Replication of fifteen loci involved in human plasma protein N-glycosylation in 4,802 samples from four cohorts

Word count: 3949 words

Key words: genetic association study / glycosylation / locus / total plasma N-glycome / replication

Sodbo Zh. Sharapov¹, Alexandra S. Shadrina¹, Yakov A. Tsepilov²,³, Elizaveta E. Elgaeva¹, Evgeny S. Tiys¹, Sofya G. Feoktistova¹, Olga O. Zaytseva⁴, Frano Vuckovic⁵, Rafael Cuadrat⁵, Susanne Jäger⁵,⁶, Clemens Wittenbecher⁵,⁶,⁷, Lennart C. Karssen⁸, Maria Timofeeva⁹,¹⁰, Therese Tillin¹¹, Irena Trbojević-Akmačić⁴, Tamara Štambuk⁴, Najda Rudman¹², Jasminka Krištić⁴, Jelena Šimunović⁴, Ana Momčilović⁴, Marija Vilaj⁴, Julija Jurić⁴, Anita Slana⁴, Ivan Gudelj⁴, Thomas Klarić⁴, Livia Puljak¹³, Andrea Skelin⁴,¹⁴, Antonia Jeličić Kadić¹⁵, Jan van Zundert¹⁶,¹⁷, Nishi Chaturvedi¹¹, Harry Campbell¹⁸, Malcolm Dunlop⁹, Susan M. Farrington⁹, Margaret Doherty¹⁹,²⁰, Concetta Dagostino²¹, Christian Gieger²², Massimo Allegri²³, Frances Williams²⁴, Matthias Schulze⁵,⁶, Gordan Laue⁴,≜, Yuriii Aulchenko¹#*

¹ Laboratory of Glycogenomics, Institute of Cytology and Genetics, Novosibirsk, 630090, Russia
² Laboratory of Theoretical and Applied Functional Genomics, Novosibirsk State University, Novosibirsk, 630090, Russia
³ Laboratory of Recombination and Segregation Analysis, Institute of Cytology and Genetics, Novosibirsk, 630090, Russia
⁴ Genos Glycoscience Research Laboratory, Borongajska cesta 83h, 10000 Zagreb, Croatia
⁵ Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, 14558 Nuthetal, Germany
⁶ German Center for Diabetes Research (DZD), Neuherberg, Germany
⁷ Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA 02115, U.S.
⁸ PolyOmica, ’s-Hertogenbosch 5237 PA, The Netherlands
⁹ Colon Cancer Genetics Group, Cancer Research UK Edinburgh Centre, MRC Institute of Genetics & Molecular Medicine, The University of Edinburgh, Edinburgh EH4 2XU, UK
¹⁰ D-IAS, Danish Institute for Advanced Study, Department of Public Health, University of Southern Denmark, J.B. Winsløws Vej 9, DK-5000 Odense C
¹¹ MRC Unit for Lifelong Health & Ageing University College London London, UK
¹² Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb, Croatia
¹³ Catholic University of Croatia, Ilica 242, Zagreb, Croatia
¹⁴ St. Catherine Specialty Hospital, 10000 Zagreb & 49210 Zabok, Croatia
¹⁵ Clinical hospital u University hospital Center Split, Split, Croatia
#These authors jointly supervised this work and contributed equally.

*Correspondence:
Yurii S. Aulchenko
yurii@bionet.nsc.ru

**Running title:** Replication of loci associated with plasma protein N-glycome

**Supplementary data list:**
- **Table S1.** Description of plasma protein N-glycosylation traits.
- **Table S2.** The list of associations between 15 replicated loci and plasma protein N-glycome traits.
- **Table S3.** Description of the cohorts.
- **Table S4.** Table of correspondence of UHPLC measured glycan peaks between studies.
Abstract

Human protein glycosylation is a complex process, and its in vivo regulation is poorly understood. Changes in glycosylation patterns are associated with many human diseases and conditions. Understanding the biological determinants of protein glycome provides a basis for future diagnostic and therapeutic applications. Genome-wide association studies (GWAS) allow to study biology via a hypothesis-free search of loci and genetic variants associated with glycome composition. Sixteen loci were identified by three previous GWAS of human plasma proteome N-glycosylation. However, the possibility that some of these loci are false positives needs to be eliminated by replication studies, which have been limited so far.

Here, we use the largest set of samples so far (4,802 individuals) to replicate the previously identified loci. For all but one locus, the expected replication power exceeded 95%. Of the sixteen loci reported previously, fifteen were replicated in our study. For the remaining locus (near the KREMEN1 gene) the replication power was low, and hence replication results were inconclusive. The very high replication rate highlights the general robustness of the GWAS findings as well as the high standards adopted by the community that studies genetic regulation of protein glycosylation. The fifteen replicated loci present a good target for further functional studies. Among these, eight genes encode glycosyltransferases: MGAT5, B3GAT1, FUT8, FUT6, ST6GAL1, B4GALT1, ST3GAL4, and MGAT3. The remaining seven loci offer starting points for further functional follow-up investigation into molecules and mechanisms that regulate human protein N-glycosylation in vivo.
**Introduction**

Glycosylation is defined as covalent attachment of carbohydrates (glycans) to a substrate. Glycosylation is a common co- and posttranslational modification that influences physical properties of proteins and their biological function (Ohtsubo and Marth 2006; Skropeta 2009; Lauc et al. 2015; Takeuchi et al. 2017). Altered glycosylation is observed in many human diseases such as type 1 and type 2 diabetes (Lemmers et al. 2017), rheumatoid arthritis (Gudelj et al. 2018), hepatic and gastrointestinal pathology (Verhelst et al. 2020) (including inflammatory bowel disease (Trbojevic Akmacic et al. 2015)), Parkinson's disease (Russell et al. 2017), and cancer (Munkley and Elliott 2016). Changes in glycosylation profile are now considered potential biomarkers for different conditions, e.g., for HNF1A-MODY diabetes (Thanabalasingham et al. 2013; Juszczak et al. 2019) and chronic inflammatory gastrointestinal and liver disease (Verhelst et al. 2020).

The process of human protein glycosylation is complex, and mechanisms of its in vivo regulation are currently poorly understood. Genome-wide association studies (GWAS) allow a hypothesis-free search of genetic loci and variants associated with glycome composition and have the potential to enrich our knowledge of molecules controlling glycosylation pathways. The gold-standard GWAS design involves identification of the loci that are significantly associated at the genome-wide level with a trait of interest in a discovery set and subsequent replication of the identified associations in an independent set (Bush and Moore 2012). Replication improves the reliability of GWAS findings and helps to reduce the chance that the observed genotype-phenotype association is a chance finding or an analysis artifact (Kraft et al. 2009). Replication is, therefore, an essential step before complex and expensive follow-up studies.

To date, three GWAS of human plasma protein N-glycosylation have been published (Lauc et al. 2010; Huffman et al. 2011; Sharapov et al. 2019). In the first two studies performed by Lauc et al. (Lauc et al. 2010) and Huffman et al. (Huffman et al. 2011), N-glycans were measured using high-performance liquid chromatography (HPLC) technology. A pilot GWAS by Lauc et al. (Lauc et al. 2010) analyzed 13 glycan traits in 2,705 individuals and discovered 3 genome-wide significant loci, one of which included the gene HNF1A and others including genes encoding fucosyltransferases FUT6 and FUT8. Furthermore, Lauc et al. performed a knockdown of HNF1A in the human liver cancer cell line HepG2 and experimentally demonstrated that HNF1A is a master regulator of plasma protein fucosylation. In 2011, Huffman et al. extended the analysis to 46 glycan traits (33 directly measured and 13 derived from original traits, that average glycosylation features such as branching, galactosylation, sialylation and other features across different individual glycan structures) and increased the sample size to 3,533 individuals. These refinements led to the identification of three additional loci. However, none of these studies used independent samples to replicate their findings. Finally, the recent study by Sharapov et al. (Sharapov et al. 2019) analyzed a total of 113 glycan traits, of which 36 were directly measured by ultra-high-performance liquid chromatography (UHPLC – a method more advanced and accurate than HPLC (Ahn et al. 2010)) and 77 were derived traits. Examining data on 2,763 individuals, Sharapov et al. replicated 5 loci discovered in previous work (all except SLC9A9) and identified 10 new loci. Seven out of these ten new
loci were replicated in an independent cohort of 1,048 samples. Thus, 16 loci associated with plasma protein N-glycome composition have been identified to date, and 12 of them have been replicated in independent samples (Figure 1).

Here, we aimed to verify the findings of previous GWAS using an independent set of 4,802 individuals from 4 studies. Our study relied on the largest collection of samples with genotype and plasma protein N-glycome data available to date. In addition, we used an updated annotation of UHPLC peaks which refines the glycan species in each glycan peak and, consequently, refines the interpretation of N-glycan measurement (Zaytseva et al. 2020). In total, we analyzed 117 glycan traits, from which 36 were directly measured glycan traits and 81 were derived traits (Table S1). Our objectives were as follows: a) to replicate the association of the four loci which have not been replicated previously; b) to provide additional confirmation for the twelve previously replicated loci; c) to establish which updated glycan traits are associated with the analyzed loci.

**Results and discussion**

In our study, we used data from the EPIC-Potsdam (N = 2,192), PainOmics (N = 1,874), SOCCS (N = 459), and SABRE (N = 277) cohorts with a total sample size of N = 4,802 to analyze the association of 16 single nucleotide polymorphisms (SNPs) with 117 plasma protein N-glycosylation traits. With this sample size, our study had > 95 % statistical power to replicate a true positive association signal. Of note, in the study of Sharapov et al. (Sharapov et al. 2019), statistical power of replication analysis was 80 %, meaning that on average one out of five true associations would not be expected to replicate.

As we aimed not only to replicate the loci of interest but also to refine the spectrum of associated traits, we tested association of all 117 traits within each locus. We considered a locus to be replicated if the nominal P-value of association with at least one of the 117 traits passed the Bonferroni corrected conservative threshold of $P < 0.05 / (117 \times 16) = 2.67 \times 10^{-5}$, where 117 is a number of traits and 16 is a number of loci. Fifteen out of sixteen loci reported in previous human plasma protein N-glycome GWAS demonstrated an association with the studied N-glycosylation traits at $P < 2.67 \times 10^{-5}$ in our study and were therefore considered replicated (Table I; all significant locus-trait associations are listed in Table S2).

One locus near the KREMEN1 gene did not pass the statistical significance threshold. While this result is consistent with the study of Sharapov et al. (Sharapov et al. 2019) in which this locus was also not replicated, it cannot yet be taken as strong evidence that the KREMEN1 locus is not involved in N-glycosylation because the minor allele frequency of the investigated variant was low (MAF = 2%) and only the SOCCS study (N = 459) had genotypic data of sufficient quality available. Consequently, the replication power was limited to 9% and hence the results concerning KREMEN1 are inconclusive.

To summarize our results, first, our study provided additional independent replication of the 12 loci replicated previously in (Sharapov et al. 2019) (Figure 1). Second, we confirmed the association of three loci near the genes PRRC2A (HLA region), RUNX3/MAN1C1, and SLC9A9 which have not been replicated before. Finally, we updated the set of glycan traits influenced by variation in the 15 replicated loci (Table S2).
Independent replication of 12 previously replicated loci provides strong evidence that these associations are true positive findings. Overall, in this study, the 12 previously replicated loci showed association with glycan features which were the same or similar to those reported previously by Sharapov et al. (see Table S2). Six of these loci (near the genes MGAT5, ST6GAL1, B4GALT1, IKZF1, IGH/TMEM121, and SMARCB1/DERL3/CHCHD10) showed the most significant association with the same top traits as reported by Huffman et al. or Sharapov et al. (Table I). For the remaining six loci (near the genes B3GAT1, FUT8, FUT6, HNF1A, ST3GAL4, and MGAT3, see Table I and Table S2), we have refined the most significantly associated glycan trait. For the loci near B3GAT1, FUT8, FUT6, and HNF1A, there is a clear explanation for the difference between the top associated glycan features found in the Huffman et al. study (Huffman et al. 2011) and ours. The top glycan traits from the Huffman et al. study were measured by HPLC after the plasma N-glycan samples have been treated with an enzyme removing terminal sialic acid residues. In the present study, such treatment has not been performed. Thus, the top glycan traits from the (Huffman et al. 2011) study were not present in our analysis set. It should be noted that, in the present study, eight loci containing glycosyltransferase genes (B3GAT1, FUT8, FUT6, ST3GAL4, ST6GAL1, B4GALT1, MGAT5, and MGAT3) showed consistency between the enzymatic activity of the corresponding proteins and the spectra of associated glycan traits (see Table S2 for the details).

The current study is the first to replicate the association between plasma protein N-glycosylation and the loci near the genes PRRC2A (HLA region), RUNX3/MAN1C1, and SLC9A9. For two of these loci - PRRC2A and RUNX3/MAN1C1- we confirmed association with the same top traits, as reported by (Sharapov et al. 2019) - M9 (high mannose glycan) and FBG1n/G1n (percentage of core fucosylated bisected glycans among monogalactosylated neutral glycans), respectively.

Compared to the previous replication study of these loci (Sharapov et al. 2019), our sample size was more than four times larger for the loci near PRRC2A and RUNX3/MAN1C1, and almost twice as large for the locus near the SLC9A9 gene, thus, increasing the sample size was critical for confirmation of the effect of these loci. The locus near SLC9A9 (solute carrier family 9 member A9) does not contain any gene that is an obvious candidate for involvement in glycosylation. Along with Huffman et al., we consider the gene nearest to the top associated SNP – SLC9A9 – as the causal gene. This gene encodes a sodium/proton exchanger which is suggestively involved in the regulation of endosomal pH (Roxrud et al. 2009). Glycosylation is likely to be highly sensitive to changes in Golgi luminal pH (Kellokumpu 2019). Golgi acidity influences the formation of heteromeric complexes of enzymes involved in glycosylation pathways (Hassinen et al. 2011). Rivinoja et al. (Rivinoja et al. 2009) showed that elevated pH in the Golgi apparatus can impair protein terminal N-glycosylation (including sialylation) by inducing mislocalization of Golgi glycosyltransferases. This is consistent with the association of SLC9A9 locus with the trait tetrasialylation of N-glycans reported in the Huffman et al. study and sialylation-related traits (including the top associated trait FBS2/FS2 – the ratio of core-fucosylated disialylated structures with and without bisecting GlcNAc) revealed in our work (Table S2). We speculate that alterations in SLC9A9 function might affect sialylation of bisected
(addition of bisected GlcNAc) glycans via modulation of Golgi pH. Interestingly, strong defects in another solute carrier, a manganese transporter SLC38A9 (Mn$^{2+}$ cations are important cofactors for normal Golgi functioning (Breton et al. 2006)), lead to a congenital disorder of glycosylation (Park et al. 2015), while the common missense variant rs13107325 in this gene was found to be associated with weak changes in the plasma protein N-glycome (Mealer et al. 2019). In Sharapov et al. (Sharapov et al. 2019), rs13107325 was associated with the percentage of trisialylated structures, although this association did not reach genome-wide significance (nominal $P = 6 \times 10^{-5}$).

The genes RUNX3 and MANIC1 located near index rs186127900 encode proteins both of which are implicated in glycosylation or its regulation. RUNX3 is a member of the runt domain-containing family of transcription factors. Klarić et al. performed a functional network analysis and suggested that RUNX3, together with RUNX1 and SMARCB1, may regulate the expression of glycosyltransferase MGAT3 (Klarić et al. 2020). MANIC1 (mannosyl-oligosaccharide 1,2-alpha-mannosidase IC) is a Golgi mannosidase involved in the maturation of Asn-linked oligosaccharides. Current evidence does not allow us to prioritize one of these genes over another, highlighting the need for further functional studies. The remaining locus (near the PRRC2A gene) resides in the HLA region. Given the very special structure of this region, prioritization of causal genes in this locus using only genetics methods is difficult.

Our study has several limitations. First, the analyzed cohorts included only European-ancestry individuals. Therefore, the generalizability of our findings to populations with a different ethnic background is limited. Second, we investigated the association of SNPs with total plasma protein glycosylation, which implies that we cannot detect genetic effects on protein-specific glycosylation (although there is evidence that some of the UHPLC glycan structures are protein specific (Clerc et al. 2016)). Finally, our study had only 9% statistical power to replicate the association of the locus near the KREMEN1 gene.

In summary, our study provides strong evidence that 15 out of the 16 loci discovered in previous studies are robustly associated with human plasma protein N-glycome composition. For the locus near the KREMEN1 gene, the results remain inconclusive. The very high replication rate highlights the robustness and generalizability of GWAS findings in general, as well as high standards adopted by previous work (Lauc et al. 2010; Huffman et al. 2011; Sharapov et al. 2019) in particular. The 15 replicated loci present a good target for further functional studies. Among these, eight contain an “obvious” candidate gene (a glycosyltransferase involved in the N-glycan biosynthesis). Previous in vitro studies confirmed the role of two loci, HNF1A (Lauc et al. 2010) and IKZF1 (Klarić et al. 2020), in regulation of N-glycosylation (see Figure 1). The association of the remaining five loci on plasma protein N-glycosylation may serve as a starting point for further functional studies.

**Materials and methods**

For the replication of previously reported associations, we used data (combined association results) collected in four studies: EPIC-Potsdam ($N = 2,192$), PainOmics ($N = 1,874$), SOCCS ($N = 459$), and SABRE ($N = 277$) with a total sample size of $N = 4,802$ samples. Description of the analyzed cohorts, phenotyping (including plasma N-glycome measurement
and quality control), genotyping, and imputation is provided in Supplementary Methods and Table S3. All participants enrolled in the studies gave written informed consent, and all studies were approved by the corresponding Ethics Committees/Institutional Review Boards.

**Association analysis and meta-analysis**

We performed association analysis for tag SNPs, representing sixteen loci, previously reported to be associated with total plasma \(N\)-glycome traits (Lauc et al. 2010; Huffman et al. 2011; Sharapov et al. 2019). Tag SNPs for the loci near the genes \(FUT3/FUT5/FUT6, FUT8, HNF1A, B3GAT1, MGAT5,\) and \(SLC9A9\), associations found in the series of GWAS conducted on HPLC-measured total plasma \(N\)-glycome data, were selected by choosing the SNP with the strongest association as reported by Sharapov et al. (Sharapov et al. 2019). Tag SNPs for other loci, associations found in the first GWAS of UHPLC-measured total plasma \(N\)-glycome data, were selected by choosing the SNP with the strongest association as reported by Huffman et al. (Huffman et al. 2011). The rs59111563 SNP tagging \(ST6GAL1\) locus was not present in our replication set, and we, therefore, replaced it with rs17775791 (LD \(r^2 = 0.9898\), as calculated by LDLink, EUR subset of samples from 1000 Genomes Project phase 3 version 5, allele rs59111563 delT is positively correlated with allele rs17775791 T). The association analysis for tag SNPs was conducted assuming an additive model of genetic effects. We conducted an inverse-variance weighted fixed-effect meta-analysis of sixteen tag SNPs among four cohorts. The meta-analysis was performed using the GWAS-MAP platform (Gorev et al. 2018).

**Replication**

Since there is no direct trait-to-trait correspondence between glycan traits measured by HPLC and UHPLC technologies, we tested the association of tag SNPs with all 117 glycan traits. We considered a locus to be replicated if the tag SNP showed association with at least one of 117 glycan traits with a replication threshold. The summary of the association between 16 loci and glycome traits that passed the replication threshold is provided in Supplementary Table 2. The statistical power calculation is given in the Supplementary Methods.

**Author contribution**

YSA together with GL conceived and supervised this study. SZhSh coordinated this study. SZhSh, EEE, SGF, EST, YAT, ASSh, and FV performed centralized data analyses. ITA, TS, NR, JK, JS, AM, MV, JJ, ASl, IG, and TK analyzed the samples. RC, SJ, CW, MT, SZhSh, EEE, SGF, EST, and LCK contributed to the analysis of data of individual cohorts. MSch, FW, MA, CG, CD, MDoj, SFar, MDun, HC, and NC contributed the data and resources to analysis. SZhSh, ASSh, OZ, GL and YSA contributed to the interpretation the results. SZhSh, ASSh, and YSA wrote the initial version of this manuscript. All co-authors contributed to the discussion of the final text.

**Funding**

This work was supported by the Russian Science Foundation (grant number 19-15-00115 to YSA) that funded contribution of SZhSh, EEE, ASSh, SGF, EST, and YSA, who coordinated and supervised this study, performed centralized data harmonization, quality control, meta-
analysis and in-silico functional follow-up, interpretation of the results and writing of the initial text of the manuscript.

The work of YAT, who contributed in interpretation of the results, was supported by the Russian Ministry of Science and Education under the 5-100 Excellence Programme and by the Federal Agency of Scientific Organizations via the Institute of Cytology and Genetics (project 0324-2019-0040-C-01/AAAA-A17-117092070032-4). The work of LCK, who contributed to the analysis of PainOmics cohort, was funded by PolyOmica, The Netherlands.

The generation of the data used in this study was supported by the European Structural and Investment Funds IRI grant (#KK.01.2.1.01.0003) and Croatian National Centre of Research Excellence in Personalized Healthcare grant (#KK.01.1.1.01.0010) to GL by grants from Cancer Research UK (C348/A3758, C348/A8896, C348/A18927) to MD and HC; by European Community's Seventh Framework Programme funded PainOmics project (Grant agreement n. 602736 to MA and FW) that also funded contribution of MA and CD. SABRE was funded at baseline by the Medical Research Council, Diabetes UK, and British Heart Foundation and at follow-up by the Wellcome Trust (082464/Z/07/Z) and British Heart Foundation (SP/07/001/23603, PG/08/103, PG/12/29/29497 and CS/13/1/30327). N.C. receives support from the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

The recruitment phase of the EPIC-Potsdam Study was supported by the Federal Ministry of Science, Germany (01 EA 9401) and the European Union (SOC 95201408 05F02). The follow-up of the EPIC-Potsdam Study was supported by German Cancer Aid (70-2488-Ha I) and the European Community (SOC 98200769 05F02). This work was furthermore supported by a grant from the German Ministry of Education and Research (BMBF) and the State of Brandenburg (DZD grant 82DZD00302).

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) and by funding for the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre.

**Acknowledgements**

We thank the Human Study Centre (HSC) of the German Institute of Human Nutrition Potsdam-Rehbrücke, namely the trustee and the data hub for the processing, and the participants for the provision of the data, the biobank for the processing of the biological samples and the head of the HSC, Manuela Bergmann, for the contribution to the study design and leading the underlying processes of data generation. We thank Dr. Roel van Reij from the Department of Anesthesiology and Pain Medicine, Maastricht University Medical Centre, for the useful comments and suggestions. The SOCCS study acknowledges the excellent technical support from Stuart Reid. We are grateful to Donna Markie and all those who continue to contribute to recruitment, data collection, and data curation for the Study of Colorectal Cancer in Scotland studies. We acknowledge that these studies would not be possible without the patients and surgeons who take part. We acknowledge the expert support on sample preparation from the
Genetics Core of the Edinburgh Welcome Trust Clinical Research Facility. The SOCCS study acknowledges excellent support of the IGMM IT group and John Ireland.

**Conflict of interest statement**

YSA and LCK are owners of Maatschap PolyOmica and PolyKnomics BV, private organizations providing services, research and development in the field of computational and statistical, quantitative and computational (gen)omics. GL is the founder and owner of Genos Glycoscience Research Laboratory, a private research organization that specializes in high-throughput glycomic analysis and has several patents in this field. ITA, TS, FV, OOZ, JK, JS, AM, MV, JJ, ASI, IG, and TK are employees of Genos Ltd. The rest of the authors declare no potential conflict of interest.

**Abbreviations**

GWAS, genome-wide association study; HPLC, high-performance liquid chromatography; SNP, single nucleotide polymorphism; UHPLC, ultra-high-performance liquid chromatography; LD, linkage disequilibrium.
References


Figure legends

Figure 1. Schematic overview of the plasma protein N-glycome-associated loci revealed in previous GWAS and replicated in the present study. Chr – chromosome. *The role of the MAN1C1 (mannosidase alpha class 1C member 1) gene in glycosylation is known, but the association of the MAN1C1 locus was not replicated in the Sharapov et al. study (Sharapov et al. 2019). Green boxes highlight loci associated with N-glycosylation of IgG as reported by (Lauc et al. 2013; Shen et al. 2017; Klarić et al. 2020). The (enz.) label highlights genes that encode enzymes with known roles in glycan biosynthesis. The (in vitro) label highlights genes whose regulatory role in the glycosylation process was shown in vitro (Lauc et al. 2010; Klarić et al. 2020).
- Chr 2  \textit{MGAT5} (enz.)
- Chr 11  \textit{B3GAT1} (enz.)
- Chr 12  \textit{HNF1A} (in vitro)
- Chr 14  \textit{FUT8} (enz.)
- Chr 19  \textit{FUT3/FUT5/FUT6} (enz.)
- Chr 3  \textit{ST6GAL1} (enz.)
- Chr 9  \textit{B4GALT1} (enz.)
- Chr 11  \textit{ST3GAL4} (enz.)
- Chr 22  \textit{MGAT3} (enz.)
- Chr 7  \textit{IKZF1} (in vitro)
- Chr 14  \textit{IGH,TMEM121}
- Chr 22  \textit{SMARCBI, DERL3, CHCHD10}
- Chr 1  \textit{MAN1C1} (enz.), \textit{RUNX3}
- Chr 6  \textit{PRRC2A} (HLA region)
- Chr 3  \textit{SLC9A9}
- Chr 22  \textit{KREMER1}

- First found in Lauc et al. 2010 and Huffman et al. 2011
- First found in Sharapov et al. 2019

\textbf{Gene label}  – locus is associated with \textit{N}-glycosylation of IgG
\textbf{(in vitro)}  – genes implicated as regulators of glycosylation based on \textit{in vitro} experiments
\textbf{(enz.)}  – genes encoding glycosyltransferases

\textbf{replicated in Sharapov et al. 2019}
\textbf{replicated in the present study}
## Tables

### Table I. Replication of sixteen loci reported to be associated with human plasma protein $N$-glycome in previous (Lauc et al. 2010; Huffman et al. 2011; Sharapov et al. 2019).

<table>
<thead>
<tr>
<th>Index SNP</th>
<th>Chr:position*</th>
<th>EA/RAb</th>
<th>EAFc</th>
<th>Genesd</th>
<th>Top traitc</th>
<th>$\beta$ (SE)</th>
<th>$P$</th>
<th>Top traitd</th>
<th>$\beta$ (SE)</th>
<th>$P$</th>
<th>EAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1257220</td>
<td>2:135015347</td>
<td>A/G</td>
<td>0.26</td>
<td>MGAT5</td>
<td>Tetra-antennary glycans</td>
<td>0.19 (0.03)</td>
<td>$1.80 \times 10^{-10}$</td>
<td>G4total, A4totalº</td>
<td>0.22 (0.02)</td>
<td>$6.11 \times 10^{-20}$</td>
<td>0.26</td>
</tr>
<tr>
<td>rs4839604</td>
<td>3:142960273</td>
<td>C/T</td>
<td>0.77</td>
<td>SLC9A9</td>
<td>Terasialylated glycans</td>
<td>-0.22 (0.03)</td>
<td>$3.50 \times 10^{-13}$</td>
<td>FBS2/FS2</td>
<td>-0.20 (0.03)</td>
<td>$3.87 \times 10^{-11}$</td>
<td>0.80</td>
</tr>
<tr>
<td>rs7928758</td>
<td>11:134265967</td>
<td>T/G</td>
<td>0.88</td>
<td>B3GAT1</td>
<td>A4F2G4 (DG13)</td>
<td>0.23 (0.04)</td>
<td>$1.66 \times 10^{-8}$</td>
<td>A4G4S3</td>
<td>0.36 (0.03)</td>
<td>$6.43 \times 10^{-27}$</td>
<td>0.85</td>
</tr>
<tr>
<td>rs735396</td>
<td>12:121438844</td>
<td>T/C</td>
<td>0.61</td>
<td>HNF1A</td>
<td>A2F1G2 (DG7)</td>
<td>0.18 (0.03)</td>
<td>$7.81 \times 10^{-12}$</td>
<td>G3Fa/G3total</td>
<td>0.21 (0.02)</td>
<td>$4.91 \times 10^{-20}$</td>
<td>0.65</td>
</tr>
<tr>
<td>rs1162121</td>
<td>14:65822493</td>
<td>C/T</td>
<td>0.43</td>
<td>FUT8</td>
<td>A2 (DG1)</td>
<td>0.27 (0.03)</td>
<td>$1.69 \times 10^{-23}$</td>
<td>FG3/G3total</td>
<td>-0.31 (0.02)</td>
<td>$8.94 \times 10^{-45}$</td>
<td>0.42</td>
</tr>
<tr>
<td>rs3760776</td>
<td>19:5839746</td>
<td>G/A</td>
<td>0.87</td>
<td>FUT6</td>
<td>A3F1G3 (DG9)</td>
<td>0.44 (0.04)</td>
<td>$3.18 \times 10^{-20}$</td>
<td>G3Fa/G3total</td>
<td>0.48 (0.05)</td>
<td>$3.85 \times 10^{-23}$</td>
<td>0.91</td>
</tr>
</tbody>
</table>

### Sharapov et al. 2019

<table>
<thead>
<tr>
<th>Index SNP</th>
<th>Chr:position*</th>
<th>EA/RAb</th>
<th>EAFc</th>
<th>Genesd</th>
<th>Top traitc</th>
<th>$\beta$ (SE)</th>
<th>$P$</th>
<th>Current study</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs186127900</td>
<td>1:25318225</td>
<td>G/T</td>
<td>0.99</td>
<td>RUNX3, MAN1C1</td>
<td>FBG1n/G1n</td>
<td>-1.26 (0.12)</td>
<td>$4.04 \times 10^{-24}$</td>
<td>FA2G2S2</td>
</tr>
<tr>
<td>rs59111563¹</td>
<td>3:186722848</td>
<td>Del/Ins</td>
<td>0.74</td>
<td>ST6GAL1</td>
<td>FG1S1/(FG1+FG1S1)</td>
<td>0.34 (0.03)</td>
<td>$1.09 \times 10^{-26}$</td>
<td>FG1S1/(FG1+FG1S1)</td>
</tr>
<tr>
<td>rs3115663</td>
<td>6:31601843</td>
<td>T/C</td>
<td>0.80</td>
<td>PRRC2A</td>
<td>M9</td>
<td>0.26 (0.04)</td>
<td>$7.65 \times 10^{-11}$</td>
<td>M9</td>
</tr>
<tr>
<td>rs6421315</td>
<td>7:50355207</td>
<td>G/C</td>
<td>0.59</td>
<td>IKZF1</td>
<td>A2[6]BG1n</td>
<td>0.19 (0.03)</td>
<td>$7.57 \times 10^{-11}$</td>
<td>A2[6]BG1n</td>
</tr>
<tr>
<td>rs13297246</td>
<td>9:33128617</td>
<td>G/A</td>
<td>0.83</td>
<td>B4GALT1</td>
<td>FA2G2n</td>
<td>-0.26 (0.04)</td>
<td>$4.11 \times 10^{-12}$</td>
<td>FA2G2n</td>
</tr>
<tr>
<td>rs3967200</td>
<td>11:126232385</td>
<td>C/T</td>
<td>0.88</td>
<td>ST3GAL4</td>
<td>A2G2S3[6,3,6]</td>
<td>-0.49 (0.04)</td>
<td>$1.51 \times 10^{-27}$</td>
<td>G4S3/G4S4</td>
</tr>
<tr>
<td>rs35590487</td>
<td>14:105989599</td>
<td>C/T</td>
<td>0.77</td>
<td>IGH, TEMEM121</td>
<td>FA2[3]G1n</td>
<td>-0.24 (0.03)</td>
<td>$7.98 \times 10^{-12}$</td>
<td>FA2[3]G1n</td>
</tr>
<tr>
<td>rs9624334</td>
<td>22:24166256</td>
<td>G/C</td>
<td>0.85</td>
<td>DERL3, SMARCB1, CHCHD10</td>
<td>FA2[6]BG1n</td>
<td>0.28 (0.04)</td>
<td>$8.38 \times 10^{-12}$</td>
<td>FA2[6]BG1n</td>
</tr>
<tr>
<td>rs140053014</td>
<td>22:29550676</td>
<td>Ins/Del</td>
<td>0.98</td>
<td>KREMM1</td>
<td>G3S2/G3S3</td>
<td>-0.67 (0.11)</td>
<td>$4.05 \times 10^{-10}$</td>
<td>FA2[3]G1n</td>
</tr>
<tr>
<td>rs909674</td>
<td>22:39859169</td>
<td>C/A</td>
<td>0.27</td>
<td>MGAT3</td>
<td>FBS2/FS2</td>
<td>0.22 (0.03)</td>
<td>$7.72 \times 10^{-11}$</td>
<td>FBn</td>
</tr>
</tbody>
</table>

Locus which was not replicated in the present study is in bold italics. All statistically significant associations identified in the present study are provided in Table S2.

* Chromosome: position on chromosome according to GRCh37.p13 assembly
b Effect allele/reference allele
c Effect allele frequency
d Nearest genes or genes prioritized in previous GWAS for human plasma protein $N$-glycome
e Trait associated with index SNP at the highest level of statistical significance (as reported in the studies by Huffman et al. (Huffman et al. 2011) and Sharapov et al. (Sharapov et al. 2019))

f Trait associated with the index SNP at the highest level of statistical significance in our study. Description of traits is provided in Table S1
Two traits (G4total – tetragalactosylated glycans and A4total – tetra-antennary glycans) were associated with the index SNP with the same level of statistical significance and effect size.

SNP rs59111563 was not present in our replication set. We performed a replication analysis for SNP rs17775791, which is in high LD with rs59111563 ($r^2 = 0.99$ in European ancestry populations, allele rs59111563 Del is positively correlated with allele rs17775791 T according to LDlink, https://analysistools.nci.nih.gov/LDlink/). Therefore, top trait, $\beta$ (SE), $P$ and EAF indicated in the “our study” Table cells correspond to rs17775791.