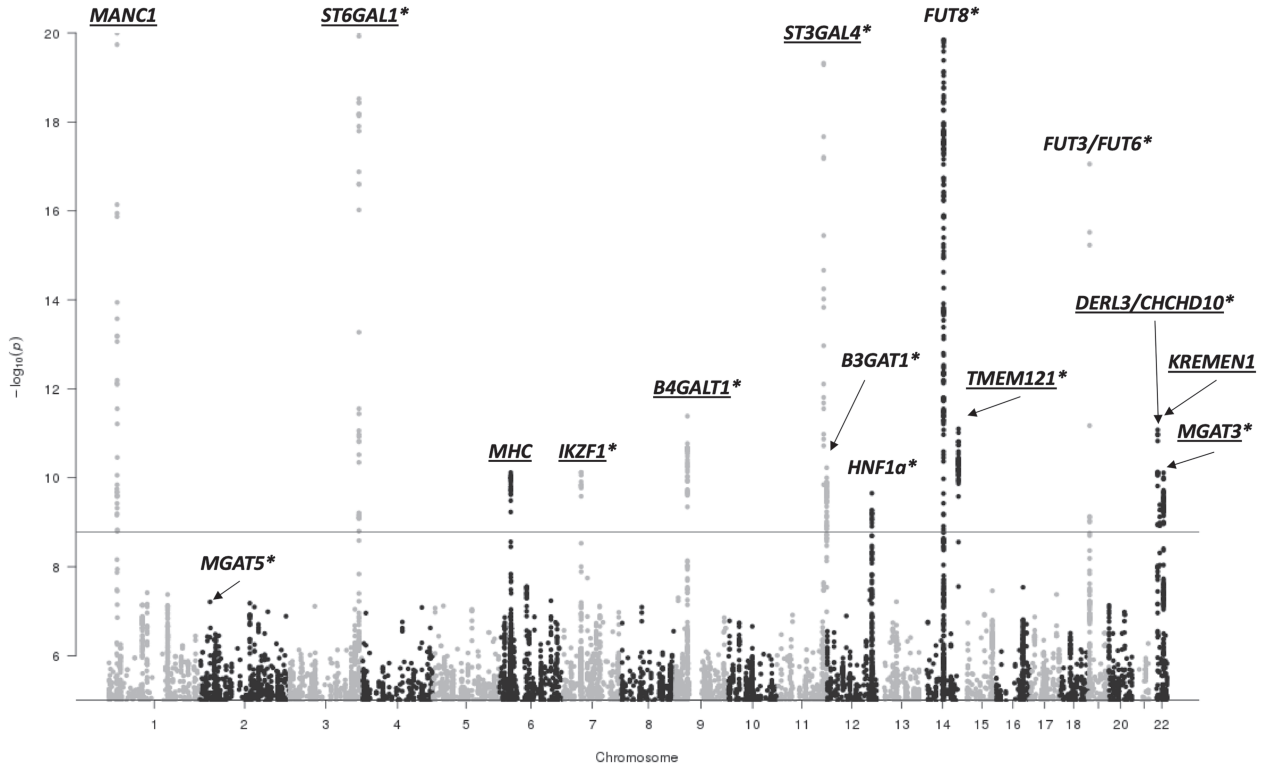

Figure 2. Manhattan plot of discovery GWAS (after correction for genomic control). Horizontal line corresponds to the genome-wide significance threshold of 1.7×10^{-9}. For each SNP the lowest *P*-value among 113 traits is shown. Only SNPs with $P < 1 \times 10^{-5}$ are shown. Points with −log10(*P*) > 20 are depicted at −log10 (*P*-value) = 20. Gene labels mark loci that were found in previous GWAS (27); underlined gene labels mark novel loci; *loci replicated either in the TwinsUK cohort (for previously reported loci) or in the replication meta-analysis of three cohorts PainOR, QMDiab and SOCCS (novel loci identified in the present study).

Table 2. Fourteen loci genome-wide significantly associated with at least one of the 113 traits in this study

				Discovery					Replication			
SNP	CHR:POS	Gene	Eff/Ref	EAF	BETA (SE)	P-value	Top trait	N traits	EAF	BETA (SE)	P	N
Novel loci												
rs186127900	1:25318225	*AL445471.2 (MAN1C1)*	G/T	0.99	−1.26 (0.119)	4.04E-24	FBG1n/G1n	26	0.99	−0.35 (0.224)	1.22E-01	1093
rs59111563	3:186722848	*ST6GAL1*	D/I	0.74	0.34 (0.031)	1.09E-26	FG1S1/(FG1+FG1S1)	3	0.73	0.32 (0.048)	9.50E-12	1088
rs3115663	6:31601843	*PRRC2A*	T/C	0.80	0.26 (0.040)	7.65E-11	M9	1	0.83	0.06 (0.059)	3.01E-01	1093
rs6421315	7:50355207	*IKZF1*	G/C	0.59	0.19 (0.029)	7.57E-11	A2[6]BG1n	2	0.60	0.27 (0.043)	5.67E-10	1077
rs13297246	9:33128617	*B4GALT1*	G/A	0.83	−0.26 (0.038)	4.11E-12	FA2G2n	2	0.83	−0.26 (0.059)	8.66E-06	1093
rs3967200	11:126232385	*ST3GAL4*	C/T	0.88	−0.49 (0.043)	1.51E-27	A2G2S[3,6+3]2	7	0.86	−0.53 (0.062)	6.85E-18	1093
rs35590487	14:105989599	*IGH / TMEM121*	C/T	0.77	−0.24 (0.034)	7.98E-12	FA2[3]G1n	2	0.78	−0.17 (0.058)	3.67E-03	1093
rs9624334	22:24166256	*DERL3 / CHCHD10*	G/C	0.85	0.28 (0.040)	8.38E-12	FA2[6]BG1n	2	0.86	0.42 (0.062)	2.09E-11	1086
rs140053014	22:29550678	*KREMEN1*	I/D	0.98	−0.67 (0.106)	4.05E-10	G3S2/G3S3	1	0.98	−0.24 (0.165)	1.50E-01	1079
rs909674	22:39859169	*MGAT3*	C/A	0.27	0.22 (0.033)	7.72E-11	FBS2/FS2	3	0.25	0.18 (0.053)	5.70E-04	1045
Previously implicated loci												
rs1866767	11:134274763	*B3GAT1*	C/T	0.87	0.28 (0.043)	5.95E-11	A4G4S[3,3,3,3]4	3		—		
rs1169303	12:121436376	*HNF1A*	A/C	0.51	0.19 (0.029)	2.23E-10	A4G4S[3,3,3]3	2				
rs7147636	14:66011184	*FUT8*	T/C	0.33	−0.39 (0.030)	6.63E-37	FA2G2S[3+6,6+3]2	17				
rs7255720	19:5828064	*FUT6*	G/C	0.96	1.14 (0.068)	2.53E-55	G4S3/G4S4	18				

Ten loci in the upper part of the table are novel for N-glycome traits, and four loci in the lower part of the table have been found previously. Replicated novel loci are in bold. CHR:POS—chromosome and position of SNP according to GRCh37 human genome build; Gene—suggested candidate genes for replicated loci (see Table 3) or the nearest gene for non-replicated loci. For the locus on chromosome 1 we additionally report *MAN1C1* gene as its product has known role in glycosylation processes; Eff/Ref—effective and reference allele; EAF—effective allele frequency; BETA (SE)—effect (in SD units) and standard error of effect; P-value—P-value after GC correction; Top trait—glycan trait with the strongest association (the lowest P-value); N traits—total number of traits significantly associated with given locus; N—sample size of replication. Description of glycan traits is provided in Supplementary Material, Table 10.

analysis. Beside this, we explored gene set/tissue/cell type enrichment and investigated potential pleiotropic effects of these loci on other complex human traits.

Analysis of possible effects of genetic variants

Within each of replicated loci, we identified a set of SNPs that are in high LD ($r^2 > 0.6$) with 12 lead variants. In total, we identified 619 SNPs that were further subjected to variant effect prediction analysis using VEP (33), FATHMM-XF (34), and FATHMM-INDEL (35). Full results of variant effect prediction annotation of 619 SNPs are presented in Supplementary Material, Table 4A and B. We detected four potentially pathogenic variants in four genes—coding variant rs17855739 (p.E247K, substitution of negatively charged Glu with positively charged Lys) in the *FUT6* gene, insertion/deletion variant rs149306472 (p.G204Lfs*35, deletion of Gly) in the *SYNGR1* gene, non-coding variant rs7423 in the *SYNGR1* gene and coding variant rs3177243 (p.F149 L, substitution of Phe with Leu) in the *DERL3* gene. However, it should be mentioned that recent study (36) highlighted the danger, at least for common variants, of pinpointing coding variants as likely to be causal.

Gene-set and tissue/cell enrichment analysis

For prioritizing genes in associated regions (based on their predicted function) and gene set and tissue/cell type enrichment analyses we used DEPICT software (32). When running DEPICT analyses on the 14 genome-wide significant loci (from Table 1), we identified tissue/cell type enrichment (with false discovery rate (FDR) < 0.05) for six tissue/cell types: plasma cells, plasma, parotid gland, salivary glands, antibody producing cells and B-lymphocytes (see Supplementary Material, Table 5C). We did not identify any significant enrichment for gene-sets (all FDR > 0.2, Supplementary Material, Table 5B). DEPICT suggestively prioritized three genes—*FUT3*, *DERL3* and *FUT8*—for three loci (on chromosome 19 at 58 Mb, on chromosome 22 at 24 Mb and on chromosome 14 at 65/66 Mb) with FDR < 0.20 (see Supplementary Material, Table 5A). We have also analyzed 93 loci with $P < 1 \times 10^{-5}/30$ (Supplementary Material, Table 6); however, all results had FDR > 0.2.

Next, we performed gene ontology (GO) gene-set enrichment analysis using Multi-marker Analysis of GenoMic Annotation (MAGMA) approach (37) for the loci with $P < 1 \times 10^{-5}$ (Supplementary Material, Table 7A). We observed strong enrichment of genes involved in the glycan synthesis pathways—such as 'oligosaccharide metabolic process' (P-value of enrichment after correction for multiple testing, 1.56×10^{-29}), 'protein N-linked glycosylation' (P-value, 4.45×10^{-22}) and 'N-glycan biosynthesis' (P-value, 1.69×10^{-20}). After exclusion of 14 genome-wide significant loci, the significance of enrichment of glycome-related pathways has reduced (corresponding P-values without correction for multiple testing were 0.57, 0.097 and 0.0063; see Supplementary Material, Table 7B).

Pleiotropy with expression quantitative trait loci

We next attempted to identify genes whose expression level could potentially mediate the association between SNPs and plasma N-glycome. To do this, we performed a summary data-based Mendelian randomization (SMR) analysis followed by heterogeneity in dependent instruments (HEIDI) test (38) using a collection of eQTL (expression quantitative trait loci) data for blood (39), 44 tissues provided in the GTEx database version 6p (40) and six blood cell types collected in the CEDAR study [see Supplementary Material, Note 3 and (41)]—five immune cell populations (CD4+, CD8+, CD19+, CD14+ and CD15+) and platelets. In short, SMR tests the association between gene expression in a particular tissue/cell type and a trait using the top associated SNP as a genetic instrument. Significant SMR test may indicate that the same functional variant influences both expression and the trait of interest (causality or pleiotropy) but may also indicate that functional variants underlying gene expression are in linkage disequilibrium with those controlling the traits. Inferences whether functional variant may be shared between plasma glycan trait and expression were made based on HEIDI test: $P_{HEIDI} > 0.05$ (likely shared), $0.05 > P_{HEIDI} > 0.001$ (possibly shared) and $P_{HEIDI} < 0.001$ (sharing is unlikely).

We applied SMR/HEIDI analyses for replicated loci that demonstrated genome-significant association in our discovery data (11 loci). In total, we included in the analysis

expression levels of 20 448 transcripts (probes). For 15 probes located in 7 loci associated with plasma glycosylation, we observed significant ($P_{SMR} < 0.05/20\,448 = 2.445 \times 10^{-6}$) association with the top SNPs associated with plasma N-glycome (see Supplementary Material, Table 8). Subsequent HEIDI test showed that the hypothesis of shared functional variant between plasma glycan traits and expression was most likely ($P_{HEIDI} > 0.05$) for four probes: *ST6GAL1* in whole blood [from Westra *et al.* (39)], *TMEM121* in whole blood [GTEx (40)], *MGAT3* in CD19+ cells [CEDAR (41)] and *CHCHD10* in whole blood [from Westra *et al.* (39)]. For other five probes, we conclude that the functional variant is possibly shared ($0.001 < P_{HEIDI} < 0.05$) between glycan traits and expression of *ST3GAL4* [in two different tissues: muscle skeletal and pancreas; GTEx (40)], *B3GAT1* [in two tissues: whole blood from Westra *et al.* (39) and lung tissue from GTEx (40)] and *SYNGR1* [in tibial nerve tissue from GTEx (40)].

Overlap with complex traits

We next investigated the potential pleiotropic effects of our loci on other complex human traits and diseases using PhenoScanner v1.1 database (42). For 12 replicated SNPs (Tables 1 and 2), we looked up traits that were genome-wide significantly ($P < 5 \times 10^{-8}$) associated with the same SNP or SNP in a strong ($r^2 > 0.7$) linkage disequilibrium. The results are summarized in Supplementary Material, Table 9. For 8 out of 12 loci, we observed associations with a number of complex traits. Four loci (near *IKZF1*, *FUT8*, *MGAT3* and *DERL3*) were associated with levels of glycosylation of IgG (43). Two loci (on chromosome 12 at 121 Mb and on chromosome 11 at 126 Mb, containing *HNF1a* and *ST3GAL4* genes, respectively) were associated with LDL and total cholesterol levels (44,45). The locus-containing *HNF1a* was additionally associated with the level of plasma C reactive protein (46,47) and gamma glutamyl transferase (48). Locus on chromosome 22 at 39 Mb (containing *MGAT3*) was associated with adult height (49). Locus on chromosome 14 at 65/66 Mb (near *FUT8*) was associated with age at menarche (50). Note, however, that PhenoScanner analysis does not allow distinguishing between pleiotropy of a variant shared between the traits and linkage disequilibrium between different functional variants affecting separate traits.

Summary of *in-silico* follow-up

We compared the genes suggested by our *in silico* functional investigation with the candidate genes suggested previously for five known loci (see Table 3). For three out of five loci (*B3GAT1, FUT8* and *FUT6/FUT3*), we selected the same genes as suggested by the authors of the previous study (27). All three genes are known to be involved in the glycan synthesis pathways. *B3GAT1* gene encodes beta-1,3-glucuronyltransferase 1. According to SMR/HEIDI analysis, the same functional variant possibly mediates the association of *B3GAT1*-containing locus on chromosome 11 with glycan trait as well as with the level of *B3GAT1* expression in whole blood and lung tissue. The *FUT8* locus was associated mostly with core-fucosylated biantennary glycans, which are known to be linked to the immunoglobulins (51). Since the *FUT8* gene encodes fucosyltransferase 8, an enzyme responsible for the addition of core fucose to glycans, this gene is the most biologically plausible in this locus. Evidence for prioritization of this gene in our study was also provided by DEPICT.

FUT3 and *FUT6* encode fucosyltransferases 3 and 6 that catalyze the transfer of fucose from GDP-beta-fucose to alpha-2,3 sialylated substrates. The *FUT3/FUT6* locus was associated with antennary fucosylation of tri- and tetra-antennary sialylated glycans, and therefore we consider these genes as good candidates. Moreover, in the *FUT6* gene (chromosome 19, 58 Mb), we found the missense variant rs17855739 (substitution G > A) that leads to the amino acid change from negatively charged glutamic acid to positively charged lysine. FATHMM-XF predicted this variant as pathogenic for transcripts of *FUT6* gene. Thus, we can consider this SNP as a potentially causal functional variant. DEPICT prioritized the *FUT3* gene.

For two other loci (on chromosome 2 at 135 Mb and on chromosome 12 at 121 Mb), we were not able to prioritize genes by the DEPICT and eQTL analyses. However, the first locus contained the *MGAT5* gene encoding mannosyl-glycoprotein-N-acetyl glucosaminyl-transferase that is involved in the glycan synthesis pathways. The second locus contained several genes including *HNF1A*, which was previously shown to co-regulate the expression of most fucosyltransferase (*FUT3, FUT5, FUT6, FUT8, FUT10* and *FUT11*) genes in a human liver cancer cell line (HepG2 cells) as well as to co-regulate expression genes encoding key enzymes needed for synthesis of GDP-fucose, the substrate for fucosyltransferases, thereby regulating multiple stages in the fucosylation process (26). Thus, we considered *HNF1A* as the candidate gene for this locus.

Four of the seven novel loci contain genes that are known to be involved in glycan synthesis pathways—*ST6GAL1, ST3GAL4, B4GALT1* and *MGAT3* (see Table 3 and Fig. 1). Moreover, SMR and HEIDI analyses have shown that expression of *ST6GAL1, ST3GAL4* and *MGAT3* genes may mediate the association between corresponding loci and plasma N-glycome. *ST6GAL1* and *ST3GAL4* genes encode sialyltransferases, enzymes that catalyze the addition of sialic acid to various glycoproteins. The locus-containing *ST6GAL1* was associated with ratio of sialylated and non-sialylated galactosylated biantennary glycans. The locus containing *ST3GAL4* was associated with galactosylated sialylated tri- and tetra-antennary glycans. The locus containing *MGAT3* was associated with core-fucosylation of bisected glycans, which is in line with the known effect of GnT-III (product of *MGAT3* gene) on FUT8 activity (52). For this locus, we have found two possible functional variants: non-coding variant rs7423 and in-frame deletion rs149306472 that was predicted to be pathogenic for the product of the *SYNGR1* gene. The *SYNGR1* gene encodes an integral membrane protein associated with presynaptic vesicles in neuronal cells. Since *MGAT3* has a known role in glycan biosynthesis, we choose *MGAT3* as the candidate gene for this locus. The *B4GALT1* gene encodes galactosyltransferase, which adds galactose during the biosynthesis of different glycoconjugates. This gene was associated with galactosylation of biantennary glycans. Thus, we observe consistency between known enzymatic activities of the products of selected candidate genes and the spectrum of glycans that are associated with corresponding loci.

The other three novel loci do not contain genes that are known to be directly involved in glycan synthesis (Fig. 1). Variant rs9624334 (chromosome 22 at 24 Mb) is located in the intron of *SMARCB1* gene that is known to be important in antiviral activity, inhibition of tumor formation, neurodevelopment, cell proliferation and differentiation (53). However, gene prioritization analysis (DEPICT) showed that the possible candidate gene is *DERL3*, which encodes a functional component of endoplasmic reticulum (ER)-associated degradation for misfolded luminal glycoproteins (54) (see Table 3). Additionally,

Table 3. Summary of *in-silico* functional annotation for 12 replicated loci

Locus	Nearest gene	Candidate Gene	CV	SMR/HEIDI	D	Funct. studies	Glycan synth.	Prev. annot.
Previously implicated loci								
2:135015347	*MGAT5*	*MGAT5*	–				+	Pl
11:134274763	*B3GAT1*	*B3GAT1*	–	Whole blood/lung			+	Pl
12:121436376	*HNF1A*	*HNF1A*	–			(27)	+	Pl
14:66011184	*FUT8*	*FUT8*	–		FDR < 20%		+	Pl, IgG
19:5828064	*NRTN*	*FUT3*	–		FDR < 20%		+	Pl
		FUT6	rs17855739				+	Pl, IgG
Novel loci								
3:186722848	*ST6GAL1*	*ST6GAL1*	–	Whole blood			+	IgG
7:50355207	*IKZF1*	*IKZF1*	–					IgG
9:33128617	*B4GALT1*	*B4GALT1*	–				+	IgG
11:126232385	*ST3GAL4*	*ST3GAL4*	–	Muscle skeletal/pancreas			+	
14:105989599	*C14orf80*	*TMEM121*	–	Whole blood				
		IGH	–					IgG
22:24166256	*SMARCB1*	*DERL3*	rs3177243		FDR < 20%			IgG
		CHCHD10	–	Whole blood				
22:39859169	*MGAT3*	*MGAT3*	–	CD19+ (B cells)			+	IgG
		SYNGR1	rs149306472, rs7423	Nerve tibial				

For each locus, we report the gene nearest to the top SNP and plausible candidate genes. CV—variant with predicted (by FATHMM-XF or FATHMM-INDEL) pathogenic impact on the gene; SMR/HEIDI—evidence for pleiotropy with expression demonstrated by SMR-HEIDI analysis; D—evidence provided by DEPICT analysis; Funct. studies—evidence from functional studies; Glycan synth.—known genes involved in glycan synthesis or its regulation are present in the locus; Prev. annot.—the region has previously been revealed in glycome GWAS, and the gene was suggested as candidate [Pl—gene was reported as affecting plasma protein N-glycome by Huffman *et al.* (27); IgG—gene was reported as affecting IgG glycome either by Lauc *et al.* (43) and/or by Shen *et al.* (56)].

FATHMM analysis demonstrated missense-coding *DERL3* variant rs3177243 to be potentially pathogenic. This polymorphism is in strong LD ($r^2 = 0.98$ in 1000 Genome EUR samples) with the leading SNP rs9624334. However, the SMR/HEIDI analysis suggested that the association with N-glycome could be (also) mediated by expression of *CHCHD10* gene, which encodes a mitochondrial protein that is enriched at cristae junctions in the intermembrane space. The *CHCHD10* gene has the highest expression in heart and liver and the lowest expression in spleen (55). While the role of mitochondrial proteins in glycosylation processes remains speculative, we propose *CHCHD10* as a candidate based on our eQTL pleiotropy analysis. Thus, we consider two genes—*DERL3* and *CHCHD10*—as possible candidate genes in this locus. Interestingly, this and the *MGAT3* loci were associated with similar glycan traits (core-fucosylation of bisected glycans). This indicates that core fucosylation of bisected glycans is under joint control of *MGAT3* and *DERL3/CHCHD10*.

The locus on chromosome 14 at 105 Mb contains the *IGH* gene that encodes immunoglobulin heavy chains. This locus was associated with sialylation of core-fucosylated biantennary monogalactosylated structures that are biochemically close to those influenced by the *ST6GAL1*-containing locus. Since IgG is the most prevalent glycosylated plasma protein (51), we can consider *IGH* a good candidate, as indeed was suggested by Shen and colleagues (56). However, our functional annotation results (SMR/HEIDI) suggested that association of this locus with plasma N-glycome may be mediated by *TMEM121* gene. Therefore, we consider two genes—*IGH* and *TMEM121*—as candidate genes for this locus.

For the locus on chromosome 7 at 50 Mb, we were not able to select a candidate gene based on the results of our *in-silico* functional annotation. This locus was previously reported to be associated with glycan levels of IgG (43), and the authors suggested that *IKZF1* may be considered a candidate gene in the region. The *IKZF1* gene encodes the DNA-binding protein Ikaros that acts as a transcriptional regulator and is associated with chromatin remodeling. It is considered an important regulator of lymphocyte differentiation. Taking into account that IgG [the most abundant glycoprotein in the blood plasma (51)] is secreted by B cells (57), *IKZF1* seems to be a plausible candidate gene.

Gene network regulating N-glycosylation

To identify possible clusters in the gene network of plasma protein N-glycosylation, we draw a graph in which 11 genome-wide significant loci and genome-wide significantly ($P < 1.66 \times 10^{-9}$) associated glycan traits were presented as nodes and edges represent observed associations (Fig. 3). We labeled each glycan trait as 'immunoglobulin-linked' (Ig-linked), 'non-immunoglobulin-linked' (non-Ig-linked) or mixed depending on the contribution of Ig- and non-Ig-linked glycans to the trait value (Supplementary Material, Table 10), which was inferred based on the information about protein-specific glycosylation reported previously (51). For more details about the procedure of Ig/non-Ig/mixed assignment, see Supplementary Material, Note 4.

The resulting network shows that candidate genes and glycan traits cluster into two major subnetworks or hubs (Fig. 3). The first subnetwork contained the six loci: *FUT8, DERL3/CHCHD10, IKZF1, TMEM121, ST6GAL1* and *MGAT3*, with *FUT8* as a hub. These loci, as well as the locus containing *B4GATL1*, were associated with core-fucosylated biantennary glycans. It is known that the majority of plasma core-fucosylated biantennary glycans are linked to immunoglobulins (51). Moreover, in previous studies, these seven genes were found to be associated with N-glycosylation of IgG (43,56). At the same time, these genes were associated with non-Ig-linked glycans. We can consider this cluster (7 genes out of 11) as related to both IgG and non-IgG glycosylation. Taking into account that IgG is the most prevalent glycosylated plasma protein, it is not surprising that more than a half of replicated loci are actually associated with immunoglobulins glycosylation. However, previous GWAS on HPLC plasma N-glycome reported only one locus—*FUT8*—as overlapping with IgG glycosylation-associated loci.

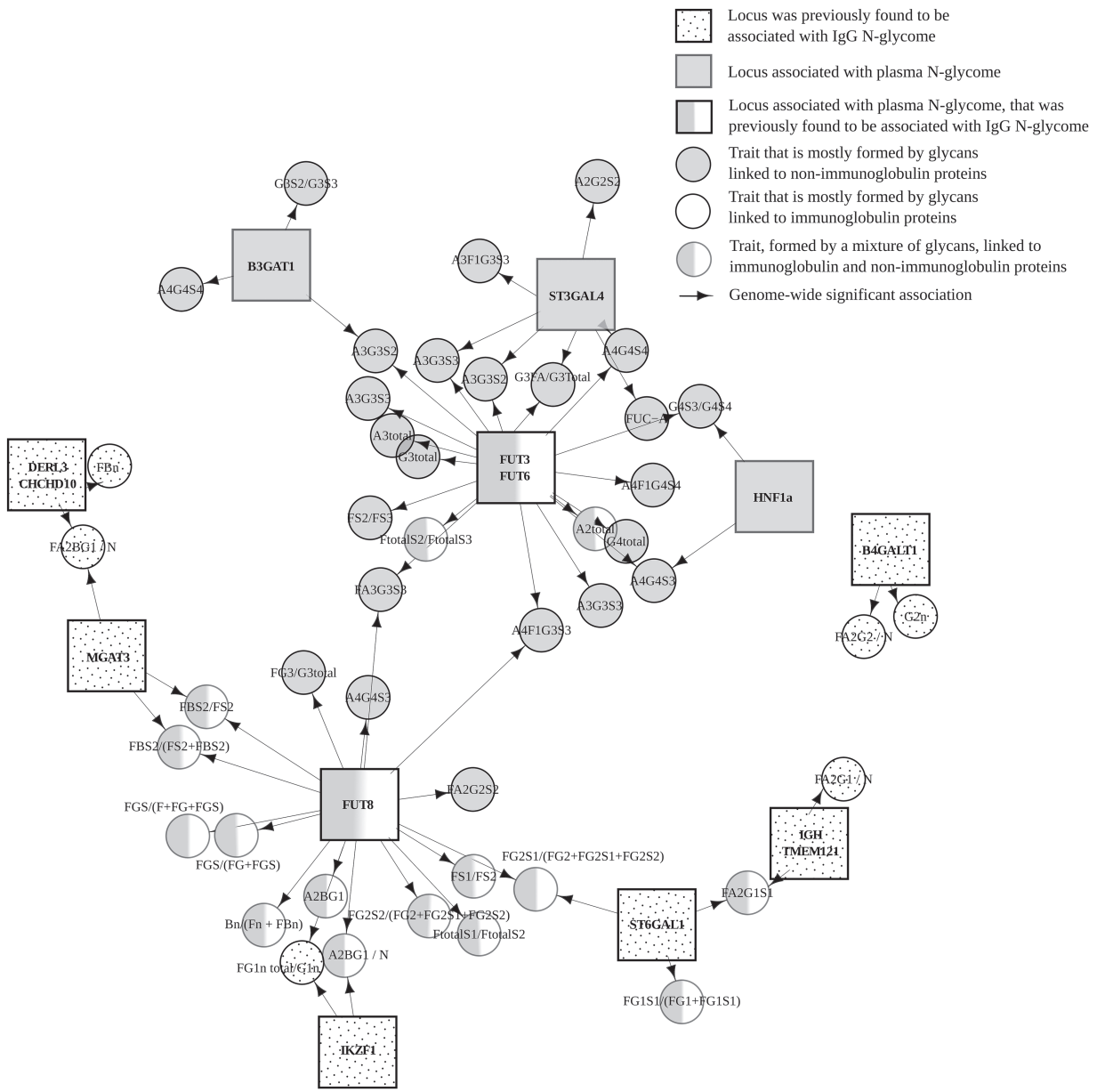

Figure 3. A network view of associations between loci and glycan traits. Square nodes represent genetic loci labeled with the names of candidate gene(s), circle nodes represent glycan traits. Squares with polka dot pattern represent candidate genes, located in genomic regions that were previously found to be associated with IgG N-glycome. Grey squares represent candidate genes, located in genomic regions associated with plasma N-glycome. Grey/white squares represent candidate genes, located in genomic regions associated with plasma N-glycome, that were previously found to be associated with IgG N-glycome. Circles with polka dot pattern highlights glycan traits mostly containing glycans that are linked to immunoglobulins. Grey circles represent traits that are mostly formed by glycans linked to other (not immunoglobulin) proteins. Grey/white circles represent glycan traits, formed by a mixture of glycans that are linked to immunoglobulin and non-immunoglobulin proteins. Arrows represent genetic association ($P < 1.66 \times 10^{-9}$) between gene and specific glycan.

The second subnetwork contained four loci (*ST3GAL4*, *HNF1A*, *FUT3/FUT6* and *B3GAT1*, with *FUT3/FUT6* as a hub) associated with tri- and tetra-antennary glycans. It is known that these types of glycans are linked to plasma proteins other than IgG (51). Thus, we attributed this cluster to non-IgG plasma protein N-glycosylation. Among these four loci, we report *ST3GAL4* as the novel locus controlling N-glycosylation of non-IgG plasma proteins. We attribute it to non-immunoglobulins plasma protein N-glycosylation owing to its association with tetra-antennary glycans.

Discussion

We conducted the first GWAS of total plasma N-glycome measured by UPLC technology. Our efforts brought the number of loci significantly associated with total plasma N-glycome from 6 (26,27) to 16, of which 12 were replicated in our study. This allowed us to use a range of *in silico* functional genomics analyses and identify candidate genes in the established loci.

Compared to the HPLC glycan measurement technology used in previous GWAS of plasma N-glycome (26,27), UPLC technology provides better resolution and quantification of glycan

structures, resulting in increased power of association testing. Thus, despite the reduced sample size in our study [2763 samples here vs. 3533 samples in the study reported by Huffman *et al.* (27)], we have detected 14 vs. 6 plasma N-glycome QTLs. In particular, we revealed and replicated novel loci containing *ST3GAL4* and *ST6GAL1*. An interesting question is why these two strongly associated loci were not detected in previous HPLC-based GWAS. In our study, *ST3GAL4* locus showed the strongest association with PGP17 peak (which corresponds to A2G2S[3,6+3]2 trait; $P = 8.6 \times 10^{-28}$). On an average UPLC chromatogram, a nearby PGP19 peak has 8× and 28× bigger area than PGP17 and PGP18 peaks, respectively. On the HPLC chromatogram, these three peaks are merged into one GP9 peak. Thus, GP9 HPLC peak is mostly formed by PGP19. In our study, we revealed association of *ST3GAL4* locus neither with PGP18 ($P = 6.14 \times 10^{-1}$) nor with PGP19 ($P = 8.25 \times 10^{-4}$) peaks. We can therefore assume that association in the previous study (27) was not detected due to a small signal-to-noise ratio. We suggest the same reasoning for the *ST6GAL1* locus. In our study it was associated with PGP13 peak ($P = 3.12 \times 10^{-23}$), which together with PGP12 forms GP6 HPLC peak. *ST6GAL1* locus showed association neither with GP6 in Huffman *et al.* study ($P = 0.012$) (27) nor with PGP12 in our study ($P = 3.19 \times 10^{-2}$). PGP12 peak has 1.5× bigger area than PGP13. Thus, we propose that *ST6GAL1* has an exclusive effect on PGP13 (FA2[3]G1S[3+6]1) trait.

It should be noted that we used a new imputation panel (1000 Genomes instead of HapMap in the previous studies) that more than tripled the number of polymorphisms analyzed (from 2.4 M SNPs to 8 M). That may have contributed to the higher power of our study as well.

In addition to detecting novel loci, we were able to replicate five (*HNF1A*, *FUT6*, *FUT8*, *B3GAT1* and *MGAT5*) out of six loci that were reported previously to be associated with human plasma N-glycome measured using the HPLC technology (26,27). However, 3 out of 10 novel loci have not been replicated. For two of them, associated SNPs had relatively low (<2%) MAF. For rare variants, multiple testing burden is increased compared to common variants, which may lead to higher than expected false positive rate (58) in discovery. At the same time, relative effects of drift that are more pronounced for rare variants (59,60) may increase false negative rate in replication. The third unreplicated locus is located in HLA region known to have very complex structure (61). This locus showed highest (although not significant) heterogeneity of association (*P*-value of Cochran's Q-test = 0.07) among loci in the replication meta-analysis.

Among six plasma glycome loci that were identified as genome-wide significant previously (26,27), only one (region of *FUT8*) had overlap with a locus identified as associated with IgG glycome composition (43). A recent multivariate GWAS study of plasma IgG glycome composition (56) identified five new loci, including the region of *FUT3/FUT6*, thus bringing the overlap between plasma and IgG glycome loci to two. In our study, among 12 replicated loci, the majority (8 loci) overlapped with loci that were reported to be associated with IgG glycome composition (43,56) (Fig. 3). We therefore established a strong overlap between IgG and plasma glycome loci.

In a way, this overlap is to be expected. It is known that majority of serum (and therefore plasma) glycoproteins are either immunoglobulins produced by B-lymphocytes or glycoproteins secreted by the liver (62). We thus expected overlap between IgG and total plasma glycome loci, and we expected that loci associated with the plasma N-glycome would be enriched by genes with tissue specific expression in liver and B cells. Indeed, we find that plasma N-glycome loci are enriched by genes expressed in plasma cells, antibody producing-cells and B-lymphocytes, and we also find overlap between plasma N-glycome loci and CD19+ eQTLs. However, we neither find enrichment of genes that are expressed in liver (Supplementary Material, Table 5C) nor overlap between plasma N-glycome loci and liver eQTLs. In the future, it will be important to achieve better resolution and separation of loci that are related to glycosylation of non-immunoglobulin glycoproteins. This could be achieved either technologically (e.g. performing analyses of IgG-free fractions of proteins), or this could be attempted via statistical modeling.

The genetic variation in the *FUT3/FUT6* locus is a major (in terms of proportion of variance explained and number of glycans affected) genetic factor for non-immunoglobulins glycosylation. According to current knowledge, these enzymes catalyze fucosylation of antennary GlcNAc32, resulting in glycan structures that are not found on IgG (51,63). This is consistent with the spectrum of glycan traits associated with *FUT3/FUT6* locus in our work (Fig. 3). However, this locus was recently found to be associated with IgG glycosylation (56). The authors could not explain this finding because at that time IgG glycans were not known to contain antennary fucose. Two explanations could have been proposed for this surprising finding: either enzymes encoded by *FUT3/FUT6* locus exhibit non-canonical activity of core fucosylation or some IgG glycans actually do contain antennary fucose. Recently, the latter was demonstrated in the study by Russell *et al.* (10). Based on our results showing no evidence for association between *FUT3/FUT6* locus and core fucosylation, we can speculate that association of this locus with IgG glycosylation could be explained by the presence of antennary fucose on some IgG-linked glycans.

An interesting pattern starts emerging out of study of genetic control of plasma glycosylation. We now see a clear overlap in genetic control between plasma and IgG glycosylation, which calls for future studies that would help distinguish among global, cell-specific, tissue-specific and protein-specific pathways of protein glycosylation. Many (8 out of 12) replicated loci contained genes that encode enzymes directly involved in glycosylation (*FUT3/FUT6*, *FUT8*, *B3GAT1*, *ST6GAL1*, *B4GALT1*, *ST3GAL4*, *MGAT3* and *MGAT5*). However, glycosylation results from a complex interplay not only of enzymes responsible for transfer of monosaccharides to a growing glycan chain but also of enzymes involved in biosynthesis of individual monosaccharides and mechanism ensuring timely localization of all components involved in the process. Moreover, a number of environmental factors, such as diet and smoking, were found to be associated with plasma N-glycome composition (25). Studies in mouse models suggested complex compensatory mechanisms to pay role in glycosylation, which can complicate interpretation of the observed associations (64,65). We now start seeing loci and genes, which are likely to reflect complex aspects of plasma protein glycosylation, such as regulation of fucosylation by HNF1a (26). Such regulatory genes, in our view, are plausible candidates that will help linking glycans with complex human diseases. The results of our study provide evidence for potential role of genes *DERL3*, *CHCHD10*, *TMEM121*, *IGH* and *IKZF1*, although for some loci we could not prioritize only one particular gene (e.g. for locus containing *DERL3* and *CHCHD10*). These genes can be considered candidates for future experimental research. To facilitate further studies of glycosylation and of the role of glycome in human health and diseases we have made full results of our plasma N-glycome GWAS (almost 1 billion of trait–SNP associations) freely available to the scientific community via GWAS archive.

Previous GWAS of HPLC measured plasma N-glycome (27) identified six genes controlling plasma N-glycosylation of which four implicated genes with obvious links to the glycosylation process. Here, using a smaller sample but more precise UPLC technology and new GWAS imputation panels, we confirmed the association of five known loci and identified and replicated additional seven loci. Our results support the idea that genetic control of plasma protein N-glycosylation is a complex process, which is under control of genes that belong to different pathways and are expressed in different tissues. Further studies with larger sample size are warranted to further decrypt the genetic architecture of the glycosylation process and explain the relations between glycosylation and mechanisms of human health and disease.

Materials and Methods

Study cohort description

This work is based on analysis of data from four cohorts—TwinsUK, PainOR, SOCCS and QMDiab. Sample demographics can be found in Supplementary Material, Table 11.

TwinsUK. The TwinsUK cohort (66) (also referred to as the UK Adult Twin Register) is a nationwide registry of volunteer twins in the UK, with about 13 000 registered twins (83% female, equal number of monozygotic and dizygotic twins, predominantly middle-aged and older). The Department of Twin Research and Genetic Epidemiology at King's College London hosts the registry. From this registry, a total of 2763 subjects had N-linked total plasma glycan measurements that were included in the analysis.

QMDiab. The QMDiab is a cross-sectional case–control study with 374 participants. QMDiab has been described previously and comprises male and female participants in near equal proportions, aged between 23 and 71 years, mainly of Arab, South Asian and Filipino descent (67,68). The initial study was approved by the institutional review boards of Hamad Medical Corporation (HMC) and Weill Cornell Medicine—Qatar (WCM-Q) (research protocol #11131/11). Written informed consent was obtained from all participants. All study participants were enrolled between February 2012 and June 2012 at the Dermatology Department of HMC in Doha, Qatar. Inclusion criteria were a primary form of type 2 diabetes (for cases) or an absence of type 2 diabetes (for controls). Sample collection was conducted in the afternoon, after the general operating hours of the morning clinic. Patient and control samples were collected in a random order as they became available and at the same location using identical protocols, instruments and study personnel. Samples from cases and controls were processed in the laboratory in parallel and in a blinded manner. Data from five participants were excluded from the analysis because of incomplete records, leaving 176 patients and 193 controls. Of the 193 control participants initially enrolled, 12 had HbA1c levels above 6.5% (48 mmol/mol) and were subsequently classified as cases, resulting in 188 cases and 181 controls.

SOCCS. SOCCS study (69,70) comprised 2057 (colorectal cancer) CRC cases (61% male; mean age at diagnosis, 65.8 ± 8.4 years) and 2111 population controls (60% males; mean age ,67.9 ± 9.0 years) as ascertained in Scotland. Cases were taken from an independent, prospective, incident CRC case series and aged <80 years at diagnosis. Control subjects were population controls matched by age (±5 years), gender and area of residence within Scotland. All participants gave written informed consent and study approval was from the MultiCentre Research Ethics Committee for Scotland and Local Research Ethics committee. Sample collection is described in (69,70).

PainOR. The PainOR (71) is the University of Parma cohort of patients of a retrospective multicenter study (ClinicalTrials.gov Identifier: NCT02037789) part of the PainOMICS project funded by European Community in the Seventh Framework Programme (Project ID: 602736). The primary objective is to recognize genetic variants associated with chronic low back pain (CLBP); secondary objectives are to study glycomics and Activomics profiles associated with CLBP. Glycomic and Activomic approaches aim to reveal alterations in proteome complexity that arise from post-translational modification that varies in response to changes in the physiological environment, a particularly important avenue to explore in chronic inflammatory diseases. The study was firstly approved by the institutional review boards of Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Foundation San Matteo Hospital Pavia and then by the institutional review boards of all clinical centers that enrolled patients. Copies of approvals were provided to the European Commission before starting the study. Written informed consent was obtained from all participants. In the period between September 2014 and February 2016, 1000 patients (including 38.1% male and 61.9% female, averaging 65 ± 14.5 years) were enrolled at the Anesthesia, Intensive Care and Pain Therapy Department of University Parma Hospital. Inclusion criteria were adult Caucasian patients who were suffering of lower back pain (pain between the costal margins and gluteal fold, with or without symptoms into one or both legs) more than 3 months who were admitted at Pain Department of University Parma Hospital. We exclude patients with recent history of spinal fractures or lower back pain due to cancer or infection. Sample collection was performed in all patients enrolled, according to the standard operating procedures published in *PlosOne* in 2017 (72). Samples were processed in PainOmics laboratory in a blinded manner in University of Parma.

Genotyping

For full details of the genotyping and imputation see Supplementary Material, Table 12.

TwinsUK. Genotyping was carried out using combination Illumina SNP arrays: HumanHap300, HumanHap610Q, 1 M-Duo and 1.2MDuo 1 M. Standard quality control of genotyped data was applied, with SNPs filtered by sample call rate > 98%; MAF > 1%; SNP call rate, >97% (for SNP with MAF ≥ 5%) or >99% (for SNPs with 1% ≤ MAF < 5%); and HWE, $P < 1 \times 10^{-6}$. In total 275 139 SNPs passed criteria. Imputation was done using IMPUTE2 software with 1000G phase 1 version 3 and mapped to the GRCh37 human genome build. Imputed SNPs were filtered by imputation quality (SNPTEST proper-info) > 0.7, MAF ≥ 1% and MAC ≥ 10, leading to 8 557 543 SNPs passed to the GWAS analysis.

QMDiab. Genotyping was carried out using Illumina Omni array 2.5 (version 8). Standard quality control of genotyped data was applied, with SNPs filtered by sample call rate > 98%; MAF > 1%; SNP call rate, > 98%; and HWE $P \times < \times 1 \times 10^{-6}$. In total 1 223 299 SNPs passed criteria. Imputation was done using SHAPEIT software with 1000G phase 3 version 5 and mapped to the GRCh37 human genome build. Imputed SNPs were filtered

by imputation quality >0.7, leading to 20 483 276 SNPs passed to the GWAS analysis.

SOCCS. Details of the genotyping procedure can be found here (73). Genotyping was carried out using Illumina SNP arrays: HumanHap300 and HumanHap240S. Standard quality control of genotyped data was applied, with SNPs filtered by sample call rate > 95%; MAF > 1%; SNP call rate, >95%; and HWE, $P < 1 \times 10^{-6}$. In total 514 177 SNPs passed criteria. Imputation was done using SHAPEIT and IMPUTE2 software with 1000 Genomes, phase 1 (Integrated haplotypes, released June 2014) and mapped to the GRCh37 human genome build. Imputed SNPs were not filtered, leading to 37 780 221 SNPs passed to the GWAS analysis.

PainOR. Genotyping was carried out using Illumina HumanCore BeadChip. Standard quality control of genotyped data was applied with SNPs filtered by sample call rate > 98%; MAF > 0.625%; SNP call rate, >97%; and HWE, $P < \times 1 \times 10^{-6}$. In total 253 149 SNPs passed criteria. Imputation was done using Eagle software with HRC r1.1 2016 reference and mapped to the GRCh37 human genome build. Imputed SNPs were not filtered, leading to 39 127 685 SNPs passed to the GWAS analysis.

Phenotyping

Plasma N-glycome quantification

Plasma N-glycome quantification of samples from TwinsUK, PainOR and QMDiab were performed at Genos by applying the following protocol. Plasma N-glycans were enzymatically released from proteins by PNGase F, fluorescently labeled with 2-aminobenzamide and cleaned up from the excess of reagents by hydrophilic interaction liquid chromatography solid phase extraction (HILIC-SPE), as previously described. (74). Fluorescently labeled and purified N-glycans were separated by HILIC on a Waters BEH Glycan chromatography column, 150 × 2.1 mm, 1.7 µm BEH particles, installed on an Acquity UPLC instrument (Waters, Milford, MA, USA) consisting of a quaternary solvent manager, sample manager and a fluorescence detector set with excitation and emission wavelengths of 250 and 428 nm, respectively. Following chromatography conditions previously described in details (74), total plasma N-glycans were separated into 39 peaks for QMDiab, TwinsUK and PainOR cohorts. The amount of N-glycans in each chromatographic peak was expressed as a percentage of total integrated area. Glycan peaks (GPs)—quantitative measurements of glycan levels—were defined by automatic integration of intensity peaks on chromatogram. The number of defined GPs varied among studies from 36 to 42 GPs.

Plasma N-glycome quantification for SOCCS samples were done at the National Institute for Bioprocessing Research and Training (NIBRT) by applying the same protocol as for TwinsUK, PainOR and QMDiab, with the only difference in the excitation wavelength (330 nm instead of 250 nm).

Harmonization of GPs

The order of the GPs on a UPLC chromatogram was similar among the studies. However, depending on the cohort some peaks located near one another might have been indistinguishable. The number of defined GPs varied among studies from 36 to 42. To conduct GWAS on TwinsUK following by replication in other cohorts, we harmonized the set of peaks (or GPs). According to the major glycostructures within the GPs we manually created the table of correspondence between different GPs (or sets of GPs) across all cohorts, where plasma glycome was measured using UPLC technology. Then, based on this table of correspondence, we defined the list of 36 harmonized GPs (Supplementary Material, Table 13) and the harmonization scheme for each cohort. We validated the harmonization protocol by comparing with manual re-integration of the peaks on chromatogram level using 35 randomly chosen samples from three cohorts: TwinsUK, PainOR and QMDiad. We show the full concordance between two approaches (Pearson correlation coefficient R > 0.999, see Supplementary Material, Table 14 for the details). We applied this harmonization procedure for the four cohorts: TwinsUK, QMDiab, CRC and PainOR, leading to the set of 36 glycan traits in each cohort.

Normalization and batch correction of GPs

Normalization and batch correction were performed on harmonized UPLC glycan data for four cohorts: TwinsUK, PainOR, SOCCS and QMDiab. We used total area normalization (the area of each GP was divided by the total area of the corresponding chromatogram). Normalized glycan measurements were log10-transformed due to right skewness of their distributions and the multiplicative nature of batch effects. Prior to batch correction, samples with outlying measurements were removed. Outlier was defined as a sample that had at least one GP that is out of three standard deviations from the mean value of GP. Batch correction was performed on log10-transformed measurements using the ComBat method (75), where the technical source of variation (batch and plate number) was modeled as a batch covariate. Again, samples with outlying measurements were removed.

From the 36 directly measured glycan traits, 77 derived traits were calculated (see Supplementary Material, Table 10). These derived traits average glycosylation features such as branching, galactosylation and sialylation across different individual glycan structures, and consequently, they may be more closely related to individual enzymatic activity and underlying genetic polymorphism. As derived traits represent sums of directly measured glycans, they were calculated using normalized and batch-corrected glycan measurements after transformation to the proportions (exponential transformation of batch-corrected measurements). The distribution of 113 glycan traits can be found in Supplementary Material, Figure 2.

Prior to GWAS, the traits were adjusted for age and sex by linear regression. The residuals were rank transformed to normal distribution [rntransform function in GenABEL (76,77) R package].

Genome-wide association analysis

Discovery GWAS was performed using TwinsUK cohort ($N = 2763$) for 113 GP traits. Genome-wide Efficient Mixed Model Association algorithm (GEMMA) (78) was used to estimate the kinship matrix and to run linear mixed model regression on SNP dosages assuming additive genetic effects. Obtained summary statistics were corrected for genomic control inflation factor λ_{GC} to account for any residual population stratification. An association was considered statistically significant at the genome-wide level if the P-value for an individual SNP was less than $5 \times 10^{-8}/(29 + 1) = 1.66 \times 10^{-9}$, where 29 is an effective number of tests (traits) that was estimated as the number of principal components that jointly explained 99% of the total plasma glycome variance in the TwinsUK sample.

Locus definition

In short, we considered SNPs located in the same locus if they were located within 500 Kb from the leading SNP (the SNP with lowest *P*-value). Only the SNPs and the traits with lowest *P*-values are reported (leading SNP-trait pairs). The detailed procedure of locus definition is described in Supplementary Material, Note 1.

Replication

We have used TwinsUK cohort for the replication of six previously described loci (27) affecting plasma N-glycome. From each of six loci we have chosen leading SNP with the strongest association as reported by authors (27). Since there is no direct trait-to-trait correspondence between glycan traits measured by HPLC and UPLC technologies we tested the association of the leading SNPs with all 113 glycan traits in TwinsUK cohort. We considered locus as replicated if its leading SNP showed association with at least one of 113 glycan traits with replication threshold of $P < 0.05/(6*30) = 2.78 \times 10^{-4}$, where six is number of loci and 30 is a number of principal components that jointly explained 99% of the total plasma N-glycome variance.

For the replication of novel associations, we used data from three cohorts: PainOR ($N = 294$), QMDiab ($N = 327$) and SOCCS ($N = 472$) with total replication sample size of $N = 1048$ samples that have plasma UPLC N-glycome and genotype data (for details of genotyping, imputation and association analysis; see Supplementary Material, Table 12). We used only the leading SNPs and traits for the replication that were identified in the discovery step. For these SNPs we conducted a fixed-effect meta-analysis using METAL software (79) combining association results from three cohorts. The replication threshold was set as $P < 0.05/10 = 0.005$, where 10 is the number of replicated loci. Moreover, we checked whether the sign of estimated effect was concordant between discovery and replication studies.

Functional annotation *in silico*

Functional annotation of associated variants. All SNPs and indels in high LD ($r^2 > 0.6$) with the 12 lead variants at replicated loci were selected. LD was calculated using genotype data for 503 samples with European descent from 1000 Genomes phase 3 version 5 data and Plink tool (80) (version 1.9) using *–show-tags* option. Additionally, for each of the 12 replicated loci we have selected the set of SNPs that had strong associations, defined as those located within +/−250 kbp window from the strongest association, and having *P*-value < T, where log10(T) = log10(Pmin) + 1, where Pmin is the *P*-value of the strongest association in the locus. This additional inclusion criteria was applied since genotype data for TwinsUK samples was imputed using 1000 Genomes phase 1 version 3 panel and some of the SNPs from this panel are not exists in 1000 Genomes phase 3 version 5 panel. The list of selected variants can be found in Supplementary Material, Table 4A. Next all selected variants passed to functional annotation using Ensembl Variant Effect Predictor (VEP) method (33). Then, we used FATHMM-XF (34) and FATHMM-INDEL (35) methods to predict the impact of SNPs and small indels. Predictions are given as *P*-values in the range [0, 1]: values above 0.5 are predicted to be deleterious, while those below 0.5 are predicted to be neutral or benign. *P*-values close to the extremes (0 or 1) are the highest-confidence predictions that yield the highest accuracy.

Gene-set and tissue/cell enrichment analysis. To prioritize genes in associated regions, gene set enrichment and tissue/cell type enrichment analyses were carried out using DEPICT software v. 1 rel. 194 (32). For the analysis we have chosen independent variants (see 'Locus definition') with $P < 5 \times 10^{-8}/30$ (14 SNPs) and $P < 1 \times 10^{-5}/30$ (93 SNPs). We used 1000G dataset for calculation of LD (81). GO enrichment analysis was performed using FUMA GENE2FUNC (82) analysis based on MsigDB c5 (83) and MAGMA (37) with default parameters and '*All genes*' as background genes.

Pleiotropy with complex traits. We have investigated the overlap between associations obtained here and elsewhere, using PhenoScanner v1.1 database (42). For 12 replicated SNPs (Table 1; Table 2) we looked up traits that have demonstrated genome-wide significant ($P < 5 \times 10–8$) association at the same or at strongly ($r^2 > 0.7$) linked SNPs.

Pleiotropy with eQTLs. To identify genes whose expression levels could potentially mediate the association between SNPs and plasma glycan traits we performed a SMR analysis followed by HEIDI method (38). In short, SMR test aims at testing the association between gene expression (in a particular tissue) and a trait using the top associated eQTL as a genetic instrument. Significant SMR test indicates not only evidence of causality or pleiotropy but also the possibility that SNPs controlling gene expression are in linkage disequilibrium with those associated with the traits. These two situations can be disentangled using the HEIDI test.

The SMR/HEIDI analysis was carried out for leading SNPs that were replicated and were genome-wide significant ($P < 1.7 \times 10^{-9}$) on discovery stage (11 loci in total, see Table 1). We checked for overlap between these loci and eQTLs in blood (39), 44 tissues provided by the GTEx database (40) and in nine cell lines from CEDAR dataset (41), including six circulating immune cell types (CD4+ T-lymphocytes, CD8+ T lymphocytes, CD19+ B lymphocytes,CD14+ monocytes, CD15+ granulocytes and platelets). Technical details of the procedure may be found in Supplementary Material, Note 2. Following Bonferroni procedure, the results of the SMR test were considered statistically significant if $P_{SMR} < 2.445 \times 10^{-6}$ (0.05/20448, where 20448 is a total number of probes used in analysis for all three data sets). Inferences whether functional variant may be shared between plasma glycan trait and expression were made based on HEIDI test: $P_{HEIDI} > 0.05$ (likely shared), $0.05 > P_{HEIDI} > 0.001$ (possibly shared) and $P_{HEIDI} < 0.001$ (sharing is unlikely).

Data availability

Summary statistics from our plasma N-glycome GWAS for 113 glycan traits are available for interactive exploration at the GWAS archive (http://gwasarchive.org). The dataset was also deposited at Zenodo (http://doi.org/10.5281/zenodo.1298406) (84). The data generated in the secondary analyses of this study are included with this article in the supplementary tables.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We thank all staff at Weill Cornell Medicine - Qatar and Hamad Medical Corporation, and especially all study participants who made the QMDiab study possible.

The SOCCS study was supported by grants from Cancer Research UK (C348/A3758, C348/A8896, C348/ A18927); Scottish Government Chief Scientist Office (K/OPR/2/2/D333, CZB/4/94); Medical Research Council (G0000657-53203, MR/K018647/1); Centre Grant from CORE as part of the Digestive Cancer Campaign (http://www.corecharity.org.uk).

TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

Conflict of Interest statement. Y.A. is a founder and co-owner of Maatschap PolyOmica, a private organization, providing services, research and development in the field of computational and statistical (gen)omics, and PolyKnomics, one of the organization providing the "GWASArchive". G.L. is a founder and owner of Genos Ltd, a biotech company that specializes in glycan analysis and has several patents in the field. All other authors declare no conflicts of interest. Other authors declare no competing financial interests.

Funding

European Community's Seventh Framework Programme-funded project PainOmics (602736); European Structural and Investment Funds IRI (KK.01.2.1.01.0003); Croatian National Centre of Research Excellence in Personalized Healthcare (KK.01.1.1.01.0010); Russian Ministry of Science and Education under the 5-100 Excellence Programme (to A.S.S.); Federal Agency of Scientific Organizations via the Institute of Cytology and Genetics (project #0324-2019-0040 to S.Z.S., Y.A.T. and Y.A.); the RCUK Innovation Fellowship from the National Productivity Investment Fund (MR/R026408/1 to L.K.); 'Biomedical Research Program' funds at Weill Cornell Medicine—Qatar, a program funded by the Qatar Foundation (to K.S. and G.T.).

Author Contributions

S.Z.S. and Y.A.T. contributed to the design of the study, carried out statistical analysis and produced the figures; S.Z.S., Y.A.T. and A.S.S. contributed to interpretation of the results; S.Z.S., Y.A.T., L.K., K.S. and Y.A. wrote the first version of the manuscript; S.Z.S., A.S.S., Y.A.T., L.K. and Y.A. wrote the revised second version of the manuscript; L.K., F.V., S.Z.S. and J.K. contributed to data harmonization and quality control; M.S., M.V., F.V., T.P., J.Stambuk, I.T.-A., J.K., J.Simunovic, M.P.-B. and G.L. contributed to plasma N-glycome measurements; M.M. and T.S. analyzed TwinsUK dataset and contributed to interpretation of the results; L.K., A.M., H.C., M.D. and S.M.F. analyzed SOCCS dataset and contributed to interpretation of the results; M.A., F.M.K.W. and C.D. designed PainOR study and contributed to interpretation of the results; K.S. and G.T. analyzed QMDiab dataset and contributed to interpretation of the results; E.L., J.D. and M.G. designed CEDAR study and contributed to interpretation of the results; Y.A. and G.L. conceived and oversaw the study, contributed to the design and interpretation of the results; all co-authors contributed to the final manuscript revision.

References

1. Varki, A. (1993) Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, **3**, 97–130.
2. Ohtsubo, K. and Marth, J.D. (2006) Glycosylation in cellular mechanisms of health and disease. *Cell*, **126**, 855–867.
3. Skropeta, D. (2009) The effect of individual N-glycans on enzyme activity. *Bioorg. Med. Chem.*, **17**, 2645–2653.
4. Takeuchi, H., Yu, H., Hao, H., Takeuchi, M., Ito, A., Li, H. and Haltiwanger, R.S. (2017) O-glycosylation modulates the stability of epidermal growth factor-like repeats and thereby regulates notch trafficking. *J. Biol. Chem.*, **292**, 15 964–15 973.
5. Lauc, G., Pezer, M., Rudan, I. and Campbell, H. (2015) Mechanisms of disease: the human N-glycome. *Biochim. Biophys. Acta*, **1860**, 1574–1582.
6. Poole, J., Day, C.J., von, M., Paton, J.C. and Jennings, M.P. (2018) Glycointeractions in bacterial pathogenesis. *Nat. Rev. Microbiol.*, **16**, 440–452.
7. Hofsteenge, J., Blommers, M., Hess, D., Furmanek, A. and Miroshnichenko, O. (1999) The four terminal components of the complement system are C-mannosylated on multiple tryptophan residues. *J. Biol. Chem.*, **274**, 32786–32794.
8. Khoury, G.A., Baliban, R.C. and Floudas, C.A. (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.*, **1**, 90.
9. Craveur, P., Rebehmed, J. and de Brevern, A.G. (2014) PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins. *Database (Oxford)*, **2014**, bau041.
10. Russell, A.C., Šimurina, M., Garcia, M.T., Novokmet, M., Wang, Y., Rudan, I., Campbell, H., Lauc, G., Thomas, M.G. and Wang, W. (2017) The N-glycosylation of immunoglobulin G as a novel biomarker of Parkinson's disease. *Glycobiology*, **27**, 501–510.
11. Trbojević-Akmačić, I., Vučković, F., Vilaj, M., Skelin, A., Karssen, L.C., Krištić, J., Jurić, J., Momčilović, A., Šimunović, J., Mangino, M. *et al.* (2018) Plasma N-glycome composition associates with chronic low back pain. *Biochim. Biophys. Acta. Gen. Subj.*, **1862**, 2124–2133.
12. Gudelj, I., Salo, P.P., Trbojević-Akmačić, I., Albers, M., Primorac, D., Perola, M. and Lauc, G. (2018) 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.
13. Trbojevic Akmacic, I., Ventham, N.T., Theodoratou, E., Vučković, F., Kennedy, N.A., Krištić, J., Nimmo, E.R., Kalla, R., Drummond, H., Štambuk, J. *et al.* (2015) Inflammatory bowel disease associates with proinflammatory potential of the immunoglobulin G glycome. *Inflamm. Bowel Dis.*, **21**, 1237–1247.
14. Lemmers, R.F.H., Vilaj, M., Urda, D., Agakov, F., Šimurina, M., Klaric, L., Rudan, I., Campbell, H., Hayward, C., Wilson, J.F. *et al.* (2017) IgG glycan patterns are associated with type 2 diabetes in independent European populations. *Biochim. Biophys. Acta Gen. Subj.*, **1861**, 2240–2249.
15. Vajaria, B.N. and Patel, P.S. (2017) Glycosylation: a hallmark of cancer? *Glycoconj. J.*, **34**, 147–156.
16. Munkley, J. and Elliott, D.J. (2016) Hallmarks of glycosylation in cancer. *Oncotarget*, **7**, 35478–35489.
17. Taniguchi, N. and Kizuka, Y. (2015) Glycans and cancer: role of N-glycans in cancer biomarker, progression and metastasis, and therapeutics. *Adv. Cancer Res.*, **126**, 11–51.
18. Rodríguez, E., Schetters, S.T.T. and van Kooyk, Y. (2018) The tumour glyco-code as a novel immune checkpoint for immunotherapy. *Nat. Rev. Immunol.*, **18**, 204–211.

19. Thanabalasingham, G., Huffman, J.E., Kattla, J.J., Novokmet, M., Rudan, I., Gloyn, A.L., Hayward, C., Adamczyk, B., Reynolds, R.M., Muzinic, A. *et al.* (2013) Mutations in HNF1A result in marked alterations of plasma glycan profile. *Diabetes*, **62**, 1329–1337.
20. Adamczyk, B., Tharmalingam, T. and Rudd, P.M. (2012) Glycans as cancer biomarkers. *Biochim. Biophys. Acta Gen. Subj.*, **1820**, 1347–1353.
21. Shinohara, Y., Furukawa, J. and Miura, Y. (2015) Glycome as Biomarkers. In Preedy, V.R. and Patel, V.B. (eds), *General Methods in Biomarker Research and their Applications*. Springer Netherlands, Dordrecht, pp. 111–140.
22. Peng, W., Zhao, J., Dong, X., Banazadeh, A., Huang, Y., Hussien, A. and Mechref, Y. (2018) Clinical application of quantitative glycomics. *Expert Rev. Proteomics*, **15**, 1007–1031.
23. Lauc, G., Vojta, A. and Zoldoš, V. (2014) Epigenetic regulation of glycosylation is the quantum mechanics of biology. *Biochim. Biophys. Acta Gen. Subj.*, **1840**, 65–70.
24. Moremen, K.W., Tiemeyer, M. and Nairn, A.V. (2012) Vertebrate protein glycosylation: diversity, synthesis and function. *Nat. Rev. Mol. Cell Biol.*, **13**, 448–462.
25. Knežević, A., Polašek, O., Gornik, O., Rudan, I., Campbell, H., Hayward, C., Wright, A., Kolčić, I., O'Donoghue, N., Bones, J. *et al.* (2009) Variability, heritability and environmental determinants of human plasma N-Glycome. *J. Proteome Res.*, **8**, 694–701.
26. Lauc, G., Essafi, A., Huffman, J.E., Hayward, C., Knežević, A., Kattla, J.J., Polašek, O., Gornik, O., Vitart, V., Abrahams, J.L. *et al.* (2010) Genomics meets Glycomics—the first GWAS study of human N-Glycome identifies HNF1α as a master regulator of plasma protein Fucosylation. *PLoS Genet.*, **6**, e1001256.
27. Huffman, J.E., Knežević, A., Vitart, V., Kattla, J., Adamczyk, B., Novokmet, M., Igl, W., Pučić, M., Zgaga, L., Johannson, Å. *et al.* (2011) Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults. *Hum. Mol. Genet.*, **20**, 5000–5011.
28. Huffman, J.E., Pučić-Baković, M., Klarić, L., Hennig, R., Selman, M.H.J., Vučković, F., Novokmet, M., Krištić, J., Borowiak, M., Muth, T. *et al.* (2014) Comparative performance of four methods for high-throughput glycosylation analysis of immunoglobulin G in genetic and epidemiological research. *Mol. Cell. Proteomics*, **13**, 1598–1610.
29. Knežević, A., Bones, J., Kračun, S.K., Gornik, O., Rudd, P.M. and Lauc, G. (2011) High throughput plasma N-glycome profiling using multiplexed labelling and UPLC with fluorescence detection. *Analyst*, **136**, 4670–4673.
30. Durbin, R.M., Altshuler, D.L., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De La Vega, F.M., Donnelly, P. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
31. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K. *et al.* (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279–1283.
32. Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.-J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T. *et al.* (2015) Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.*, **6**, 5890.
33. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P. and Cunningham, F. (2016) The Ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
34. Rogers, M.F., Shihab, H.A., Mort, M., Cooper, D.N., Gaunt, T.R. and Campbell, C. (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, **34**, 511–513.
35. Ferlaino, M., Rogers, M.F., Shihab, H.A., Mort, M., Cooper, D.N., Gaunt, T.R. and Campbell, C. (2017) An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. *BMC Bioinformatics*, **18**, 442.
36. Mahajan, A., Wessel, J., Willems, S.M., Zhao, W., Robertson, N.R., Chu, A.Y., Gan, W., Kitajima, H., Taliun, D., Rayner, N.W. *et al.* (2018) Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.*, **50**, 559–571.
37. de, C.A., Mooij, J.M., Heskes, T. and Posthuma, D. (2015) MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.*, **11**, e1004219.
38. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M. *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, **48**, 481–487.
39. Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E. *et al.* (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, **45**, 1238–1243.
40. GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI; NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
41. Momozawa, Y., Dmitrieva, J., Théâtre, E., Deffontaine, V., Rahmouni, S., Charloteaux, B., Crins, F., Docampo, E., Elansary, M., Gori, A.-S. *et al.* (2018) IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.*, **9**, 2427.
42. Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J. *et al.* (2016) PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics*, **32**, 3207–3209.
43. Lauc, G., Huffman, J.E., Pučić, M., Zgaga, L., Adamczyk, B., Mužinić, A., Novokmet, M., Polašek, O., Gornik, O., Krištić, J. *et al.* (2013) Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet.*, **9**, e1003225.
44. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J. *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.
45. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.
46. Shah, T., Zabaneh, D., Gaunt, T., Swerdlow, D.I., Shah, S., Talmud, P.J., Day, I.N., Whittaker, J., Holmes, M.V., Sofat, R. *et al.* (2013) Gene-centric analysis identifies variants

associated with Interleukin-6 levels and shared pathways with other inflammation markers. *Circ. Cardiovasc. Genet.*, **6**, 163–170.
47. Ridker, P.M., Pare, G., Parker, A., Zee, R.Y.L., Danik, J.S., Buring, J.E., Kwiatkowski, D., Cook, N.R., Miletich, J.P. and Chasman, D.I. (2008) Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's genome health study. *Am. J. Hum. Genet.*, **82**, 1185–1192.
48. Chambers, J.C., Zhang, W., Sehmi, J., Li, X., Wass, M.N., Van Der, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S.E. *et al.* (2011) Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat. Genet.*, **43**, 1131–1138.
49. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z. *et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, **46**, 1173–1186.
50. Perry, J.R.B., Day, F., Elks, C.E., Sulem, P., Thompson, D.J., Ferreira, T., He, C., Chasman, D.I., Esko, T., Thorleifsson, G. *et al.* (2014) Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, **514**, 92–97.
51. Clerc, F., Reiding, K.R., Jansen, B.C., Kammeijer, G.S.M., Bondt, A. and Wuhrer, M. (2016) Human plasma protein N-glycosylation. *Glycoconj. J.*, **33**, 309–343.
52. Brockhausen, I. and Schachter, H. Glycosyltransferases involved in N- and O-Glycan biosynthesis. In *Glycosciences*. Wiley-VCH GmbH, Weinheim, Germany, pp. 79–113.
53. Pottier, N., Cheok, M.H., Yang, W., Assem, M., Tracey, L., Obenauer, J.C., Panetta, J.C., Relling, M.V. and Evans, W.E. (2007) Expression of SMARCB1 modulates steroid sensitivity in human lymphoblastoid cells: identification of a promoter snp that alters PARP1 binding and SMARCB1 expression. *Hum. Mol. Genet.*, **16**, 2261–2271.
54. Oda, Y., Okada, T., Yoshida, H., Kaufman, R.J., Nagata, K. and Mori, K. (2006) Derlin-2 and Derlin-3 are regulated by the mammalian unfolded protein response and are required for ER-associated degradation. *J. Cell Biol.*, **172**, 383–393.
55. GTEx Consortium (2015) Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
56. Shen, X., Klarić, L., Sharapov, S., Mangino, M., Ning, Z., Wu, D., Trbojević-Akmačić, I., Pučić-Baković, M., Rudan, I., Polašek, O. *et al.* (2017) Multivariate discovery and replication of five novel loci associated with immunoglobulin G N-glycosylation. *Nat. Commun.*, **8**, 447.
57. Slack, J.M.W. (2014) Molecular biology of the cell. In *Principles of Tissue Engineering*. Academic Press, Boston, pp. 127–145.
58. Auer, P.L. and Lettre, G. (2015) Rare variant association studies: considerations, challenges and opportunities. *Genome Med.*, **7**, 16.
59. Pardo, L.M., MacKay, I., Oostra, B., van Duijin, C.M. and Aulchenko, Y.S. (2005) The effect of genetic drift in a young genetically isolated population. *Ann. Hum. Genet.*, **69**, 288–295.
60. Crow, J.F. (1954) Statistics and mathematics in biology. In *Breeding Structure of Populations. II. Effective Population Number*. Iowa State College Press, Ames, IA, pp. 543–556.
61. Kennedy, A.E., Ozbek, U. and Dorak, M.T. (2017) What has GWAS done for HLA and disease associations? *Int. J. Immunogenet.*, **44**, 195–211.
62. Bekesova, S., Kosti, O., Chandler, K.B., Wu, J., Madej, H.L., Brown, K.C., Simonyan, V. and Goldman, R. (2012) N-glycans in liver-secreted and immunoglogulin-derived protein fractions. *J. Proteomics*, **75**, 2216–2224.
63. Ma, B., Simala-Grant, J.L. and Taylor, D.E. (2006) Fucosylation in prokaryotes and eukaryotes. *Glycobiology*, **16**, 158R–184R.
64. Takamatsu, S., Antonopoulos, A., Ohtsubo, K., Ditto, D., Chiba, Y., Le, D.T., Morris, H.R., Haslam, S.M., Dell, A., Marth, J.D. *et al.* (2010) Physiological and glycomic characterization of N-acetylglucosaminyltransferase-IVa and -IVb double deficient mice. *Glycobiology*, **20**, 485–497.
65. Kurimoto, A., Kitazume, S., Kizuka, Y., Nakajima, K., Oka, R., Fujinawa, R., Korekane, H., Yamaguchi, Y., Wada, Y. and Taniguchi, N. (2014) The absence of Core Fucose up-regulates GnT-III and Wnt target genes. *J. Biol. Chem.*, **289**, 11704–11714.
66. Moayyeri, A., Hammond, C.J., Hart, D.J. and Spector, T.D. (2013) The UK adult twin registry (TwinsUK resource). *Twin Res. Hum. Genet.*, **16**, 144–149.
67. Mook-Kanamori, D.O., Selim, M.M.E.-D., Takiddin, A.H., Al-Homsi, H., Al-Mahmoud, K.A.S., Al-Obaidli, A., Zirie, M.A., Rowe, J., Yousri, N.A., Karoly, E.D. *et al.* (2014) 1,5-Anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. *J. Clin. Endocrinol. Metab.*, **99**, E479–E483.
68. Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., DeLisle, R.K. *et al.* (2017) Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.*, **8**, 14357.
69. Vuckovic, F., Theodoratou, E., Thaci, K., Timofeeva, M., Vojta, A., Stambuk, J., Pucic-Bakovic, M., Rudd, P.M., Erek, L., Servis, D. *et al.* (2016) IgG Glycome in colorectal cancer. *Clin. Cancer Res.*, **22**, 3078–3086.
70. Theodoratou, E., Thaçi, K., Agakov, F., Timofeeva, M.N., Štambuk, J., Pučić-Baković, M., Vučković, F., Orchard, P., Agakova, A., Din, F.V.N. *et al.* (2016) Glycosylation of plasma IgG in colorectal cancer prognosis. *Sci. Rep.*, **6**, 28098.
71. Allegri, M., De, M., Minella, C.E., Klersy, C., Wang, W., Sim, M., Gieger, C., Manz, J., Pemberton, I.K., MacDougall, J. *et al.* (2016) 'Omics' biomarkers associated with chronic low back pain: protocol of a retrospective longitudinal study. *BMJ Open*, **6**, e012070.
72. Dagostino, C., De, M., Gieger, C., Manz, J., Gudelj, I., Lauc, G., Divizia, L., Wang, W., Sim, M., Pemberton, I.K. *et al.* (2017) Validation of standard operating procedures in a multicenter retrospective study to identify -omics biomarkers for chronic low back pain. *PLoS One*, **12**, e0176372.
73. Tenesa, A., Farrington, S.M., Prendergast, J.G.D., Porteous, M.E., Walker, M., Haq, N., Barnetson, R.A., Theodoratou, E., Cetnarskyj, R., Cartwright, N. *et al.* (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.*, **40**, 631–637.
74. Trbojević Akmačić, I., Ugrina, I., Štambuk, J., Gudelj, I., Vučković, F., Lauc, G. and Pučić-Baković, M. (2015) High-throughput glycomics: optimization of sample preparation. *Biochem.*, **80**, 934–942.
75. Johnson, W.E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
76. Aulchenko, Y.S., Ripke, S., Isaacs, A. and van, C.M. (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

77. Karssen, L.C., van, C.M. and Aulchenko, Y.S. (2016) The GenABEL project for statistical genomics. *F1000Res.*, **5**, 914.
78. Zhou, X. and Stephens, M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods*, **11**, 407–409.
79. Willer, C.J., Li, Y. and Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.
80. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.
81. Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
82. Watanabe, K., Taskesen, E., van Bochoven, A. and Posthuma, D. (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.*, **8**, 1826.
83. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P. and Mesirov, J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
84. Sharapov, S., Tsepilov, Y., Klaric, L., Mangino, M., Thareja, G., Simurina, M., Dagostino, C., Dmitrieva, J., Vilaj, M., Vuckovic, F., *et al.* (2018) Genome-wide association summary statistics for human blood plasma glycome. [Data set]. Zenodo. 10.5281/ZENODO.1298406.