

A genome-wide association study in 10,000 individuals links plasma N-glycome to liver disease and anti-inflammatory proteins

Sodbo Sharapov¹, Anna Timoshchuk¹, Olga Zaytseva², Denis Maslov¹, Anna Soplenskova¹, Elizaveta E. Elgaeva^{3,4}, Evgeny S. Tiys³, Massimo Mangino^{5,6}, Clemens Wittenbecher⁷, Lennart Karssen⁸, Maria Timofeeva^{9,10}, Arina Nostaeva^{3,8}, Frano Vuckovic², Irena Trbojević-Akmačić², Tamara Štambuk², Sofya Feoktistova³, Nadezhda A. Potapova¹¹, Viktoria Voroshilova^{3,12}, Frances Williams⁵, Dragan Primorac¹³, Jan Van Zundert^{14,15}, Michel Georges¹⁶, Karsten Suhre¹⁷, Massimo Allegri¹⁸, Nishi Chaturvedi¹⁹, Malcolm Dunlop⁹, Matthias B. Schulze^{20,21,22}, Tim Spector⁵, Yakov A. Tsepilov^{3,23}, Gordan Lauc², Yurii S. Aulchenko^{1,3}

1 MSU Institute for Artificial Intelligence, Lomonosov Moscow State University, Moscow, Russia;

2 Genos Glycoscience Research Laboratory, Borongajska cesta 83H, 10000 Zagreb, Croatia;

3 Institute of Cytology and Genetics, Novosibirsk, 630090, Russia;

4 Novosibirsk State University, Novosibirsk, 630090, Russia;

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 21 9RT, UK;

7 Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany;

8 PolyOmica, 's-Hertogenbosch 5237 PA, Netherlands;

9 Colon Cancer Genetics Group, Cancer Research UK Scotland Centre, Institute of Genetics & Cancer, Western General Hospital, 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. Winsløvs Vej 9, DK-5000 Odense C, Denmark;

11 Lopukhin Federal Research and Clinical Center of Physical-Chemical Medicine of Federal Medical Biological Agency, Moscow, Russia;

12 Vavilov Institute of General Genetics Russian Academy of Sciences, 119991 Moscow, Russia;

13 St. Catherine Specialty Hospital, Ulica Kneza Branimira 71E, Zagreb, Croatia;

14 Department of Anesthesiology and Multidisciplinary Paincentre, ZOL, Genk/Lanaken, Belgium;

15 Department of Anesthesiology and Pain Medicine, Maastricht University Medical Centre, P. Debyelaan 25, Maastricht, 6229 HX, The Netherlands;

16 Unit of Animal Genomics, WELBIO, GIGA-R and Faculty of Veterinary Medicine, University of Liège, (B34) 1 Avenue de l'Hôpital, Liège 4000, Belgium;

17 Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, P.O. Box 24144 Doha, Qatar;

18 Centre Lemanique d'antalgie et neuromodulation – EHC - Morges - CH;

19 MRC Unit for Lifelong Health & Ageing University College London, London, UK;

20 Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam- Rehbruecke, 14558 Nuthetal, Germany;

21 German Center for Diabetes Research (DZD), Neuherberg, 85764, Germany;

22 Institute of Nutrition Science, University of Potsdam, Potsdam, Germany;

23 Wellcome Sanger Institute, Cambridge, CB10 1RQ, the UK

Correspondence should be addressed to:

Dr. Yurii S. Aulchenko

yurii@bionet.nsc.ru

Current address:

GSK Medicines Research Centre

Gunnels Wood Road, Stevenage, SG1 2NY, UK

54 **More than a half of plasma proteins are N-glycosylated. Most of them are**
55 **synthesized, glycosylated, and secreted to the bloodstream by liver and lymphoid**
56 **tissues. While associations with N-glycosylation are implicated in the rising**
57 **number of liver, cardiometabolic, and immune diseases, little is known about the**
58 **genetic regulation of this process. Here, we performed the largest genome-wide**
59 **association study of N-glycosylation of the blood plasma proteome in 10,000**
60 **individuals. We doubled the number of genetic loci known to be associated with**
61 **blood N-glycosylation by identifying 16 novel loci and prioritizing 13 novel genes**
62 **contributing to N-glycosylation. Among these were the *GCKR*, *TRIB1*, *HP*,**
63 ***SERPINA1* and *CFH* genes. These genes are predominantly expressed in the liver**
64 **and show a previously unknown genetic link between plasma protein N-**
65 **glycosylation, metabolic and liver diseases, and inflammatory response. By**
66 **integrating glycomics, proteomics, transcriptomics, and genomics, we provide a**
67 **resource that facilitates deeper exploration of disease pathogenesis and supports**
68 **the discovery of glycan-based biomarkers.**

69

70

71 During maturation, more than half of human proteins are modified by the covalent
72 linking of complex carbohydrates – glycans¹. Glycoproteins comprise various secreted
73 and membrane enzymes, receptors, hormones, cytokines, immunoglobulins, as well as
74 structural and adhesion molecules². Glycans affect the physical and chemical properties
75 of proteins and their biological function³⁻⁶. Adequate glycosylation is required for the
76 normal physiological action of glycoproteins, while aberrant glycosylation is increasingly
77 implicated in human diseases⁷⁻⁹. Glycans are considered to be potential therapeutic
78 targets¹⁰⁻¹², essential part of therapeutics¹³⁻¹⁵, as well as biomarkers¹⁶⁻¹⁸, which makes
79 glycobiology a promising field for future clinical applications.

80 N-glycosylation is the most abundant type of glycosylation¹ and, unlike other
81 types, is specific to a consensus asparagine-containing sequence (Asn-X-Ser/Thr,
82 where X is any amino acid except Pro) in the protein's primary structure. Human N-
83 glycans are irregular branched polymers consisting of mannose, galactose, fucose,
84 sialic acid, and *N*-acetylglucosamine (GlcNAc) residuals, whose combinations introduce
85 a great diversity of protein glycoforms. Unlike proteins, whose primary structure is
86 encoded in the genomic DNA sequence, the occupancy of the N-glycosylation site, and
87 the abundance of specific N-glycan structures are not directly encoded in the human
88 genome. Protein glycosylation depends on the interplay of multiple enzymes catalyzing

89 glycan transfer, glycosidic linkage hydrolysis, and glycan biosynthesis. The abundance
90 of specific protein glycoforms can be influenced by various parameters, including the
91 activity of enzymes and availability of substrates, the accessibility of a glycosylation site,
92 protein synthesis, and degradation¹⁹. Overall, protein glycosylation is a complex
93 process controlled by genetic, epigenetic, and environmental factors^{20–22}.

94 While the biochemical network of human N-glycan biosynthesis is well
95 understood²³, little is known about *in vivo* regulation of this process²⁴, including tissue-
96 and protein-specific regulation. A major part of the plasma glycoproteins consists of
97 immunoglobulins, produced by antibody-producing B-cells, and secreted proteins
98 produced in the liver²⁵. Therefore, the N-glycosylation of blood plasma proteins serves
99 as an indicator of liver and B-cell function. Study of plasma protein N-glycosylation
100 potentially provides insights into the etiology and pathophysiology of liver and B-cell-
101 mediated diseases, as well as diseases where these tissues are important players, such
102 as cardiometabolic diseases and inflammatory conditions. Understanding the
103 mechanisms underlying blood plasma glycosylation and its regulation at the tissue-
104 specific level is crucial for unraveling the complex interplay between protein
105 modifications, cellular functions, and disease processes.

106 In this context, genetics offers an attractive approach to studying regulation of N-
107 glycosylation *in vivo* and sheds light on how these molecular phenotypes are linked to
108 human disease^{25,26}. Abundance of total plasma N-glycans can be quantified through
109 various analytical methods²⁷. As for other quantitative phenotypes, the genome-wide
110 association study (GWAS) and multivariate genetic association analysis^{28,29} may be
111 applied to N-glycans to identify genetic loci associated with abundance and, therefore,
112 contain genes involved in the regulation of N-glycosylation. Further integration of N-
113 glycome GWAS results with other layers of biological information (e.g., biological
114 pathways, protein-protein interactions, transcriptomics, proteomics, and others) allows
115 the discovery of novel candidate genes regulating this process and provides hypotheses
116 about biological mechanisms underlying the genetic associations^{30,31}. A joint analysis of
117 GWAS results of N-glycome and disease (e.g., pleiotropy analysis³² and analysis of
118 causal relationships using Mendelian randomization³³) can shed light on how protein
119 glycosylation is involved in pathogenesis of human disease and suggest possible
120 glycome-based biomarkers. Previous GWAS of total plasma N-glycome^{34–37} identified
121 15 genetic loci and suggested the role of 19 candidate genes. These studies were
122 supplemented with GWAS of N-glycome of immunoglobulin G (IgG)^{28,38–41} and
123 transferrin (TF)⁴² glycoproteins, identifying an additional 19 loci and prioritizing 26

124 candidate genes²⁵. The role of three candidate genes, encoding transcriptional factors
125 *HNF1A*, *IKZF1*, and *RUNX3*, in the regulation of N-glycosylation was experimentally
126 confirmed *in vitro*^{34,40}. A Mendelian randomization study of IgG N-glycome found that
127 the abundance of N-glycans with bisecting GlcNAc is a potential biomarker of systemic
128 lupus erythematosus⁴³. However, there remains a limited understanding of the role of
129 genes-regulators of N-glycosylation in health and disease.

130 The first aim of this study was to identify novel glycome quantitative trait loci
131 (glyQTLs), prioritize novel candidate genes, and reconstruct tissue-specific gene
132 networks that regulate plasma protein glycosylation. For this, we performed the largest
133 genome-wide association meta-analysis (GWAMA) of total plasma N-glycome using
134 data from seven studies (N = 10,764). For replicated glyQTLs, we prioritized candidate
135 genes using a broad spectrum of methods and explored how these genes are
136 connected in a functional tissue-specific network that regulates protein glycosylation.
137 The second aim of this study was to identify potential glycan biomarkers for disease.
138 We performed a phenome-wide association study (PheWAS) in conjunction with
139 colocalization analysis to investigate the pleiotropic effects of glyQTLs on complex
140 diseases. Next, we correlated genetically predicted glycan levels in 450,000 UK
141 Biobank samples with the disease's endpoints. Finally, we conducted a bidirectional
142 Mendelian randomization study to identify potential causal effects between glycans and
143 disease. This strategy not only resulted in the discovery of new candidate genes but
144 also suggested how some of these genes might regulate glycosylation enzymes and
145 how they are linked to the aberrant glycosylation observed in disease.

146 Results

147 Single- and multi-trait GWASs for 138 N-glycome traits

148 The levels of 36 N-glycan structures (Supplementary Table 3a, 3c) linked to various
149 plasma glycoproteins were measured by ultra-high performance liquid chromatography
150 in seven participating cohorts from six countries. Majority of 10,764 participants (94.5%)
151 were of European ancestry. From the 36 directly measured N-glycans, we computed 81
152 derived N-glycome traits such as the total level of fucosylation, galactosylation,
153 sialylation and others, reflecting pathways of N-glycan biosynthesis (Supplementary
154 Table 3a). We conducted GWAS for each of these 117 N-glycome traits in each of the
155 seven participating cohorts, assuming an additive model of the genetic effect. We then
156 performed a fixed-effect discovery meta-analysis of the subcohorts that included

157 participants of European descent (N = 7,540) (Supplementary Table 1b). After meta-
158 analysis, we took advantage of the correlation structure between 117 N-glycome traits
159 and performed GWAS of 21 multivariate N-glycome traits defined based on their
160 biochemical similarities (Supplementary Table 1b).

161 The size of the replication sample (N = 3,224, Supplementary Table 1b) was
162 defined as to achieve 80% statistical power for a replication of the true association
163 signal (Supplementary Note).

164 The genomic control inflation factor in the discovery GWAMA varied from 1.004
165 to 1.059. By contrast, an intercept of LD score regression⁴⁴ varied from 0.996 to 1.002
166 (Supplementary Table 4c), confirming minimal impact of genetic stratification on the
167 GWAS results. Hence, implementing Genomic Control correction in the analysis was
168 unnecessary.

169 Our analyses identified and replicated a total of 40 loci (**Fig. 1a**, Supplementary
170 Table 5a, Supplementary Table 5b) that were significantly associated with at least one
171 of 117 N-glycome traits and 21 multivariate N-glycome traits. The association of 25 loci
172 with total plasma N-glycome was shown and replicated for the first time (**Table 1**), while
173 the association of 15 loci confirms previous findings^{36,37}.

174 We performed an approximate conditional and joint analysis implemented in
175 GCTA-COJO⁴⁵ to identify conditionally independent association signals in the replicated
176 loci on discovery GWAMA. We found evidence of multiple SNPs contributing
177 independently to glycan level variation for nine loci (Supplementary Table 6a). Seven of
178 these loci span glycosyltransferase genes, coding for enzymes directly involved in the
179 biosynthesis of glycans. Two sentinel associations were observed in the loci containing
180 fucosyltransferases *FUT8* and *FUT6*, sialyltransferases *ST6GAL1* and *ST3GAL4*,
181 galactosyltransferase *B4GALT1*, glycuronyltransferase *B3GAT1*, and the
182 acetylglucosaminyltransferase *MGAT5*. Beyond glycosyltransferase loci, the locus
183 spanning the human leukocyte antigen (*HLA*) and the locus containing *HPR* gene
184 showed secondary associations.

185 SNP-based heritability and whole genome polygenic scores for 117 N-glycome traits

186 For 117 N-glycome traits we estimated SNP-based heritability using LD Score
187 regression⁴⁴. For 68 N-glycome traits SNP-based heritability was above zero at nominal
188 $P \leq 0.05$, varying from 10.2% to 33.4% ($19.8 \pm 10.3\%$) (Supplementary Table 4c), which
189 is on average 2.5x lower than the narrow-sense heritability of 37 N-glycome traits,
190 estimated in a twins-based study – $50.6 \pm 14.0\%$ ⁴⁶.

191 For each of the 117 N-glycome traits, we created polygenic score (PGS) models
192 based on the GWAMA of European-ancestry participants (N = 10,172) using the
193 SBayesR method⁴⁷. We tested the out-of-sample prediction accuracy of these models in
194 the CEDAR dataset (N=187 participants of European ancestry). For 79 N-glycome traits
195 in CEDAR samples, PGS models explained from 2.4% to 20.0% of the trait variance
196 (FDR < 5%), allowing for calculation of genetically predicted glycan levels in large scale
197 cohorts of European descent (e.g., UKBiobank). For the remaining 38 N-glycome traits
198 the explained variance did not deviate significantly from zero (FDR > 5%). The out-of-
199 sample prediction accuracy correlated significantly with the SNP-based heritability
200 ($R = 0.48$, $P = 4.05 \times 10^{-8}$). The implementation of SBayesR models is detailed in
201 Supplementary Table 7.

202 [Prioritization of causal genes for protein N-glycosylation](#)

203 Identification of genes, rather than genetic loci, can help to find novel protein
204 glycosylation regulators and suggest targets for intervention in glycome-related
205 diseases. To prioritize the most likely effector genes, we employed a consensus-based
206 prioritization approach selecting the gene with the highest unweighted sum of evidence
207 from eight different predictors - based on a literature search of genes encoding known
208 enzymes and regulators of N-glycan biosynthesis; genes causing congenital disorders
209 of glycosylation; colocalization of glyQTLs with eQTLs and blood plasma pQTLs;
210 annotation of putative causal variants affecting protein structure; enrichment of gene
211 sets and tissue-specific expression; and prioritization of the nearest gene (see
212 Methods). We prioritized the most likely effector gene for each locus by selecting the
213 gene with the highest unweighted sum of evidence across all eight predictors⁴⁸,
214 provided a gene was supported by at least two predictors.

215 We prioritized candidate genes in 31 of the 40 glyQTLs (Supplementary Table
216 6b). The prioritized genes may regulate the protein N-glycosylation through several
217 known general mechanisms: biosynthesis of N-glycans, abundance of N-glycoproteins
218 in the blood, regulation of transcription in lymphoid and gastrointestinal tissues, and ion
219 homeostasis in the endoplasmic reticulum and Golgi apparatus.

220 Among the 31 prioritized genes (**Fig. 2b**), we identified nine genes encoding
221 glycosyltransferases (*MGAT5*, *ST6GAL1*, *B4GALT1*, *ABO*, *ST3GAL4*, *B3GAT1*, *FUT8*,
222 *FUT6*, *MGAT3*); mutations in three are known to lead to congenital disorders of
223 glycosylation (*B4GALT1*, *FUT8*, *SLC39A8*) and four genes have strong experimental
224 support for being regulators of N-glycan biosynthesis genes (*HNF1A*, *IKZF1*, *RUNX3*,

225 *SLC39A8*)²⁵. The SMR/HEIDI approach indicated that total plasma N-glycosylation–
226 associated variants in two loci possibly had pleiotropic effects on plasma levels of two
227 blood proteins (HPT, CAFH) (Supplementary Table 6g) and transcription of 10 genes in
228 different tissues (Supplementary Table 6f). In 12 genes, associated variants were either
229 coding or were in strong LD with the variants coding for potentially deleterious amino
230 acid changes (annotated by Variant Effect Predictor, VEP⁴⁹), and in 5 genes -
231 pathogenic amino acid changes (predicted by FATHMM XF⁵⁰ and FATHMM InDel⁵¹).
232 The DEPICT gene prioritization tool³¹ provided evidence of prioritization for 19 genes in
233 18 loci at FDR < 0.2 (Supplementary Table 6h).

234 Because not all the glyQTLs were colocalized with a cis-eQTL, cis-pQTL or lay in
235 proximity to biologically relevant genes, we also utilized the nearest protein-coding
236 genes as an independent predictor. This approach was chosen due to the tendency of
237 the nearest protein-coding genes to enrich for molecular QTLs⁵².

238 In the following discussion, we focus on thirteen novel candidate genes that were
239 not identified before in GWASs of human protein N-glycosylation; for the latter, we refer
240 a reader to previous works and published reviews²⁵. We prioritized four genes
241 associated with lipid metabolism regulation - *GCKR*, *FADS2*, *TRIB1*, and *GRAMD1B*
242 which, to our knowledge, is the first time protein N-glycosylation has been linked to
243 genes involved in lipid metabolism and its regulation; four genes encoding N-
244 glycoproteins having anti-inflammatory function - *HP*, *HRP*, *SERPINA1* and *CFH*; the
245 gene *SCL39A8* encoding a zinc transporter; three genes encoding transcription factors -
246 *MAX*, *NFKB1*, *MYRF*; and a glycosyltransferase gene *ABO*, which determines an
247 individual's ABO blood type.

248 The Supplementary Note provides an in-depth account of the details of thirteen
249 newly prioritized genes. The other genes that have been previously prioritized
250 elsewhere are described in Timoshchuk et al.²⁵.

251 Tissue-specific regulation of plasma protein N-glycosylation

252 Lymphoid tissue, specifically plasma cells that produce antibodies, and liver,
253 specifically hepatocytes, contribute the majority of glycoproteins present in human
254 blood^{2,25} and are thus the primary drivers of N-glycosylation of plasma proteome.
255 However, the N-glycosylation machinery in these two cell types varies, leading to
256 distinct spectra of glycans attached to proteins produced in these two tissues^{2,46}. Many
257 glycosylation-associated genes prioritized in this study are expressed in plasma cells, or
258 hepatocytes, or both (**Fig. 2a**).

259 To gain a deeper understanding of the tissue-specific regulation of glyco-genes,
260 we constructed a gene network for N-glycosylation regulation. This network comprised
261 32 loci that were replicated in the univariate association analysis, and 117 N-glycome
262 traits as vertexes, with significant associations between them represented as edges
263 (**Fig. 1b**). The resulting network revealed two major subnetworks, wherein candidate
264 genes and glycan traits were clustered. The first subnetwork was primarily associated
265 with blood plasma N-glycans typically produced in the liver, and included 13 loci
266 (*ATF6B*, *B3GAT1*, *CHST2*, *FUT6*, *HNF1A*, *HP*, *LINC02714*, *MAX*, *MGAT5*, *MLXIPL*,
267 *SERPINA1*, *ST3GAL4*, *TRIB1*). The second subnetwork was related to blood plasma N-
268 glycans typically attached to immunoglobulins and consisted of 14 loci (*B4GALT1*,
269 *ELL2*, *HIVEP2*, *IKZF1*, *MEF2B*, *MGAT3*, *RUNX1*, *SLC38A10*, *SLC39A8*, *SMARCB1*,
270 *SMARCD3*, *TNFRSF13B*, *TMEM121*, *TXLNB*). According to classification of Clerc et
271 al.², genetic variation in five loci (containing *FUT8*, *GCKR*, *GRAMD1B*, *RUNX3*, and
272 *ST6GAL1*) had an impact on plasma N-glycans attached to both immunoglobulins and
273 liver-secreted proteins. Most of these five loci exhibited strong bias towards N-glycans
274 known to be preferentially expressed on proteins produced in one of the tissues, i.e.,
275 *GRAMD1B* was associated with 8 N-glycans, of which 7 were typical for liver proteins;
276 *GCKR* – with 9, of which only one was typical for immunoglobulins; *RUNX3* and
277 *ST6GAL1* were preferentially associated with N-glycans typically attached to
278 immunoglobulins (32/35 and 43/44 glycans, respectively). It should be noted that the
279 classification of Clerc et al.² was compiled based on a large body of literature data, and
280 we cannot exclude occasional misclassification.

281 To gain insight into the spectrum of glycans that were associated to the 8 loci that
282 were replicated in multivariate association analysis, we considered significant (at
283 $p < 0.01/36$) association of the partial regression coefficients in the multivariate analysis
284 of trait set “N-glycosylation” (36 traits) (Supplementary Table 5e). In this analysis,
285 *DIPK1A*, *FADS2*, and *CALB2* loci associated with N-glycans typical for liver-secreted
286 proteins; *LOC107985440* – with N-glycans typically observed on immunoglobulins IgG.
287 Results for *ST3GAL6* and *LOC157273* were inconclusive, although the former was
288 associated with the multivariate trait “high branching N-glycans”; such glycans are
289 typical for liver-secreted proteins.

290 Three loci showed clear effect on N-glycans found on both liver-secreted
291 glycoproteins and immunoglobulins: *FUT8* (significant effect on 9 N-glycans typically
292 attached to liver-secreted proteins and 3 typically attached to immunoglobulins); *CFH* (2
293 and 2, respectively) and *ABO* (also 2 and 2).

294 TF and IgG are two proteins secreted by hepatocytes and plasma cells,
295 respectively, and GWASs of their N-glycosylation shed light on the genetic control of
296 protein N-glycosylation in the corresponding tissues⁴². To gain further insights into the
297 mechanism of association and to support the tissue-specificity of the loci, we conducted
298 a colocalization analysis of total plasma, IgG and TF glyQTLs using the SMR- θ
299 method⁵³. The analysis was restricted to the loci that were previously implicated in TF⁴²
300 N-glycome or IgG⁴⁰ N-glycome GWASes, and reached genome-wide significance in
301 univariate association analysis in this study. Excluding *HLA*, this selection resulted in 21
302 loci, of which 15 were significant in previous IgG N-glycome GWAS only, four were only
303 significant in the TF N-glycome GWAS, and two (*FUT6* and *FUT8*) were significant in
304 both (Supplementary Table 5d)²⁵. For specific locus, we colocalized signals of genetic
305 association for traits that have reached genome-wide significance in that locus.

306 The results of colocalization analysis are presented in Supplementary Figures 3.
307 If regional genetic associations of a plasma N-glycome trait colocalized ($|\theta| > 0.7$) with
308 genetic associations of an IgG N-glycome trait, we considered this as evidence that the
309 locus is expressing its effect on plasma N-glycome through its effect on IgG N-
310 glycosylation, acting in antibody-producing cells. Similarly, colocalization with genetic
311 association signal for TF N-glycome was taken as an indication that the locus may
312 exhibit its action via effect of TF N-glycome, acting in liver. The analysis suggested that
313 the *ELL2*, *TXLNB*, *HIVEP2*, *IKZF1*, *SMARCD3*, *TMEM121*, *SLC38A10*, *MEF2B*,
314 *ATF6B*, *RUNX1*, *RUNX3*, *SMARCB1*, *MGAT3*, *ST6GAL1*, *B4GALT1* loci regulate N-
315 glycosylation of IgG while *MGAT5*, *ST3GAL4*, *B3GAT1*, *HNF1A* loci regulate N-
316 glycosylation of TF. The *FUT8* and *FUT6* act as regulators of both glycoproteins. An
317 interesting case of pleiotropy was observed in the *FUT8* locus. The colocalization signal
318 in *FUT8* split into two distinct clusters (Supplementary Fig. 3, page 120), one of which
319 was dominated by N-glycans predominantly presented on proteins produced in the liver,
320 while the other was almost exclusively presented by these on immunoglobulins. To
321 support the hypothesis of two distinct tissue-specific genetic mechanisms in the locus,
322 we combined traits from the two clusters into single traits using MANOVA approach²⁸
323 and performed a colocalization analysis between the two constructed linear
324 combinations using the SMR- θ method⁵³ and R Coloc package⁵⁴. We found strong
325 evidence *against* colocalization of N-glycans presented on liver-specific proteins and
326 these on immunoglobulins in this genomic region (PP.H3 = 100%, where PP.H3 is the
327 posterior probability that there are two distinct causal variants contributing to trait
328 variation, $\theta = 4 \times 10^{-6}$). This supports the hypothesis that genetic regulation of *FUT8* in

329 the liver and antibody-producing cells follows two distinct mechanisms, as previously
330 suggested by Landini and colleagues⁴².

331 Thus, the colocalization analysis confirms different tissue-specific mechanisms of
332 genetic regulation for each locus. With an exception of the *ATF6B* and *FUT6* loci, the
333 results from colocalization are largely consistent with the classification based on gene-
334 N-glycan association network.

335 Our findings demonstrate that the genetic regulation of protein N-glycosylation is
336 highly tissue-specific. Even glycosyltransferases such as *FUT8*, *FUT6*, *MGAT3*,
337 *ST6GAL1* and *B4GALT1*, being expressed in antibody-producing cells and hepatocytes
338 (Fig. 2a) and known to participate in the N-glycan biosynthesis in both tissues, display
339 pronounced tissue-specific genetic regulation that is not shared between different
340 tissues.

341 [PheWAS highlights loci associated with an extensive number of diseases and](#) 342 [quantitative traits](#)

343 In this study, we examined the pleiotropic effects of 40 replicated glyQTLs on over a
344 thousand diseases and quantitative traits endpoints (as listed in Supplementary Table
345 8a) using the SMR/HEIDI method. Our analysis revealed a total of 1,214 significant
346 associations, of which 781 demonstrated a non-rejection of the pleiotropy hypothesis by
347 the HEIDI test. The identified pleiotropic associations encompassed a wide range of
348 phenotypes, including type 2 and type 1 diabetes, blood glucose levels, coronary artery
349 disease, cholesterol levels, bipolar disorder, schizophrenia, gout, various oncological
350 diseases, metabolomic and anthropometric traits, lifestyle and diet-related traits, general
351 life history and overall health status, among others (as depicted in **Fig. 3a, 3b** and
352 Supplementary Table 8b). Additionally, *TRIB1* and *GCKR* showed colocalization with
353 metabolic dysfunction-associated steatotic liver disease (*TRIB1*: $P_{SMR} = 4.58 \times$
354 10^{-8} , $P_{HEIDI} = 0.03$ (possibly shared); *GCKR*: $P_{SMR} = 3.26 \times 10^{-8}$, $P_{HEIDI} = 0.73$ (likely
355 shared)).

356 Hierarchical clustering based on sets of colocalized traits allowed us to differentiate
357 four distinct groups of loci (**Fig. 3a**, right panel). The first cluster (**Fig. 3a**, topmost
358 cluster) was characterized by a high number of metabolic colocalizations and
359 encompassed eight of the ten most pleiotropic loci (**Fig. 3c**). It was also found to be
360 colocalized with a diverse range of other phenotypes, including disease and
361 anthropometric traits. Of particular note, this cluster comprised loci with prioritized lipid
362 metabolism genes, namely, *GCKR*, *TRIB1*, *FADS2* and previously known *HNF1A*.

363 Furthermore, it encompassed the locus containing the *MLXIPL* gene, a transcriptional
364 factor that induces liver glycolysis and lipogenesis, as well as *ABO*, which is known to
365 be associated with stroke⁵⁵, metabolic dysfunction-associated steatotic liver disease
366 and levels of lipids^{56,57}.

367 [Association between PGS for plasma N-glycosylation traits and ICD-10 diseases](#)

368 We analyzed associations between polygenic scores (PGS) for the 117 plasma
369 N-glycosylation traits and 167 diseases classified according to International
370 Classification of Diseases (ICD)-10 in individuals of European ancestry from the UK
371 Biobank cohort (N = 374,303) (Supplementary Note). The analysis revealed 14
372 diseases associated with PGS for at least one plasma N-glycome trait and PGS for 35
373 plasma N-glycome traits associated with at least one disease at the designated
374 significance threshold of $p < 1.07 \times 10^{-5}$ (**Fig. 4**, Supplementary Table 9). Full results
375 and overview of the analysis are provided in the Supplementary Note.

376 Notably, we observed positive associations between PGS for traits, related to the
377 increased levels of high-mannose glycans, especially to those of containing nine
378 mannose residuals (M9), with cardiovascular disease phenotypes such as essential
379 hypertension, ischemic heart disease, angina, hyperlipidaemias, as well as type 1
380 diabetes and asthma (**Fig. 4**).

381 We found negative associations for the PGS for S0 total (percentage of neutral
382 N-glycan structures, i.e. N-glycans without sialic acid, in total plasma N-glycans) with
383 primary hypertension, lipidaemias, obesity and non-insulin dependent diabetes (**Fig. 4**).
384 Moreover, PGS for the N-glycome traits describing the abundances of galactosylated
385 structures were negatively associated with obesity, lipoprotein metabolism disorders,
386 primary hypertension and type 2 diabetes (**Fig. 4**).

387 [Bidirectional Mendelian Randomization analysis of causal effects between plasma N- 388 glycosylation traits and ICD-10 diseases](#)

389 For the statistically significantly associated pairs of disease-plasma N-glycome
390 traits in UK Biobank we conducted bidirectional two-sample Mendelian randomization
391 (MR) analysis to investigate the direction of effects.

392 Using the disease phenotype as exposure, we conducted a two-sample MR
393 analysis and revealed statistically significant positive causal effect of disorders of
394 lipoprotein metabolism on M9 (N-glycan with nine mannose residuals) and on Mtotal
395 (the percentage of high-mannose structures in total plasma glycans) (**Table 2**,
396 Supplementary Table 10b, Supplementary Fig. 6a-d, 7a-d). The direction of the causal

397 effect corresponded to the signs of the beta of association between disorders of
398 lipoprotein metabolism and other lipidaemias (E78) and PGS for M9 and Mtotal.
399 Sensitivity analyses, including a two-sample MR after removal of pleiotropic IVs, as well
400 confirmed the observed causal effects of lipidaemias (E78) on M9 and Mtotal
401 (Supplementary Fig. 8a-d, 9a-d; Supplementary Tables 11a-c, 13a-c, Supplementary
402 Note).

403 Using plasma N-glycome trait as exposure we performed MR analysis in the
404 opposite direction, discovering a statistically significant positive causal effect of M6n,
405 percentage of M6 in total neutral plasma glycans on asthma and a positive effect of
406 M9n, percentage of M9 in total neutral plasma glycans, on disorders of lipoprotein
407 metabolism and other lipidaemias (**Table 2**, Supplementary Table 10a, Supplementary
408 Fig. 4a-d; 5a-c). In both cases, the direction of the effect was concordant with the
409 direction of association between corresponding pairs of disease and PGS. Sensitivity
410 analyses also confirmed the observed causal effect (Supplementary Tables 11a-c).
411 Since the number of available IVs for both M6n and M9n was not sufficient for the
412 analysis of pleiotropy among the IVs using MR-PRESSO⁵⁸, as an additional sensitivity
413 analysis for the effects of M6n on asthma and M9n on lipidaemia we performed
414 colocalization (SMR-HEIDI) analysis for the loci tagged by the genetic variants used as
415 IVs in these cases (Supplementary Table 12). SMR-HEIDI analysis provided evidence
416 for one shared causal variant (rs144126567, $p_{SMR} = 0.003$; $p_{HEIDI} = 0.30$) for M6n and
417 asthma but did not find any proof for existence of shared genetic variants influencing
418 M9n and lipidaemia (Supplementary Table 12). Therefore, we report an observed
419 positive effect of M6n on asthma, while the presence of a positive effect of M9n on
420 lipidaemia remains inconclusive. Full results of MR analysis are presented in
421 Supplementary Tables 10a-b.

422 Discussion

423 Here, we reported 40 quantitative trait loci (glyQTLs) discovered in the GWAS of
424 138 blood plasma N-glycome traits, resulting in a more than two-fold expansion of loci
425 affecting N-glycosylation of blood plasma proteins. The integration of these findings
426 with genetic information related to human diseases and other complex phenotypes
427 allowed us to show for the first time that genes involved in liver function are linked to the
428 human blood plasma protein N-glycosylation.

429 A subset of newly prioritized genes allows us to postulate a link between genetic
430 regulation of metabolic and liver diseases and blood plasma protein N-glycosylation.

431 Specifically, common genetic variation in the loci near *GCKR* and *TRIB1* is known to
432 predispose to metabolic dysfunction-associated steatotic liver disease (MASLD)⁵⁹.
433 Moreover, genetic association signal for MASLD in these loci are colocalized with the
434 corresponding glyQTLs. In the gene *SERPINA1*, rare Mendelian mutations lead to
435 alpha-1 antitrypsin (AAT) deficiency, with liver disease as part of the phenotype.
436 Common variation in this region associates with chronic elevation of alanine
437 aminotransferase (cALT) levels⁶⁰, a proxy phenotype for MASLD.

438 Although on phenotypic level, the changes of total and liver secreted protein-
439 specific N-glycosylation in liver disease are well-known⁶¹⁻⁶³, we demonstrate, for the
440 first time, that specific genes are associated with both N-glycosylation and liver disease,
441 offering a starting point for genetically-guided investigation of the functional
442 mechanisms of this phenotypic association.

443 Somewhat superficially, we may reason that liver disease is characterized by
444 hepatocyte injury and endoplasmic reticulum stress⁶⁴, which is strongly associated with
445 changes in glycosylation. A less direct mechanism could be inflammation, that is a
446 hallmark of liver disease, with proinflammatory cytokines shown to alter the substrate
447 synthesis pathways as well as the expression of glycosyltransferases required for the
448 biosynthesis of N-glycans⁶⁵. Thus, changes in N-glycosylation observed in blood
449 plasma may be at least partly explained by altered N-glycosylation of hepatocyte-
450 secreted proteins. Consistently with this hypothesis, we demonstrate that common
451 genetic variation in the three loci known to be associated with liver disease is
452 associated with variation in abundance of N-glycans, typically attached to the liver-
453 secreted proteins.

454 Other notable, partly overlapping, subset of liver-expressed genes newly
455 implicated in plasma protein N-glycosylation encodes anti-inflammatory proteins --
456 haptoglobin (*HP*, *HP*), complement factor H (*CFAH*, *CFH*), and alpha-1-antitrypsin
457 (*AAT*, *SERPINA1*). The glyQTLs located at *HP* and *CFH* are colocalized with the
458 corresponding pQTLs. At least two mechanisms may be suggested to explain such
459 colocalization. Genetic variants in these loci may affect the composition of blood N-
460 glycosylation through changes in the abundance of glycans preferentially bond to *HP*
461 and *CFAH* by regulating the level of these glycoproteins in the blood. Alternatively, the
462 genetic variation may change glycosylation of these proteins, which, in turn, could
463 change the affinity of the binding of the specific probes used by the SomaLogic assays.
464 While we did not observe colocalization between *SERPINA1* glyQTL and the *AAT*
465 pQTL, this may be a false negative, potentially explained by the low frequency of the

466 *SERPINA1* lead variant (rs28929474). Nonetheless, the rs28929474 is associated with
467 the level of glycoprotein acetyl, a mixture of N-glycoproteins, predominantly
468 alpha-1-acid glycoprotein, haptoglobin, and alpha-1-antitrypsin⁶⁶.

469 While variation at *HP* and *SERPINA1* loci is associated with changes in N-
470 glycans typically attached to liver proteins, *CFH* locus appears to also affect N-glycans
471 typically attached to immunoglobulins. While *CFH* does express at some level in plasma
472 cells (**Fig. 2a**), we speculate that perhaps a more likely mechanism is that genetic
473 variation in *CFH* affects its expression liver, and changes in N-glycans attached to
474 immunoglobulins occur through a systemic mechanism, i.e., regulation of inflammation.
475 Consistent with this hypothesis is the known association of the *CFH* locus with IgA
476 nephropathy⁶⁷, as well as indications that in mice, that CFH modulates splenic B cell
477 development and limits autoantibody production⁶⁸.

478 Our results suggest that genetic regulation of plasma protein N-glycosylation
479 predominantly occurs in lymphoid and liver tissue and exhibits strong tissue specificity.
480 Integration of evidence from transcriptomics and N-glycomics suggests that molecular
481 expression of genetic variation in the majority of glyQTLs is restricted to one tissue; the
482 effects of this variation on N-glycans in blood plasma occur either through changes of
483 N-glycosylation of the proteins secreted by the tissue, or through systemic mechanisms.
484 Of note, all of N-glycosyltransferase genes that express on RNA, protein, and glycan
485 levels in both tissues, are genetically regulated in only one of them (*B4GALT1*,
486 *ST6GAL1*, *MGAT3*), or exhibit different glyQTLs in different tissues (*FUT8*, *FUT6*). Even
487 thus, while a number of glycosyltransferases are expressed both in liver and lymphoid
488 tissues, we provide evidence that the genetic variation regulating N-glycosylation in
489 each of the two tissues is unique and does not overlap at the available resolution of the
490 analysis. To the best of our knowledge, this is the first study to analyze and reveal the
491 strong tissue-specificity of the genetic regulation of population variability of human
492 protein N-glycosylation.

493 Further studies of the genetic regulation of N-glycosylation of individual proteins
494 rather than bulk N-glycome will lead to the discovery of novel glyQTLs, which we cannot
495 observe now due to a lack of power or noise in bulk N-glycome. Quantification of the N-
496 glycome of purified proteins like IgG⁴⁰, TF⁴², IgA^{69,70}, and apolipoprotein CIII⁷¹, and
497 other proteins, will be highly relevant to understanding the etiology of such disease, as
498 rheumatoid arthritis, hepatocellular carcinoma, IgA nephropathy, endocarditis⁷.
499 Alternatively, development and application of computational deconvolution approaches
500 may be similar to those applied for bulk RNA-Seq⁷².

501 In conclusion, our study offers insight into the genetic factors influencing blood
502 plasma N-glycome, and, for the first time, establishes a genetic link between N-
503 glycosylation, liver disease, and anti-inflammatory proteins. The identification of novel
504 genes associated with metabolic and liver disease and N-glycosylation contributes to
505 deeper understanding of shared biological mechanisms and will facilitate future
506 biomarker discovery and interpretation.

507 **Methods**

508 We conducted a multicenter study using data from seven studies – TwinsUK (N =
509 3,918), EPIC-Potsdam (N = 2,192), PainOmics (N = 1,873), SOCCS (N = 1,742),
510 SABRE (N = 544), QMDiab (N = 325), and CEDAR (N = 170) with a total sample size N
511 = 10,764. Local research ethics committees approved all studies, and all participants
512 gave written informed consent. The detailed description of the cohorts is shown in
513 Supplementary Table 1 and Supplementary Notes.

514 **Glycome measurement and phenotype processing**

515 Plasma N-glycome quantification of samples from all but SOCCS studies was
516 performed at Genos Ltd using the protocol published previously⁷³. Briefly, plasma N-
517 glycans were enzymatically released from proteins by PNGase F, fluorescently labeled
518 with 2-aminobenzamide and cleaned-up. Fluorescently labeled and purified N-glycans
519 were separated by HILIC on a Waters BEH Glycan chromatography column. The
520 fluorescence detector was set with excitation and emission wavelengths of 250 nm and
521 428 nm, respectively. Plasma N-glycome quantification for SOCCS samples was done
522 at NIBRT by applying the same protocol with the only difference in the excitation
523 wavelength (330 nm instead of 250 nm). Glycan peaks (GPs) – quantitative
524 measurements of glycan levels – were defined by manual integration of intensity peaks
525 in the chromatograms or were defined by automatic integration. The number of defined
526 GPs varied among studies from 36 to 42, therefore to conduct multi-center association
527 analysis followed by meta-analysis, we harmonized the set of GPs by applying a
528 recently published protocol³⁴ to a harmonized set of 36 GPs. To reduce experimental
529 variation in glycan measurements, before genetic studies, raw glycan data were
530 probabilistic median quotient normalized^{74,75} and batch corrected centrally by the
531 phenotype provider (Genos Ltd). More detailed information on glycan preprocessing can
532 be found in the Supplementary Note. From the 36 directly measured glycan traits, 81
533 derived traits were calculated (Supplementary Table 3a). These derived traits average

534 glycosylation features such as branching, galactosylation, and sialylation, etc, across
535 different individual glycan structures and, consequently, they may be more closely
536 related to individual enzymatic activity and underlying genetic polymorphism.

537 [Discovery and replication genetic association analysis](#)

538 [Single trait association analysis](#)

539 Discovery genome-wide association studies were performed in (sub) cohorts of
540 European descent: TwinsUK (N = 2,739), EPIC-Potsdam (N = 2,192), PainOmics (N =
541 1,873), SOCCS (controls, N = 459) and SABRE (N = 277) with a combined sample size
542 of 7,540 (Supplementary Table 1b). Prior to GWAS, the total plasma N-glycome traits
543 were adjusted for sex and age, and the residuals were quantile transformed to normal
544 distribution. The genetic association analysis in each cohort was conducted using a
545 similar protocol. We assumed an additive model of genetic effects. GWAS were based
546 on the genotypes imputed from Haplotype Reference Consortium Results⁷⁶ or 1000
547 Genomes project⁷⁷. Results of GWAS in each discovery cohort passed a strict quality
548 control procedure followed by fixed-effects inverse-variance weighted meta-analysis.
549 After quality control, 8.8 M SNPs were used for the downstream analysis.

550 To define genome-wide significant glyQTLs, we used conventional genome-wide
551 significance threshold, Bonferroni corrected by 28 independent glycan traits ($P \leq 1.79 \times$
552 10^{-9}) as suggested before⁷⁸. We considered SNPs located in the same locus if they
553 were located within 250 Kb from the leading SNP (the SNP with lowest P). Only the
554 SNPs and the traits with lowest P are reported (leading SNP-trait pairs) in the **Table 1**.
555 The detailed procedure of locus definition is described in Supplementary Note.

556 Replication GWAS were performed using (sub) cohorts of samples with
557 European descent: SOCCS (colorectal cancer cases, N = 1,283), TwinsUK (N = 1,179),
558 CEDAR (N = 170); South Asian descent: SABRE (N = 267) and Arabian, Indian, Filipino
559 descent: QMDiab (N = 325) (Supplementary Table 1b). Results of GWAS in each
560 discovery cohort passed a strict quality control procedure followed by fixed-effects
561 inverse-variance weighted meta-analysis. For replication of novel glyQTL, found at the
562 discovery step, we used the leading SNP-trait pair that showed the most significant
563 association. The replication threshold was set as $P < 0.05/28 = 0.00178$, where 28 is
564 the number of replicated loci. Moreover, we checked whether the sign of estimated
565 effect was concordant between discovery and replication studies.

566 Identification of secondary associations in glyQTLs

567 To identify secondary association signals at glyQTL in univariate analysis and capture
568 the overall contribution to phenotypic variation, we performed conditional analysis using
569 GCTA-COJO software, version 1.93.2beta⁴⁵. This method uses summary-level statistics
570 from a discovery meta-analysis and LD corrections between SNPs estimated from a
571 reference sample for implementing a stepwise selection procedure including a series of
572 conditional and joint regression analyses in which the SNP with the strongest
573 association in the region is added to the regression model until no additional SNPs
574 reach genome-wide significance. We used 1,429 unrelated individuals with European
575 descent from SABRE cohort as reference samples for LD calculation. We used
576 $P \leq 1.79 \times 10^{-9}$ as a genome-wide significance level and a default window setting to
577 identify lead associations (Supplementary Table 6a).

578 Multi-trait association analysis

579 To gain additional power of glyQTL detection, we performed a multivariate
580 GWAS of total plasma N-glycome. It has been previously demonstrated that multivariate
581 genetic association analysis of N-glycome, that is, a joint analysis of multiple N-glycome
582 traits, has higher power for loci detection than a regression model under which glycome
583 traits are analyzed independently of each other^{28,41}.

584 For discovery and replication analyses, we used discovery and replication
585 GWAMA summary statistics, obtained in single-trait analysis. The discovery multivariate
586 analysis was performed using the MANOVA-based method, adopted for analysis of a
587 group of single-trait GWAS summary statistics (details are in Supplementary Note
588 Discovery multivariate analysis)²⁸. Discovery analysis was performed using the
589 MultiABEL R package. Due to method requirements, we filtered out SNPs with sample
590 size lower than 6,790 (which is 90% of 7,540 samples). The statistical significance
591 threshold for multivariate analysis was set at $P < 5 \times 10^{-8} / 21$, where 21 is the number
592 of multivariate traits described in Supplementary Table 3b and Supplementary Note.
593 GlyQTLs with significant association were defined in the same way as for single-trait
594 discovery.

595 For replication of multi-trait associated glyQTLs we used a complex four-step
596 replication strategy as proposed by Ning et al.²⁹, which consists of the following steps:
597 MANOVA, Phenotype Score, Pearson correlation method and Kendall correlation
598 method. In the first step (MANOVA) we straightforwardly checked whether the locus is
599 significantly associated with the multivariate trait in the replication cohort using the same

600 test as in the discovery stage. The replication threshold was set as $P < \frac{0.05}{(7+16)} = 0.0021$,
601 where 7 is the number of previously identified but not replicated loci and 16 is the
602 number of novel loci. Then we checked whether the effect direction is consistent
603 between the two cohorts, using the phenotype score approach²⁹. Next, we evaluated
604 the concordance of multivariate effect between two samples using Pearson and
605 Kendall's correlation coefficients. We considered an association of the locus replicated if
606 it had successfully passed MANOVA and phenotype score steps of replication. The
607 multivariate effect of the locus replicated if it additionally had passed both Pearson's
608 and Kendall's correlation steps of replication (Supplementary Note, Supplementary
609 Table 5b, and Supplementary Fig. 2). Phenotype score-based replication was
610 performed as in Shadrina et al⁴¹. For each lead pair of SNP and trait group phenotype,
611 we extracted coefficients of the linear combination of genotype onto multiple atomic
612 phenotypes, estimated for discovery cohort. We used them to construct the
613 corresponding trait group phenotypes for further testing of an association between the
614 lead SNPs and the derived linear combinations (see Supplementary Note for details). A
615 locus was replicated if the association of the SNP with the constructed linear
616 combination had the same direction of effect as in the discovery cohort and passed the
617 threshold of $P < 0.0021$.

618 To evaluate the similarity between estimates of multivariate genetic effects from
619 discovery and replication cohorts across multiple traits, we used an MC-based approach
620 implemented in MultiABEL package (MV.cor.test() function)²⁸. For both Pearson's and
621 Kendall's correlation coefficients, we considered a multivariate effect for a specific SNP
622 replicated if the 95% confidence intervals didn't include zero.

623 SNP-based heritability and polygenic scores

624 SNP-based heritability was estimated using the LD Score regression software⁴⁴
625 embedded in the GWAS-MAP platform⁷⁹. We used pre-computed LD scores that were
626 calculated from the European-ancestry samples in the 1000 Genomes Project. Only the
627 1,176,189 HapMap3 SNPs were included with a $MAF \geq 0.05$. For the purpose of
628 heritability estimation and further post-GWAS analyses, we generated GWAMA
629 summary statistics for the samples of European descent with $N = 10,172$. We used
630 GWAMA summary statistics for the analysis in order to use the largest data set with
631 homogeneous ancestry.

632 SBayesR method reweights the effect of each variant according to the marginal
633 estimate of its effect size, statistical strength of association, the degree of correlation

634 between the variant and other variants nearby, and tuning parameters. This method
635 requires a compatible LD matrix file computed using individual-level data from a
636 reference population. For these analyses, we used publicly available shrunk sparse
637 GCTB LD matrix including 1.1 million HapMap3 variants and computed from a random
638 set of 50,000 individuals of European ancestry from the UKB data set^{47,80}. SBayesR
639 (gctb_2.03) was run for each chromosome separately, and with the default parameters
640 except for the number of iterations (N = 5,000) and options for the stability of the
641 algorithm (Supplementary Table 7). The prediction accuracy was defined as the
642 proportion of the variance of a phenotype that is explained by PGS values (R²). To
643 calculate PGS based on the PGS model, we used PLINK2 software⁸¹, where PGS
644 values were calculated as a weighted sum of allele counts. Out-of-sample prediction
645 accuracy was evaluated using samples from the CEDAR cohort that were not used for
646 discovery or replication.

647 [Prioritization of candidate genes in found loci](#)

648 For the purpose of post-GWAS analyses, we generated GWAMA summary statistics for
649 the samples of European descent (N = 10,172). GWAMA summary statistics passed the
650 same QC procedure as discovery and replication GWAMA. We applied an ensemble of
651 methods to prioritize plausible candidate genes in the loci with found and replicated
652 glyQTL (32 in univariate and 8 in multivariate analysis). We applied eight approaches to
653 prioritize the most likely effector genes: (1) prioritization of the nearest gene; (2)
654 prioritization of genes with known role in biosynthesis of N-glycans; (3) genes of
655 congenital disorders of glycosylation; (4) genes with direct experimental support for
656 regulation of protein N-glycosylation; (5) prioritization of genes containing variants in
657 strong LD ($r^2 \geq 0.8$) with the lead variant, which are protein truncating variants
658 (annotated by Variant Effect Predictor, VEP⁴⁹) or predicted to be damaging by FATHMM
659 XF⁵⁰, FATHMM InDel⁵¹; (6) prioritization of genes whose eQTL and/or (7) pQTL are
660 colocalized with glyQTL; (8) prioritization of genes based on the gene set and tissue/cell
661 type enrichment, calculated by Data-driven Expression Prioritized Integration for
662 Complex Traits (DEPICT) framework³¹. We prioritized the most likely 'causal gene' for
663 each association using a consensus-based approach, selecting the gene with the
664 highest, unweighted sum of evidence across all eight predictors. In the case of equality
665 of the scores for two genes, we prioritized both genes.

666 Functional annotation of genetic variants

667 We inferred the possible molecular consequences of genetic variants in glyQTLs. We
668 focused on variants in LD with lead (for univariate and multivariate signals) and sentinel
669 variants (for univariate signals) picked by COJO. We created a “long list” of putative
670 causal variants using PLINK version 1.9 (--show-tags option), applied to whole genome
671 re-sequenced data for 503 European ancestry individuals (1000 Genomes phase 3
672 version 5 data). The size of the window to find the LD in both cases was equal to
673 500 kb. The default value of $r^2 > 0.8$ was taken as a threshold to include SNPs into the
674 credible sets. Ensembl Variant Effect Predictor (VEP) (Supplementary Table 6e) and by
675 FATHMM-XF (Supplementary Table 6c), FATHMM-INDEL (Supplementary Table 6d) to
676 reveal pathogenic point mutations.

677 Genes of N-glycan biosynthesis and Congenital Disorders of Glycosylation

678 We searched for the genes encoding glycosyltransferases – enzymes, with a known
679 role in N-glycan biosynthesis⁸², located in the ± 250 Kb-vicinity of the lead SNPs in
680 glyQTLs. Additionally, we prioritized genes with known mutations, that cause Congenital
681 Disorder of Glycosylation according to MedGen database
682 (<https://www.ncbi.nlm.nih.gov/medgen/76469>) that are located in the vicinity of ± 250 kb
683 from the lead SNPs.

684 Colocalization with eQTL and pQTL

685 To find potential pleiotropic effects of glyQTL on gene expression levels in relevant
686 tissues, we applied Summary data-based Mendelian Randomization (SMR) analysis
687 followed by the Heterogeneity in Dependent Instruments (HEIDI)³² on expression of
688 quantitative trait loci (eQTLs) obtained from Westra Blood eQTL collection⁸³ (peripheral
689 blood), GTEx (version 7) eQTL collection⁸⁴ (liver, whole blood), CEDAR eQTL
690 collection⁵³ (CD19+ B lymphocytes, CD8+ T lymphocytes, CD4+ T lymphocytes, CD14+
691 monocytes, CD15+ granulocytes) and on protein quantitative trait loci (pQTLs) using
692 SomaLogic datasets^{85,86}. As outcome variable we used univariate association results for
693 the N-glycome trait with the most significant association; in the case of glyQTLs
694 replicated only in multivariate analysis, we used summary statistics for the most
695 associated univariate trait as the primary trait in the analysis.

696 The results of the SMR test were considered statistically significant if $P_{adj} < 0.05$
697 (Benjamini-Hochberg adjusted P). The significance threshold for HEIDI tests was set at
698 $P = 0.05$ ($P < 0.05$ corresponds to the rejection of the pleiotropy hypothesis)
699 (Supplementary Table 6f, 6g).

700 DEPICT

701 Gene prioritization and gene set and tissue/cell type enrichment analyses were
702 performed using the Data-driven Expression Prioritized Integration for Complex Traits
703 framework (DEPICT)³¹. DEPICT analysis was conducted for SNPs associated with any
704 N-glycosylation trait at $P < 5 \times 10^{-8}/28$ in univariate analysis and with any N-
705 glycosylation trait group at $P < 5 \times 10^{-8}/21$ in multivariate analysis. The significance
706 threshold for DEPICT analysis was set at False Discovery Rate $FDR < 0.20$
707 (Supplementary Table 6h, 6i, 6j).

708 Colocalization with TF and IgG

709 In this study, colocalization analysis (SMR- θ)⁵³ was conducted for total plasma,
710 IgG and TF glyQTLs. The analysis was restricted to loci that were a) previously
711 implicated in TF GWAS⁴² (4 loci); IgG GWAS⁴⁰ (15 loci); both (2 loci) (Supplementary
712 Table 5d), b) reached genome-wide significance in the GWAMA of European descent
713 (N=10,172), c) replicated in this study. Statistic θ is a weighted correlation, whose
714 computation requires information on p-values and effect direction. The high absolute
715 value (e. g. $|\theta| > 0.7$) means the locus likely has a pleiotropic effect on investigated
716 traits.

717

718 Pleiotropy with disease

719 To study potential pleiotropic effects on a range of traits associated with various
720 medical conditions SMR/HEIDI analysis was carried out similarly to that for
721 colocalization with eQTL and pQTL.

722 Summary statistics for complex and medical conditions-related traits were
723 obtained from the UK Biobank⁸⁷, the CARDIoGRAM Consortium
724 (<http://www.cardiogramplusc4d.org/>), the Psychiatric Genomics consortium
725 (<https://pgc.unc.edu/>) and other trait collections from other studies (see Supplementary
726 Table 9 for the full list of the traits analyzed). We conducted analysis separately for the
727 disease-related traits and other complex traits.

728

729 Associations between PGS for plasma N-glycosylation traits and disease phenotypes

730 To test the associations between the 117 human plasma N-glycosylation traits
731 and ICD-10 disease phenotypes, we used logistic regression considering PGS for each
732 glycan trait as a predictor for each disease phenotype in turn.

733 The list of the diseases was taken from medical histories and questionnaires
734 obtained from non-related UK Biobank participants of European descent for which we
735 had PGS for N-glycosylation traits calculated (N = 374,303). All medical codes were
736 preliminary filtered by prevalence (> 0.5% and < 99.5%). For this analysis we used 167
737 groups of codes that fall into Chapters I-XV of the UK Biobank classification of
738 phenotypes. These codes describe a wide range of phenotypes including infectious
739 diseases, endocrine, nutritional and metabolic diseases, diseases of the nervous
740 system, diseases of the circulatory, respiratory, digestive and other systems, etc.

741 To perform logistic regression analyses we used the standard `glm()` function in R
742 v.4.2.2. programming language. We included sex, age, batch number and first ten
743 principal components of the kinship matrix (PC 1-10) as covariates in addition to the
744 PGS predictor. Finally, we filtered out the results not passing the significance threshold
745 for the association of $P < 0.05 / (28 \times 167) = 1.07 \times 10^{-5}$, where 28 is the number of
746 plasma N-glycome principal components explaining over 99% of the 117 N-
747 glycosylation traits variation, and 167 is the numbers of ICD-10 codes.

748

749 [Mendelian Randomization and Sensitivity Analyses](#)

750 In the previous step we identified 64 pairs of associated disease phenotypes and
751 plasma N-glycosylation traits. To investigate the causal relationships between these
752 traits we performed a bidirectional two-sample MR analysis³³: for each pair we
753 performed two MR analyses using the glycosylation trait as exposure and the disease
754 as outcome and *vice versa*.

755 As the sources of the summary statistics for MR analyses, we used the largest
756 available GWAMA for plasma N-glycosylation traits in a cohort of European descent
757 described previously in the current study (N = 10,172) and GWAS available from the UK
758 Biobank database (for more details about these cohorts see Supplementary Table 14).

759 The framework of the two-sample MR was specified before the analysis. Genetic
760 IVs for the two sample MR were identified as follows. First, the set of SNPs present both
761 in the GWAS for the exposure and outcome traits was selected. Then for this
762 overlapping set of SNPs in the GWAS for the exposure trait we performed clumping for
763 independence using PLINK2⁸¹ within a 10,000 kb window. Additional parameters for
764 clumping included an $r^2 > 0.001$ threshold for correlation, IVs with minor allele
765 frequency $MAF < 0.05$ were excluded. When plasma N-glycosylation traits were
766 considered as exposures, P threshold for clumping was defined as $5 \times 10^{-8} / 28 =$

767 1.79×10^{-9} (28 - number of plasma N-glycome principal components explaining over
768 99% of the 117 N-glycosylation traits variation). When the disease phenotypes were
769 considered as exposure, this threshold was set at $5 \times 10^{-8}/14 = 3.57 \times 10^{-9}$ (14 -
770 number of disease phenotypes significantly associated with at least one plasma N-
771 glycosylation trait in the logistic regression analysis).

772 Summary statistics for IVs in the exposure and outcome GWAS data were
773 processed using the TwoSampleMR R package³³: the data were harmonized excluding
774 ambiguous/triallelic SNPs. Only the pairs where at least 2 IVs were available for the
775 exposure trait were considered for the further analysis. MR analysis was performed
776 using `mr_report()` function from the TwoSampleMR R package. Significance thresholds
777 for the MR results were set as $\frac{0.05}{49} = 0.001$ for the analysis of 49 traits pairs where
778 glycans were considered as exposures, and as $\frac{0.05}{64} = 0.00078$ for the analysis of the 64
779 pairs where diseases were considered as exposures. If at least one of the MR methods
780 used (Inverse variance weighted, MR Egger, Weighted median, Weighted mode, or
781 Simple mode) produced a statistically significant causal estimate, that pair of traits was
782 selected for the follow-up sensitivity analyses.

783 Follow-up sensitivity analyses included those automatically implemented in the
784 `mr_report()` function, such as heterogeneity tests, test for directional horizontal
785 pleiotropy, leave-one-out analysis, forest plot and funnel plot.

786 In addition to the sensitivity analyses described above, for the pairs where
787 plasma N-glycosylation traits were used as exposures, since the number of IVs was
788 very low (2-3 SNPs), we performed colocalization analysis (SMR-HEIDI) for each of the
789 IVs. The results of the SMR test were considered statistically significant if Benjamini-
790 Hochberg adjusted $P < 0.05$. The significance threshold for HEIDI tests was set at
791 $P = 0.05$ which corresponds to the rejection of the pleiotropy hypothesis.

792 For the pairs where the disease was the upstream exposure trait (disorders of
793 lipoprotein metabolism and other lipidaemias, E78) and 27 IVs were available, we
794 identified pleiotropic IVs using MR-PRESSO R package, detecting the IVs for which P of
795 the test for outliers in MR-PRESSO were < 1 . Then we repeated the MR analysis as
796 described above excluding these SNPs.

797 [Data availability](#)

798 The full genome-wide summary association statistics for the 117 N-glycome traits will be
799 made publicly available **upon publication of the paper**. The data generated in the

800 secondary analyses of this study are included with this article in the Supplementary
801 Tables.

802 [Acknowledgements](#)

803 The work of S.Sh., A.T., D.M., A.S., Y.S.A. was supported by the Research Program at
804 the Moscow State University (MSU) Institute for Artificial Intelligence. The study was
805 conducted using the UK Biobank resource under application #59345. The work of E.E.,
806 Y.A.T was supported by the budget project of the Institute of Cytology and Genetics
807 FWNR-2022-0020. European Community's Seventh Framework Programme funded
808 project PainOmics (602736). TwinsUK is funded by the Wellcome Trust, Medical
809 Research Council, Versus Arthritis, European Union Horizon 2020, Chronic Disease
810 Research Foundation (CDRF), Zoe Ltd and the National Institute for Health Research
811 (NIHR) Clinical Research Network (CRN) and Biomedical Research Centre based at
812 Guy's and St Thomas' NHS Foundation Trust in partnership with King's College
813 London. The TwinsUK Study was approved by London-Westminster Research Ethics
814 Committee (REC reference EC04/015), and Guy's and St Thomas' NHS Foundation
815 Trust Research and Development (R&D). The TwinsUK BioBank was approved by the
816 HRA - Liverpool East Research Ethics Committee (REC reference 19/NW/0187), IRAS
817 ID 258513. All participants provide written, informed consent. We thank Toma Keser,
818 Mirna Šimurina, Marija Vilaj, Jerko Štambuk, Ivan Gudelj, Thomas S. Klarić, Jasminka
819 Krištić, Jelena Šimunović, Julija Jurić, Ana Momčilović, Najda Rudman, and Maja Hanić
820 for their assistance with glycan analysis.

821

822

823 [Author contributions](#)

824 S.Sh. coordinated this study; S.Sh., A.T., O.Z., D.M., A.S., E.E., E.T., A.N., S.F., N.A.P.,
825 Y.A.T. contributed to the design of the study, carried out statistical analysis; A.T., D.M.,
826 A.S., E.T., O.Z., S.Sh. produced the figures; S.Sh., A.T., O.Z., D.M., A.S., Y.A.T.,
827 contributed to interpretation of the results; S.Sh., A.T., O.Z., D.M., A.S., A.N. and Y.S.A.
828 wrote the first version of the manuscript; S.Sh., A.T., O.Z., D.M., A.S., and Y.S.A. wrote
829 the revised second version of the manuscript; S.Sh., E.T., E.E., S.F., V.V. and Y.A.T.
830 contributed to data harmonization and quality control; F.V., I.T.-A., T.Š. contributed to
831 plasma N-glycome measurements and quality control; M.M., T.S. analyzed TwinsUK
832 dataset and contributed to interpretation of the results; M.T., M.D. analyzed SOCCS
833 dataset and contributed to interpretation of the results; L.K., F.W., D.P., J. Van Z., M.A.

834 designed PainOmics study and contributed to interpretation of the results; K.S. analyzed
835 QMDiab dataset and contributed to interpretation of the results; M.G. designed CEDAR
836 study and contributed to interpretation of the results; C.W. and M.B.S. contributed to the
837 data collection and analyses of EPIC-Potsdam and to the interpretation of results;
838 Y.S.A. and G.L. conceived and oversaw the study, contributed to the design and
839 interpretation of the results; all co-authors contributed to the final manuscript revision.

840 [Competing interests statement](#)

841 Y.S.A. is a full-time employee of GSK PLC and receives salary and stock options as
842 compensation. G.L. is a founder and owner of Genos Ltd, a biotech company that
843 specializes in glycan analysis and has several patents in the field. O.Z., T.Š., F.V. and
844 I.T.-A are employees of Genos Ltd. All other authors declare no conflicts of interest.
845 Other authors declare no competing financial interests.

846 [References](#)

- 847 1. Apweiler, R., Hermjakob, H. & Sharon, N. On the frequency of protein glycosylation,
848 as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta*
849 **1473**, 4–8 (1999).
- 850 2. Clerc, F. *et al.* Human plasma protein N-glycosylation. *Glycoconj J* **33**, 309–343
851 (2016).
- 852 3. Ohtsubo, K. & Marth, J. D. Glycosylation in cellular mechanisms of health and
853 disease. *Cell* **126**, 855–867 (2006).
- 854 4. Skropeta, D. The effect of individual N-glycans on enzyme activity. *Bioorg Med*
855 *Chem* **17**, 2645–2653 (2009).
- 856 5. Takeuchi, H. *et al.* O-Glycosylation modulates the stability of epidermal growth
857 factor-like repeats and thereby regulates Notch trafficking. *J Biol Chem* **292**, 15964–
858 15973 (2017).
- 859 6. Lauc, G., Pezer, M., Rudan, I. & Campbell, H. Mechanisms of disease: The human
860 N-glycome. *Biochim Biophys Acta* **1860**, 1574–1582 (2016).
- 861 7. Reily, C., Stewart, T. J., Renfrow, M. B. & Novak, J. Glycosylation in health and
862 disease. *Nat Rev Nephrol* **15**, 346–366 (2019).

- 863 8. Dotz, V. & Wuhrer, M. N-glycome signatures in human plasma: associations with
864 physiology and major diseases. *FEBS Lett* **593**, 2966–2976 (2019).
- 865 9. Gudelj, I., Lauc, G. & Pezer, M. Immunoglobulin G glycosylation in aging and
866 diseases. *Cell Immunol* **333**, 65–79 (2018).
- 867 10. Dwek, R. A., Butters, T. D., Platt, F. M. & Zitzmann, N. Targeting glycosylation as a
868 therapeutic approach. *Nat Rev Drug Discov* **1**, 65–75 (2002).
- 869 11. Rodríguez, E., Schettters, S. T. T. & van Kooyk, Y. The tumour glyco-code as a
870 novel immune checkpoint for immunotherapy. *Nat Rev Immunol* **18**, 204–211
871 (2018).
- 872 12. Pagan, J. D., Kitaoka, M. & Anthony, R. M. Engineered Sialylation of Pathogenic
873 Antibodies In Vivo Attenuates Autoimmune Disease. *Cell* **172**, 564-577.e13 (2018).
- 874 13. Van Landuyt, L., Lonigro, C., Meuris, L. & Callewaert, N. Customized protein
875 glycosylation to improve biopharmaceutical function and targeting. *Curr Opin*
876 *Biotechnol* **60**, 17–28 (2019).
- 877 14. Johannssen, T. & Lepenies, B. Glycan-Based Cell Targeting To Modulate Immune
878 Responses. *Trends in Biotechnology* **35**, 334–346 (2017).
- 879 15. Paderi, J., Prestwich, G. D., Panitch, A., Boone, T. & Stuart, K. Glycan
880 Therapeutics: Resurrecting an Almost Pharma-Forgotten Drug Class. *Advanced*
881 *Therapeutics* **1**, 1800082 (2018).
- 882 16. Adamczyk, B., Tharmalingam, T. & Rudd, P. M. Glycans as cancer biomarkers.
883 *Biochimica et Biophysica Acta (BBA) - General Subjects* **1820**, 1347–1353 (2012).
- 884 17. Thanabalasingham, G. *et al.* Mutations in HNF1A result in marked alterations of
885 plasma glycan profile. *Diabetes* **62**, 1329–1337 (2013).
- 886 18. Shinohara, Y., Furukawa, J. & Miura, Y. Glycome as Biomarkers. in *General*
887 *Methods in Biomarker Research and their Applications* (eds. Preedy, V. R. & Patel,

- 888 V. B.) 111–140 (Springer Netherlands, Dordrecht, 2015). doi:10.1007/978-94-007-
889 7696-8_23.
- 890 19. Taniguchi, N. *Handbook of Glycosyltransferases and Related Genes, Second*
891 *Edition // Handbook of Glycosyltransferases and Related Genes.* (2014).
- 892 20. Lauc, G., Vojta, A. & Zoldoš, V. Epigenetic regulation of glycosylation is the
893 quantum mechanics of biology. *Biochim Biophys Acta* **1840**, 65–70 (2014).
- 894 21. Moremen, K. W., Tiemeyer, M. & Nairn, A. V. Vertebrate protein glycosylation:
895 diversity, synthesis and function. *Nat. Rev. Mol. Cell Biol.* **13**, 448–462 (2012).
- 896 22. Knezevic, A. *et al.* Variability, heritability and environmental determinants of human
897 plasma N-glycome. *J. Proteome Res.* **8**, 694–701 (2009).
- 898 23. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The
899 carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**,
900 D490-495 (2014).
- 901 24. Lauc, G., Rudan, I., Campbell, H. & Rudd, P. M. Complex genetic regulation of
902 protein glycosylation. *Mol Biosyst* **6**, 329–335 (2010).
- 903 25. Timoshchuk, A., Sharapov, S. & Aulchenko, Y. S. Twelve Years of Genome-Wide
904 Association Studies of Human Protein N-Glycosylation. *Engineering* (2023)
905 doi:10.1016/j.eng.2023.03.013.
- 906 26. Krištić, J., Sharapov, S. Z. & Aulchenko, Y. S. Quantitative Genetics of Human
907 Protein N-Glycosylation. *Adv Exp Med Biol* **1325**, 151–171 (2021).
- 908 27. Reiding, K. R. *et al.* High-throughput Serum N-Glycomics: Method Comparison and
909 Application to Study Rheumatoid Arthritis and Pregnancy-associated Changes. *Mol.*
910 *Cell. Proteomics* **18**, 3–15 (2019).
- 911 28. Shen, X. *et al.* Multivariate discovery and replication of five novel loci associated
912 with Immunoglobulin G N-glycosylation. *Nat Commun* **8**, 447 (2017).

- 913 29. Ning, Z. *et al.* Nontrivial Replication of Loci Detected by Multi-Trait Methods. *Front*
914 *Genet* **12**, 627989 (2021).
- 915 30. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary
916 association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).
- 917 31. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using
918 predicted gene functions. *Nat Commun* **6**, 5890 (2015).
- 919 32. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts
920 complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
- 921 33. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference
922 across the human phenome. *Elife* **7**, (2018).
- 923 34. Lauc, G. *et al.* Genomics Meets Glycomics—The First GWAS Study of Human N-
924 Glycome Identifies HNF1 α as a Master Regulator of Plasma Protein Fucosylation.
925 *PLoS genetics* (2010) doi:10.1371/journal.pgen.1001256.
- 926 35. Huffman, J. E. *et al.* Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are
927 associated with variation within the human plasma N-glycome of 3533 European
928 adults. *Hum Mol Genet* **20**, 5000–5011 (2011).
- 929 36. Sharapov, S. *et al.* *Defining the Genetic Control of Human Blood Plasma N-*
930 *Glycome Using Genome-Wide Association Study.* (2018). doi:10.1101/365486.
- 931 37. Sharapov, S. Z. *et al.* Replication of 15 loci involved in human plasma protein N-
932 glycosylation in 4802 samples from four cohorts. *Glycobiology* **31**, 82–88 (2021).
- 933 38. Lauc, G. *et al.* Loci associated with N-glycosylation of human immunoglobulin G
934 show pleiotropy with autoimmune diseases and haematological cancers. *PLoS*
935 *Genet.* **9**, e1003225 (2013).
- 936 39. Wahl, A. *et al.* Genome-Wide Association Study on Immunoglobulin G Glycosylation
937 Patterns. *Front. Immunol.* **9**, 277 (2018).

- 938 40. Klarić, L. *et al.* Glycosylation of immunoglobulin G is regulated by a large network of
939 genes pleiotropic with inflammatory diseases. *Sci Adv* **6**, eaax0301 (2020).
- 940 41. Shadrina, A. S. *et al.* Multivariate genome-wide analysis of immunoglobulin G N-
941 glycosylation identifies new loci pleiotropic with immune function. *Hum Mol Genet*
942 **30**, 1259–1270 (2021).
- 943 42. Landini, A. *et al.* Genetic regulation of post-translational modification of two distinct
944 proteins. *Nat. Commun.* **13**, 1586 (2022).
- 945 43. Zaytseva, O. O. *et al.* Investigation of the causal relationships between human IgG
946 N-glycosylation and 12 common diseases associated with changes in the IgG N-
947 glycome. *Hum Mol Genet* **31**, 1545–1559 (2022).
- 948 44. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from
949 polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
- 950 45. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary
951 statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**,
952 369–75, S1-3 (2012).
- 953 46. Zaytseva, O. O. *et al.* Heritability of Human Plasma N-Glycome. *J Proteome Res* **19**,
954 85–91 (2020).
- 955 47. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple
956 regression on summary statistics. *Nat Commun* **10**, 5086 (2019).
- 957 48. Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and
958 genes for coronary artery disease in over a million participants. *Nat Genet* **54**, 1803–
959 1815 (2022).
- 960 49. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122
961 (2016).
- 962 50. Rogers, M. *et al.* FATHMM-XF: accurate prediction of pathogenic point mutations
963 via extended features. *Bioinform.* (2018) doi:10.1093/bioinformatics/btx536.

- 964 51. Ferlaino, M. *et al.* An integrative approach to predicting the functional effects of
965 small indels in non-coding regions of the human genome. *BMC Bioinformatics* **18**,
966 442 (2017).
- 967 52. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal
968 genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, e3 (2019).
- 969 53. Momozawa, Y. *et al.* IBD risk loci are enriched in multigenic regulatory modules
970 encompassing putative causative genes. *Nature Communications* **9**, (2018).
- 971 54. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic
972 association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
- 973 55. Williams, F. M. K. *et al.* Ischemic stroke is associated with the ABO locus: the
974 EuroCLOT study. *Ann Neurol* **73**, 16–31 (2013).
- 975 56. Sabotta, C. M. *et al.* Genetic variants associated with circulating liver injury markers
976 in Mexican Americans, a population at risk for non-alcoholic fatty liver disease. *Front*
977 *Genet* **13**, 995488 (2022).
- 978 57. Ewald, D. R. & Sumner, S. C. J. Blood type biochemistry and human disease. *Wiley*
979 *Interdiscip Rev Syst Biol Med* **8**, 517–535 (2016).
- 980 58. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal
981 pleiotropy in causal relationships inferred from Mendelian randomization between
982 complex traits and diseases. *Nat Genet* **50**, 693–698 (2018).
- 983 59. Fairfield, C. J. *et al.* Genome-Wide Association Study of NAFLD Using Electronic
984 Health Records. *Hepatology Communications* **6**, 297 (2022).
- 985 60. Vujkovic, M. *et al.* A multiancestry genome-wide association study of unexplained
986 chronic ALT elevation as a proxy for nonalcoholic fatty liver disease with histological
987 and radiological validation. *Nat Genet* **54**, 761–771 (2022).

- 988 61. Hülsmeyer, A. J., Tobler, M., Burda, P. & Hennet, T. Glycosylation site occupancy in
989 health, congenital disorder of glycosylation and fatty liver disease. *Sci Rep* **6**, 33927
990 (2016).
- 991 62. Verhelst, X. *et al.* Protein Glycosylation as a Diagnostic and Prognostic Marker of
992 Chronic Inflammatory Gastrointestinal and Liver Diseases. *Gastroenterology* **158**,
993 95–110 (2020).
- 994 63. Blomme, B., Van Steenkiste, C., Callewaert, N. & Van Vlierberghe, H. Alteration of
995 protein glycosylation in liver diseases. *Journal of Hepatology* **50**, 592–603 (2009).
- 996 64. Gong, J., Tu, W., Liu, J. & Tian, D. Hepatocytes: A key role in liver inflammation.
997 *Front. Immunol.* **13**, (2023).
- 998 65. Radovani, B. & Gudelj, I. N-Glycosylation and Inflammation; the Not-So-Sweet
999 Relation. *Front Immunol* **13**, 893365 (2022).
- 1000 66. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci
1001 and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122 (2016).
- 1002 67. Kiryluk, K., Novak, J. & Gharavi, A. G. Pathogenesis of Immunoglobulin A
1003 Nephropathy: Recent Insight from Genetic Studies. *Annual Review of Medicine* **64**,
1004 339–356 (2013).
- 1005 68. Kiss, M. G. *et al.* Complement Factor H Modulates Splenic B Cell Development and
1006 Limits Autoantibody Production. *Front Immunol* **10**, 1607 (2019).
- 1007 69. Dotz, V. *et al.* O- and N-Glycosylation of Serum Immunoglobulin A is Associated
1008 with IgA Nephropathy and Glomerular Function. *J Am Soc Nephrol* **32**, 2455–2465
1009 (2021).
- 1010 70. Momčilović, A. *et al.* Simultaneous Immunoglobulin A and G Glycopeptide Profiling
1011 for High-Throughput Applications. *Anal Chem* **92**, 4518–4526 (2020).
- 1012 71. Demus, D. *et al.* Large-Scale Analysis of Apolipoprotein CIII Glycosylation by
1013 Ultrahigh Resolution Mass Spectrometry. *Frontiers in Chemistry* **9**, (2021).

- 1014 72. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K.
1015 Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat*
1016 *Commun* **11**, 5650 (2020).
- 1017 73. Akmačić, I. T. *et al.* High-throughput glycomics: optimization of sample preparation.
1018 *Biochemistry (Mosc.)* **80**, 934–942 (2015).
- 1019 74. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient
1020 normalization as robust method to account for dilution of complex biological
1021 mixtures. Application in 1H NMR metabonomics. *Anal Chem* **78**, 4281–4290 (2006).
- 1022 75. Benedetti, E. *et al.* Systematic Evaluation of Normalization Methods for Glycomics
1023 Data Based on Performance of Network Inference. *Metabolites* **10**, 271 (2020).
- 1024 76. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation.
1025 *Nat Genet* **48**, 1279–1283 (2016).
- 1026 77. 1000 Genomes Project Consortium *et al.* A global reference for human genetic
1027 variation. *Nature* **526**, 68–74 (2015).
- 1028 78. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues
1029 of a correlation matrix. *Heredity (Edinb)* **95**, 221–227 (2005).
- 1030 79. Shashkova, T. I. *et al.* The GWAS-MAP platform for aggregation of results of
1031 genome-wide association studies and the GWAS-MAP|homo database of 70 billion
1032 genetic associations of human traits. *Vavilovskii Zhurnal Genet Selektiv* **24**, 876–
1033 884 (2020).
- 1034 80. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic
1035 data. *Nature* **562**, 203–209 (2018).
- 1036 81. Chen, Z.-L. *et al.* A high-speed search engine pLink 2 with systematic evaluation for
1037 proteome-scale identification of cross-linked peptides. *Nat Commun* **10**, 3404
1038 (2019).

- 1039 82. Narimatsu, Y. *et al.* An Atlas of Human Glycosylation Pathways Enables Display of
1040 the Human Glycome by Gene Engineered Cells. *Mol Cell* **75**, 394-407.e5 (2019).
- 1041 83. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of
1042 known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
- 1043 84. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues.
1044 *Nature* **550**, 204–213 (2017).
- 1045 85. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human
1046 blood plasma proteome. *Nat Commun* **8**, 14357 (2017).
- 1047 86. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79
1048 (2018).
- 1049 87. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK
1050 Biobank. *Nat Genet* **50**, 1593–1599 (2018).
- 1051 88. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human
1052 and mouse. *Nat Commun* **9**, 1366 (2018).

1053

1054 **Legends to Main Figures**

1055 **Figure 1: Discovered loci.** (a) Associations of 59 loci with 138 glycomic traits labeled
1056 by the prioritized candidate or nearest gene names. Marked black: loci are discovered
1057 and replicated in this work; red: discovered, but not replicated in this work. In total, 117
1058 univariate traits were analyzed (Supplementary Table 3a), but for two of them, no
1059 genome-wide significant associations were found. Univariate traits are grouped into 4
1060 categories: glycans mostly linked to immunoglobulins (green), glycans mostly linked to
1061 non-immunoglobulin proteins (purple), glycans linked to both types of proteins (orange),
1062 not classified glycans (gray). The details of glycan classification are described in
1063 Supplementary Note. Also, the results from analysis of 21 multivariate traits (turquoise)
1064 are presented. The multivariate traits were defined based on biochemical similarities
1065 between 36 directly measured total plasma N-glycan traits (Supplementary Table 3b).
1066 (b) A network view of associations between loci and glycan traits. Rectangular nodes
1067 represent genetic loci labeled with the names of the prioritized candidate or nearest

1068 genes; circle nodes represent glycan traits. Lines represent significant genetic
1069 associations between locus and specific glycans. The colors of circle nodes are
1070 consistent with those in (a).

1071 Symbol ○ next to the candidate gene indicates that the locus was previously discovered
1072 in immunoglobulin G N-glycome GWASs; Symbol ● — the locus was previously
1073 discovered in transferrin N-glycome GWAS, as reviewed in ²⁵.

1074

1075 **Figure 2: Candidate genes.** (a) Gene expression of the candidate genes in two
1076 relevant cell types - hepatocytes and plasma cells. Expression levels are represented
1077 as the median logarithm of transcripts per million. The data for hepatocytes (N = 513)
1078 and plasma cells (N = 53) samples were obtained from the ARCHS4 portal ⁸⁸.

1079 (b) Predictors indicating the 32 candidate genes. Gene order corresponds to (a). The
1080 identical superscripts denote candidate genes inside one locus. Full details of the gene
1081 prioritization are presented in Supplementary Table 6b. Symbol ○ next to the candidate
1082 gene indicates that the locus was previously discovered in immunoglobulin G N-
1083 Glycome GWASs; Symbol ● — the locus was previously discovered in transferrin N-
1084 Glycome GWAS, as reviewed in ²⁵.

1085

1086 **Figure 3: Phenome-wide colocalization results for significant and replicated loci.**

1087 (a) Heatmap of traits with pleiotropic associations ($P_{adjSMR} < 0.05$ and $P_{HEIDI} \geq$
1088 0.05). The row order is determined by clustering on the set of significantly colocalized
1089 traits (Jaccard similarity index and Ward's linkage). For the sake of readability similar
1090 traits are grouped into broader categories, e.g., such traits as fat-free mass of left leg,
1091 trunk fat mass are combined under the name "Body mass". The numbers in the cells
1092 represent the number of pleiotropic associations, grouped under a common name,
1093 confirmed for a given locus. Color does not carry semantic load.

1094 (b) Heatmap providing details of pleiotropic associations with diseases from (a). Color
1095 does not carry semantic load.

1096 (c) Count of traits for each locus with pleiotropic associations.

1097 Symbol ○ next to the candidate gene indicates that the locus was previously discovered
1098 in immunoglobulin G N-Glycome GWASs; Symbol ● — the locus was previously
1099 discovered in transferrin N-Glycome GWAS, as reviewed in ²⁵. Abbreviations: BMI: body
1100 mass index; BN: benign neoplasm; CLL: chronic leukocytic leukemia; T1D: type 1
1101 diabetes; T2D: type 2 diabetes; BD: bipolar disorder; BP: blood pressure; MI:
1102 myocardial infarction; CAD: coronary artery's disease; CD: Crohn disease; IBD:

1103 Inflammatory bowel disease; IHD: Ischemic heart disease; NC: non-cancer, non-I: non-
 1104 iodine dependent; RA: rheumatoid arthritis; SCZ: schizophrenia; UC: ulcerative colitis.
 1105

1106 **Figure 4: Significant associations between PGS for plasma N-glycosylation traits**
 1107 **and ICD-10 diseases.**

1108 Heatmap of associations between PGS for plasma N-glycosylation traits and ICD-10
 1109 diseases that were statistically significant at the designated significance threshold of
 1110 1.07×10^{-5} . Every column represents an N-glycosylation trait, every row – an ICD-10
 1111 disease. The colour of the cell represents the effects estimated in the logistic regression
 1112 of the disease phenotype incidence on the PGS for each glycan trait – blue hues
 1113 represent negative effects, red hues – positive, and the intensity of the colour
 1114 represents the absolute value of the effect (larger values are shown in darker hues).
 1115 Grey cells represent the non-significant associations. The bands under the heatmap
 1116 depict the groups of N-glycosylation traits that are related to certain structural features
 1117 of the N-glycans. T1D – type 1 diabetes.

1118

1119 **Table 1: Twenty-five novel loci associated with total plasma N-glycosylation**
 1120 **discovered and replicated in this study.** Full results of discovery and replication are
 1121 provided in Supplementary Table 5a, 5b. CHR:POS—chromosome and position of SNP
 1122 according to GRCh37 human genome build; EA/RA—effective and reference allele;
 1123 Gene—prioritized or nearest gene for a locus (Supplementary Table 6b); N—sample
 1124 size; EAF—effective allele frequency; BETA (SE)—effect (in SD units) and standard
 1125 error of effect; *P*—*P*-value; Top trait—glycan trait with the strongest association (the
 1126 lowest *P*); Type of association—univariate or multivariate. Description of glycan traits is
 1127 provided in Supplementary Table 3. Symbol ○ — the locus was previously discovered in
 1128 immunoglobulin G N-glycome GWASs, as reviewed in ²⁵.

1129

SNP info				Discovery					
SNP	CHR:POS	EA/RA	Gene	N	EAF	BETA (SE)	<i>P</i>	Top trait	Type of association
rs12726286	1:93334379	C/T	DIPK1A	7540	73.52%	0.022 (0.002)	2.24E-21	sialylation of antennary branches	multivariate
rs1329427	1:196704559	C/T	CFH	7540	57.63%	0.016 (0.003)	5.61E-10	N-glycosylation	multivariate
rs1260326	2:27730940	C/T	GCKR	7081	57.51%	-0.168 (0.018)	9.65E-22	G3total	univariate
rs2470750	3:98690592	A/T	ST3GAL6	6790	59.04%	0.012 (0.002)	7.23E-10	high branching glycans	multivariate
rs3774964	4:103519487	A/G	SLC39A8	7164	63.69%	0.109 (0.018)	1.49E-09	G0n	univariate
rs7705720	5:95280033	C/T	ELL2 ○	7540	20.03%	-0.133 (0.021)	1.32E-10	FG1S1/(FG1+FG1S1)	univariate
rs4543384	6:139629524	C/T	TXLNB ○	6790	55.74%	0.143 (0.018)	9.67E-16	FA2BG1n	univariate

rs7758383	6:143169723	A/G	HIVEP2 ◊	7540	49.87%	0.153 (0.017)	3.55E-20	FA2G2n	univariate
rs34166762	7:73018524	C/T	MLXIPL	7540	27.58%	-0.121 (0.019)	1.09E-10	A3G3S2	univariate
rs7781265	7:150950940	A/G	SMARCD3 ◊	7540	10.99%	0.223 (0.027)	4.75E-16	M64	univariate
rs4841133	8:9183664	A/G	LOC157273	7540	8.37%	0.013 (0.002)	6.85E-10	sialylation of antennary branches	multivariate
rs28601761	8:126500031	C/G	TRIB1	6790	57.56%	-0.143 (0.018)	7.10E-15	G3Fa/G3total	univariate
rs582118	9:136145471	A/G	ABO	7540	65.12%	0.023 (0.003)	2.00E-17	N-glycosylation	multivariate
rs174528	11:61543499	C/T	FADS2	7540	35.83%	0.013 (0.002)	1.50E-09	sialylation of antennary branches	multivariate
rs36020612	11:123344435	C/T	GRAMD1B	7319	79.96%	-0.169 (0.022)	1.66E-14	FBG2S1/ (FBG2+FBG2S1+FBG2S2)	univariate
rs11223982	11:134612702	A/G	LINC02714	6280	12.47%	0.256 (0.029)	3.45E-19	A4G4S3	univariate
rs7161378	14:65450780	C/T	MAX	7540	24.74%	-0.180 (0.019)	2.05E-20	FG3/G3total	univariate
rs28929474	14:94844947	C/T	SERPINA1	5135	97.71%	0.437 (0.072)	1.01E-09	A2G2S2	univariate
rs8046823	16:71400131	A/G	CALB2	7540	51.53%	0.021 (0.003)	1.46E-16	galactosylation of antennary branches	multivariate
rs217184	16:72105965	C/T	HP	7540	19.90%	0.211 (0.021)	5.44E-23	S3total	univariate
rs4500785	17:16848565	C/G	TNFRSF13B ◊	7540	88.63%	-0.180 (0.027)	1.33E-11	FA2BG1	univariate
rs2659007	17:79217478	A/G	SLC38A10 ◊	7540	46.44%	0.108 (0.017)	4.07E-10	FA2BG1n	univariate
rs11669860	19:19277296	A/G	MEF2B ◊	7540	55.43%	-0.113 (0.017)	7.14E-11	FA2BG1n	univariate
rs2618588	20:17832658	C/T	LOC107985440 ◊	7540	39.69%	0.011 (0.002)	2.02E-10	core-fucosylation	multivariate
rs2834847	21:36588180	A/C	RUNX1 ◊	7540	77.22%	0.160 (0.020)	2.58E-15	FBn	univariate

1130

1131

1132

1133

1134 **Table 2: Causal associations between plasma N-glycosylation traits and ICD-10**

1135 **diseases discovered in this study.** Full results of the analysis are provided in

1136 Supplementary Tables 10a, 10b. Exposure – upstream trait in the analysis (cause);

1137 Outcome – downstream trait in the analysis; IV- instrumental variable (independent

1138 SNPs associated with upstream trait); MR – Mendelian randomization; Causal BETA

1139 (SE) – causal effect of the exposure on the outcome (in log OR units of the disease unit

1140 per SD units of glycan trait if glycan trait was an exposure, or in SD units of glycan trait

1141 per log OR unit of the disease phenotype if disease phenotype was an exposure) with

1142 its standard error; *P* - P-value of the MR analysis.

1143

Exposure	Outcome	MR method	Number of IVs	Causal BETA (SE)	<i>P</i>
PGP64 - The percentage of M6 in total neutral plasma glycans (GPn)	J45 - Asthma	Weighted median	3	0.012 (0,0036)	9.78E-04
PGP69 - The percentage of M9 in total neutral plasma glycans (GPn)	E78 - Disorders of lipoprotein metabolism and other lipidaemias	Inverse variance weighted	2	0.013 (0,0035)	1.25E-04
E78 - Disorders of lipoprotein metabolism and other lipidaemias	PGP18 - The percentage of M9	Inverse variance weighted	27	4.09 (1,1772)	5.16E-04

E78 - Disorders of lipoprotein metabolism and other lipidaemias	PGP107 - The percentage of high-mannose structures in total plasma glycans	Weighted median	27	3.01 (0,8756)	5.87E-04
---	--	-----------------	----	---------------	----------

1144





