

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

3

4 Arianna Landini^{1,2}, Paul R.H.J. Timmers^{1,2}, Azra Frkatović-Hodžić³, Irena Trbojević-
5 Akmačić³, Frano Vučković³, Tea Pribić, Regeneron Genetics Center⁴, Gannie Tzoneva⁴, Alan
6 R. Shuldiner⁴, Ozren Polašek^{5,6}, Caroline Hayward¹, Gordan Lauc^{3,7}, James F. Wilson*^{1,2} &
7 Lucija Klarić*¹

8 1 MRC Human Genetics Unit, Institute for Genetics and Cancer, University of Edinburgh,
9 Edinburgh, United Kingdom

10 2 Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh,
11 United Kingdom

12 3 Genos Glycoscience Research Laboratory, Zagreb, Croatia

13 4 Regeneron Genetics Center, Tarrytown, NY, USA

14 5 Department of Public Health, School of Medicine, University of Split, Split, Croatia

15 6 Algebra University College, Zagreb, Croatia

16 7 Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb, Croatia

17

18 * Authors contributed equally.

19 Correspondence to: J.F.W (jim.wilson@ed.ac.uk) or L.K. (lucija.klaric@ed.ac.uk)

20

21 **Abstract**

22 It is often difficult to be certain which genes underlie the effects seen in association studies.
23 However, variants that disrupt the protein, such as predicted loss of function (pLoF) and
24 missense variants, provide a shortcut to identify genes with a clear biological link to the
25 phenotype of interest. Glycosylation is one of the most common post-
26 translational modifications of proteins, and an important biomarker of both disease and its
27 progression. Here, we utilised the power of genetic isolates, gene-based aggregation tests and
28 intermediate phenotypes to assess the effect of rare (MAF<5%) pLoF and missense variants
29 from whole exome sequencing on the N-glycome of plasma transferrin (N=1907) and
30 immunoglobulin G (N=4912), and their effect on diseases. We identified significant gene-
31 based associations for transferrin glycosylation at 5 genes ($p < 8.06 \times 10^{-8}$) and for IgG glycan
32 traits at 4 genes ($p < 1.19 \times 10^{-7}$). Associations in three of these genes (*FUT8*, *MGAT3* and
33 *RFXAP*) are driven by multiple rare variants simultaneously contributing to protein
34 glycosylation. Association at *ST6GALI*, with a 300-fold up-drifted variant in the Orkney
35 Islands, was detectable by a single-point exome-wide association analysis. Glycome-associated
36 aggregate associations are located in genes already known to have a biological link to protein
37 glycosylation (*FUT8*, *RFXAP* for transferrin, *FUT8*, *MGAT3* and *ST6GALI* for IgG) but also in

38 genes which have not been previously reported (e.g. *RFXAP* for IgG). To assess the potential
39 impact of rare variants associated with glycosylation on other traits, we queried public
40 repositories of gene-based tests, discovering a potential connection between transferrin
41 glycosylation, *MSRI*, galectin-3, insulin-like growth factor 1 and diabetes. However, the exact
42 mechanism behind these connections requires further elucidation.

43 Introduction

44 Genome-wide association studies (GWAS) have so far identified thousands of loci associated
45 with human complex traits and diseases. However, the large majority of these variants are
46 found in noncoding regions of the genome¹, posing a challenge when attempting to uncover
47 their functional impact on the phenotype. On the contrary, whole-exome sequencing (WES)
48 studies offer the opportunity to identify rare variants of larger effect on the encoded protein,
49 such as predicted loss of function (pLoF) and missense variants, for which causal biological
50 mechanisms are generally easier to elucidate². Methods for exome-wide rare variant analysis
51 have been successfully employed to discover variants and genes associated with both complex
52 molecular traits³ and diseases^{4,5}. While single-variant tests, such as GWAS, are largely adopted
53 to explore associations of common genetic variants with phenotypes of interest, they have little
54 power to identify rare variant associations, due to the low number of observations. Therefore,
55 a set of methods testing cumulative effects of multiple rare variants in genetic regions, where
56 rare variants are grouped at the gene level (also known as ‘masks’) via a collapsing test, such
57 as burden tests, or variance-component tests (e.g. sequence kernel association test, SKAT⁶)
58 were developed. In addition to increasing the statistical power by aggregating multiple rare-
59 variants, using genetically isolated populations can provide unique opportunities for novel
60 discovery in an association study⁷. Recent bottlenecks, restricted immigration and limited
61 population size lead to increased genetic drift. Consequently, in such populations some
62 otherwise rare variants can substantially increase in frequency compared to the general
63 population, therefore increasing association power for these variants.

64 Glycosylation is one of the most frequent post-translational modifications, where sugar
65 residues, called glycans, are attached to the surface of proteins. Changes in protein N-
66 glycosylation patterns have been described in the ageing process^{8,9} and in a wide variety of
67 complex diseases, including autoimmune diseases¹⁰, diabetes¹¹, cardiovascular diseases¹²,
68 neurodegenerative diseases¹³ and cancer¹⁴. Despite glycans having an important role in human
69 health and serving as potential biomarkers in clinical prognosis and diagnosis¹⁵, we have just
70 started scratching the surface of the complex network of genes regulating protein glycosylation.
71 All studies published to date exploring the genetic regulation of total plasma protein,
72 immunoglobulin G (IgG) and transferrin N-glycosylation have employed single variant-based
73 GWAS tests, mostly uncovering common variants located in non-coding regions of the
74 genome^{16–23}. Rare variants contributing to glycan variation, and their impact on human health,
75 thus remain unexplored.

76 To address this knowledge gap, we used multiple gene-based aggregation tests to investigate
77 how rare (MAF<5%) pLoF and missense variants from whole exome sequencing affect 51
78 transferrin (N = 1907) and 94 IgG (N = 4912) glycan traits in European-descent cohorts. IgG
79 is both the most abundant antibody and one of the most abundant proteins in human serum. It
80 contains evolutionary conserved N-glycosylation sites in the constant region of each of its
81 heavy chains, occupied by biantennary, largely core-fucosylated and partially truncated glycan
82 structures, that may carry a bisecting N-acetylglucosamine and sialic acid residues^{24,25}.
83 Transferrin is a blood plasma glycoprotein that binds iron (Fe) and consequently mediates its

84 transport through blood plasma. Human transferrin has two N-glycosylation sites, with
85 biantennary disialylated digalactosylated glycan structure without fucose being the most
86 abundant glycan attached^{26,27}.

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

95 Results

96 Exome variant annotation

97 To assess the effect of rare genetic variants on glycosylation of two proteins, we sequenced the
 98 exomes of 4,801 participants of European ancestry. After quality control, a total of 233,820
 99 distinct autosomal coding genetic variants were available in the ORCADES cohort (N=2090),
 100 244,649 in the VIKING cohort (N=2106) and 340,203 in the CROATIA-Korcula cohort
 101 (N=2872). Percentages of variants for each effect category in the total sequenced coding
 102 variation are similar across the three cohorts (Table 1). More than half (~53%) of the sequenced
 103 coding variants are missense variants, of which nearly half (~28% of total coding variation) are
 104 classified as likely or possibly deleterious by multiple variant effect predictor algorithms (see
 105 Methods). The second most represented effect category is synonymous mutations (~33%),
 106 followed by variants in splice regions (~8%), predicted loss of function (pLoF) (~4%) and in-
 107 frame insertions/deletions (~1.5%). Around one quarter of coding variants in the ORCADES
 108 and VIKING cohorts are singletons (minor allele count, MAC=1); this percentage is instead
 109 higher in CROATIA-Korcula cohort (~35%), possibly due to the larger sequenced sample size.

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

Variant category	ORCADES (N=2090)			CROATIA-Korcula (N=2872)			VIKING (N=2106)		
	No. of variants	% of total coding variants	Variants % with MAC=1	No. of variants	% of total coding variants	Variants % with MAC=1	No. of variants	% of total coding variants	Variants % with MAC=1
coding variants	233,820		25.1%	340,203		35.5%	244,649		28.9%
pLOF	8639	3.69%	37.1%	12,970	3.81%	47.2%	9025	3.69%	41.4%
Splice acceptor	872	0.37%	37.8%	1309	0.38%	45.5%	945	0.39%	42.7%
Splice donor	1042	0.45%	37.5%	1506	0.44%	48.1%	1079	0.44%	40.4%
Stop gained	2833	1.21%	36.8%	4171	1.23%	47%	2879	1.18%	41.8%
Frameshift	3401	1.45%	37.8%	5274	1.55%	47.9%	3583	1.46%	42.1%
Stop lost	151	0.06%	32.5%	244	0.07%	43%	182	0.07%	39%
Start lost	340	0.15%	30%	466	0.14%	44.2%	357	0.15%	30.3%
Missense	124,416	53.2%	27%	183,056	53.8%	37.6%	130,299	53.3%	30.6%
Likely benign (0-1)	56,366	24.1%	21.8%	80,777	23.7%	31.6%	59,141	24.2%	25.4%
Possibly deleterious (2-3)	31,693	13.5%	28.1%	47,235	13.9%	39.1%	33,138	13.5%	31.6%

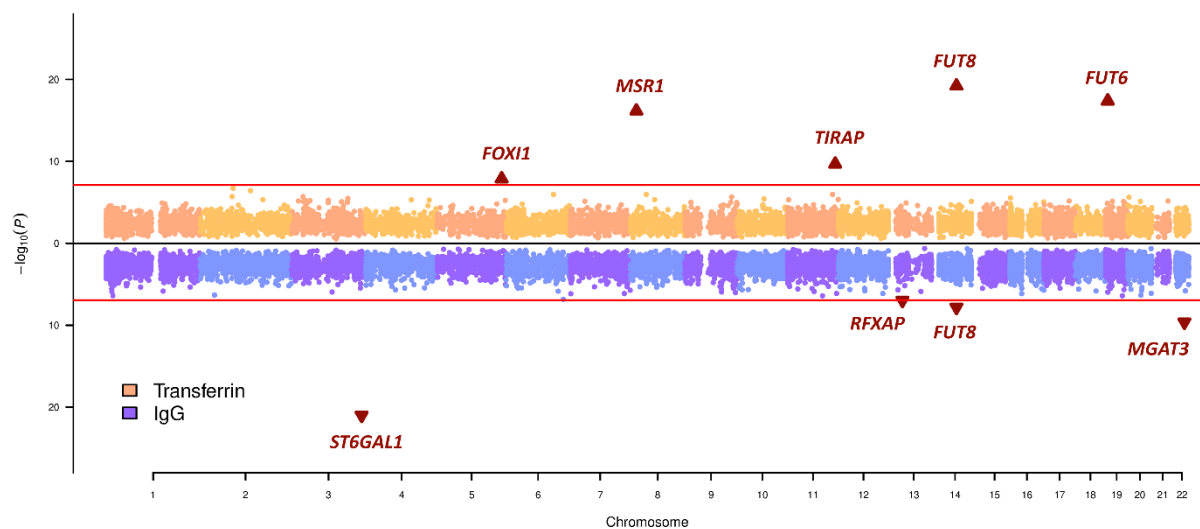
Likely deleterious (4-5)	35,728	15.3%	34.2%	54,187	15.9%	45.4%	37,384	15.3%	38.2%
Unclassified missense	629	0.27%	23%	857	0.25%	29.3%	636	0.26%	23.1%
Splice region	18,580	7.95%	24.7%	26660	7.84%	32.4%	19,297	7.89%	27.8%
In-frame indel	3383	1.45%	21.5%	5244	1.54%	29.5%	3606	1.47%	26%
Protein altering	3	0%	33.3%	4	0%	25%	2	0%	0%
Stop retained	1	0%	0%	2	0%	50%	1	0%	25.2%
Synonymous	78,798	33.7%	21.1%	112,267	33%	31.6%	82,419	33.7%	100%

114

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

116 We performed exome-wide gene-based tests across 51 transferrin traits (glycome subset of
 117 CROATIA-Korcula N = 948, VIKING N = 959) and 94 IgG glycan traits (glycome subset of
 118 ORCADES N = 1960, CROATIA-Korcula N = 1866, VIKING N = 1086), testing low
 119 frequency and rare (MAF <5%) pLoF and missense variants. In total, we identified 16
 120 significant associations for transferrin- (Supplementary Table 1) and 32 significant associations
 121 for IgG- (Supplementary table 2) glycan traits, at Bonferroni-corrected p-values of 8.06×10^{-8}
 122 and 1.19×10^{-7} , respectively (Figure 1, Table 2). Most gene-aggregated rare variants were
 123 associated with protein-specific glycans (transferrin: variants in *FUT6*, *TIRAP*, *MSR1* and
 124 *FOXII* genes, IgG: variants in *MGAT3*, *ST6GAL1* and *RFXAP* genes); only *FUT8* was
 125 associated with glycans from both proteins (Table 2, Supplementary Tables 1 and 2). Almost
 126 all identified genes encode key enzymes in protein glycosylation (*MGAT3*, *ST6GAL1*, *FUT6*,
 127 *FUT8*) or have been previously associated with transferrin and IgG glycan traits in GWAS
 128 analysis (*MSR1*, *FOXII*)^{17,18}. The exceptions are *TIRAP* and *RFXAP*, which have no previously
 129 known link to protein glycosylation. We successfully replicated (p-value < 3.2×10^{-4} for
 130 transferrin, p-value < 5.9×10^{-4} for IgG) associations of glycans with low-frequency and rare
 131 variants from 4 genes - *FUT6* and *TIRAP* with transferrin glycans, and *FUT8* and *MGAT3* for
 132 IgG glycans (Table 2) - as frequencies of variants in these genes are similar across the studied
 133 cohorts (Supplementary Table 3). While the associations of IgG glycans and variants from
 134 *FUT8* replicated, the association of transferrin glycans with variants from the same gene did
 135 not reach the significance threshold for replication (p-value in VIKING = 1.7×10^{-3}), likely
 136 because of the 7-fold decreased frequency of the rs2229678 variant in the VIKING (MAF =
 137 0.0056) compared to CROATIA-Korcula (MAF = 0.049) cohort (Supplementary Table 4).
 138 However, given the known biological role of *FUT8* in protein glycosylation as a
 139 fucosyltransferase (one of the enzymes involved in the synthesis of glycans), we believe this
 140 association to be real. Associations of rare variants from the CROATIA-Korcula cohort in the
 141 *MSR1* gene with transferrin glycosylation also did not formally replicate in the VIKING cohort
 142 (p-value = 8.6×10^{-4}) (Table 2). However, the cumulative allele count of rare variants in this
 143 gene is different between CROATIA-Korcula (MAC=46) and the VIKING cohort (MAC=38)
 144 (Supplementary Table 3), decreasing the power to replicate. We also detected a couple of

145 isolate-specific associations that are driven by variants increased in frequency compared to
146 publicly accessible biobanks and variant repositories. Namely, the rs750567016 variant in
147 *ST6GAL1* that affects IgG glycosylation is more than 300 times more common in ORCADES
148 (MAF = 3.3×10^{-3}) than in UK Biobank (MAF = 1.0×10^{-5}) or gnomAD (MAF = 9.0×10^{-6}) and
149 is absent from CROATIA-Korcula and VIKING cohorts. The rs115399307 variant in *FOXI1*,
150 associated with transferrin glycosylation, is seven times more common in VIKING (MAF =
151 2.1×10^{-2}) than in CROATIA-Korcula cohort (MAF = 2.7×10^{-3}), UK Biobank (MAF = 8.5×10^{-3})
152 and gnomAD (MAF = 7.1×10^{-3}) (Supplementary Table 4). While the role of sialyltransferase
153 *ST6GAL1* in IgG glycosylation is well described, the roles of the transcription factor *FOXI1*
154 and the regulatory factor X-associated protein *RFXAP* still need to be confirmed and
155 investigated.



156
157 **Figure 1. Miami plot summarising the results from exome-wide gene-based tests for**
158 **transferrin and IgG glycan traits.** Genomic positions of the genes, calculated as the mean
159 position of variants included in the reported mask, are labelled on the x-axis and the $-\log_{10}$ of
160 the p-value for each rare-variants aggregating test on the y-axis. For each gene-glycan
161 association, the lowest p-value across multiple masks, multiple variant aggregate tests and
162 cohorts was selected for plotting. The Bonferroni-corrected significance threshold for
163 transferrin glycan traits (horizontal red line in the top part of the plot) corresponds to 8.06×10^{-8} ,
164 while Bonferroni-corrected threshold for the IgG glycan traits (horizontal red line in the
165 bottom part of the plot) corresponds to 1.19×10^{-7} . Genes significantly associated with
166 transferrin/IgG glycan traits are indicated with a triangle and labelled, while genes not passing
167 the significance threshold are indicated with dots.

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

Lead glycan	Gene	MAF	Variants	N variants	Discovery cohort	Discovery P	Assoc. test	Discovery MAF	Discovery AC	Repl. cohort	Repl. P	Repl. MAF	Repl. AC	No. of glycans
Transferrin														
TfGP20	<i>FUT8</i>	<0.05	pLoF and deleterious (1/5)*	6	CROATIA-Korcula	6.29x10 ⁻²⁰	Burden	0.0111	124	VIKING	1.73x10 ⁻³	0.0042	8	3
TfGP32	<i>FUT6</i>	<0.05	pLoF and deleterious (1/5)*	5	CROATIA-Korcula	4.31x10 ⁻¹⁸	SKAT	0.0097	90	VIKING	1.56x10 ⁻¹⁴	0.0072	96	8
TfGP35	<i>MSR1</i>	<0.05	pLoF	3	CROATIA-Korcula	6.93x10 ⁻¹⁷	Burden	0.0083	46	VIKING	8.64x10 ⁻⁴	0.01	38	2
TfGP17	<i>TIRAP</i>	<0.05	pLoF and deleterious (1/5)*	3	VIKING	2.17x10 ⁻¹⁰	SKAT-O	0.0077	44	CROATIA-Korcula	8.12x10 ⁻⁹	0.0076	98	2
TfGP23	<i>FOX11</i>	<0.05	pLoF and missense	3	VIKING	1.37x10 ⁻⁸	SKAT-O	0.0074	42	CROATIA-Korcula	1.56x10 ⁻²	0.0014	8	1
IgG														
FG2S1/(FG2+FG2S1+FG2S2)	<i>ST6GAL1</i>	<0.01	pLoF and missense	2	ORCADES	9.82x10 ⁻²²	Burden	0.0019	15	-	-	-	-	9
Fn/(Bn+FBn)	<i>MGAT3</i>	<0.01	pLoF and deleterious (1/5)*	4	ORCADES	2.31x10 ⁻¹⁰	SMMAT-E	0.0021	33	CROATIA-Korcula	6.57x10 ⁻⁹	0.0012	29	17
FG1n total/G1n	<i>FUT8</i>	<0.05	pLoF and missense	7	CROATIA-Korcula	6.74x10 ⁻⁸	Burden	0.0072	177	ORCADES	3.04x10 ⁻⁶	0.0037	43	5

GP21	<i>RFXAP</i>	<0.01	pLoF and missense	2	ORCADES	1.04x10 ⁻⁷	SMMAT-E	0.0033	26	VIKING	6.29x10 ⁻²	0.0009	2	1
<p>Lead glycan - glycan trait reporting the strongest rare-variants association at the gene. Gene – gene for which rare variants are grouped. MAF - the highest allele frequency of variants aggregated for the mask. Variants - functional consequence of variants in the mask, aggregated in the given gene (pLoF – predicted loss of function: * denotes the number of algorithms predicting the missense variant to be deleterious). N variants - number of variants included in the mask. Discovery cohort - cohort reporting the lower p-value for the glycan-gene association. Discovery/Repl. P - p-value of the gene-based association with the lead glycan trait in the discovery/replication cohort. Assoc. test – gene-based association test for which p-value is reported. Discovery/repl. MAF - mean minor allele frequency of variants from the mask in the discovery/replication cohort. Discovery/repl. AC - sum of allele counts of variants from the mask in the discovery/replication cohort. Repl. Cohort - cohort reporting the higher p-value for the glycan-gene association. No. of glycans - number of other glycan traits associated with variants from the same mask. Results for transferrin glycome are reported at the top of the table, while results for IgG glycome are reported at the bottom of the table. Discovery significance threshold is 8.06x10⁻⁸ for transferrin and 1.19x10⁻⁷ for IgG glycans. Replication significance threshold is 3.23x10⁻⁴ for transferrin and 5.95x10⁻⁴ for IgG glycans.</p>														

169

170 **IgG glycans gene-based aggregation meta-analysis**

171 To further increase statistical power, we performed gene-based aggregation meta-analysis of
172 IgG glycan traits for ORCADES and VIKING cohorts. In addition to two genes already found
173 to be associated with IgG in the cohort-specific analysis (*MGAT3* and *ST6GALI*), the combined
174 analysis of VIKING and ORCADES cohort added *FUT6* to the list of genes whose rare variants
175 are significantly ($p\text{-value} < 1.19 \times 10^{-7}$) associated with IgG glycan traits (Supplementary Table
176 5). *FUT6* is another gene known to be involved in glycosylation^{17,20,22,23}, encoding a
177 glycosyltransferase enzyme that catalyses the transfer of fucose moieties to a growing glycan
178 chain.

179

180 **Genetic architecture of aggregated effects of rare-variants**

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

185 We first performed GWAS on imputed genotypes for each glycan trait and then repeated the
186 rare variant association tests incorporating the dosages of GWAS sentinel SNPs as additional
187 covariates. Genome-wide significant (transferrin $p\text{-value} < 1.61 \times 10^{-9}$, IgG $p\text{-value} < 2.38 \times 10^{-9}$)
188 associations reported in this study (Supplementary Table 6 for transferrin glycans and
189 Supplementary Table 7 for IgG glycans) have been described in further details in Landini *et al.*¹⁸
190 and Klarić *et al.*¹⁷. Overall, aggregated associations with variants from 3 out of 8 genes,
191 *FUT8*, *ST6GALI* and *MGAT3*, remained significant ($p\text{-value} < 8.06 \times 10^{-8}$ for transferrin and $p\text{-value} < 1.19 \times 10^{-7}$
192 for IgG) after conditioning on sentinel GWAS associations (Table 3). For one
193 gene, *RFXAP*, there were no significant associations in the GWAS analysis, while the
194 remaining four gene-based associations (*FUT6*, *MSRI*, *FOXII* and *TIRAP*) were explained by
195 sentinel GWAS variants. For two of these genes, *FUT6* and *FOXII*, the GWAS sentinel
196 variants are low frequency ($0.02 < \text{MAF} < 0.05$; Supplementary Table 8). More specifically, for
197 transferrin glycans, 14 out of the 16 glycome-gene aggregate pairs fail to reach genome-wide
198 significance ($p\text{-value} < 8.06 \times 10^{-8}$) after conditioning on GWAS sentinel SNPs (Supplementary
199 Table 9), meaning that a considerable part of the rare variant signal was dependent on variants
200 identifiable by GWAS. In contrast, for IgG glycans, 24 of the 32 glycome-gene aggregate pairs
201 remained significant ($p\text{-value} < 1.19 \times 10^{-7}$), even after adjusting for GWAS sentinel SNPs
202 (Supplementary Table 10).

203 Next, we performed single-point exome-wide association analysis (ExWAS) and repeated the
204 aggregated rare variant association tests while conditioning on the sentinel ExWAS
205 associations. In this way we tested whether the rare-variants associations with glycosylation
206 were driven by a single variant (i.e. showed an attenuated signal after conditioning on the
207 sentinel ExWAS variant) or were actually affected by multiple rare variants in concert (i.e.
208 associations remain significant after the conditioning). Two of the associations, between

209 variants in the *FUT8* gene and IgG glycans, and variants in the *MGAT3* gene and transferrin
210 glycans, remain significant after conditioning on the sentinel ExWAS variant (Table 3). Upon
211 closer inspection, for both of these genes, the sentinel ExWAS variant was a common variant
212 that is also an eQTL for the gene in blood (eQTLGen²⁸: rs35949016, *FUT8* eQTL, p-value =
213 6.5×10^{-159} ; rs6001566, *MGAT3* eQTL p-value = 3.9×10^{-230}) (Supplementary Table 8). Hence,
214 it appears that glycosylation is affected by common variants and independently by aggregates
215 of rare variants in these two genes. Indeed, by looking at the single-point effects of each rare
216 variant from the mask, we can see that multiple independent rare variants contribute to the
217 effect on glycosylation levels (Supplementary Table 11).

218 In summary, four of the identified associations, three with low-frequency variants from *FUT6*,
219 *MSRI* and *FOXII* and one with a common variant from the *TIRAP* gene, could have been
220 discovered using a GWAS of imputed genotype data. On the other hand, the rare variant
221 association at *ST6GALI* gene could only have been discovered using an ExWAS, as it is too
222 rare to be imputed well. Finally, associations with variants from two genes, *FUT8* and *MGAT3*,
223 are driven by multiple rare variants simultaneously contributing to glycosylation of IgG and
224 transferrin. Also the rare variant association at *RFXAP* gene could not have been discovered
225 by either GWAS or ExWAS as there were no significant single-point associations. However,
226 it is important to note that we could not replicate this association because the variants from its
227 mask are depleted in the other studied cohorts (Supplementary Table 4).

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

Lead glycan	Gene	MAF	Variants	Association test	cohort	Discovery P	GWAS adj p	ExWAS adj p
Transferrin								
TfGP20	<i>FUT8</i>	<0.05	pLoF and deleterious (1/5)*	Burden	CROATIA-Korcula	6.29x10 ⁻²⁰	2.75x10 ⁻¹²	2.70 x10 ⁻¹⁵
TfGP32	<i>FUT6</i>	<0.05	pLoF and deleterious (1/5)*	SKAT	CROATIA-Korcula	4.31x10 ⁻¹⁸	3.07x10 ⁻¹	2.94 x10 ⁻¹
TfGP35	<i>MSRI</i>	<0.05	pLoF	Burden	CROATIA-Korcula	6.93x10 ⁻¹⁷	6.62x10 ⁻⁷	3.70 x10 ⁻³
TfGP17	<i>TIRAP</i>	<0.05	pLoF and deleterious (1/5)*	SKAT-O	VIKING	2.17x10 ⁻¹⁰	8.61x10 ⁻¹	1.38 x10 ⁻¹
TfGP23	<i>FOXII</i>	<0.05	pLoF and missense	SKAT-O	VIKING	1.37x10 ⁻⁰⁸	6.85x10 ⁻¹	6.38 x10 ⁻¹
IgG								
FG2S1/(FG2+F G2S1+FG2S2)	<i>ST6GAL1</i>	<0.01	pLoF and missense	Burden	ORCADES	9.82x10 ⁻²²	6.99x10 ⁻¹⁹	1.44 x10 ⁻²
Fn/(Bn+FBn)	<i>MGAT3</i>	<0.01	pLoF and deleterious (1/5)*	SMMAT-E	ORCADES	2.31x10 ⁻¹⁰	5.47x10 ⁻¹⁰	5.68 x10 ⁻¹⁰
FG1n total/G1n	<i>FUT8</i>	<0.05	pLoF and missense	Burden	CROATIA-Korcula	6.74x10 ⁻⁸	2.31x10 ⁻⁶	8.4x10 ⁻⁶
GP21	<i>RFXAP</i>	<0.01	pLoF and missense	SMMAT-E	ORCADES	1.04x10 ⁻⁷	1.04x10 ^{-7***}	1.04x10 ^{-7***}
Lead glycan - glycan trait reporting the strongest rare-variants association at the gene. Gene – gene for which rare variants are grouped. MAF - the highest allele frequency of variants aggregated for the mask. Variants - functional consequence of variants in the mask, aggregated in the given gene (pLoF – predicted loss of function: * denotes the number of algorithms predicting the missense variant to be deleterious). Association test – gene-based association test for which p-value is reported. Cohort - cohort reporting the lower p-value for the glycan-gene association. Discovery P - p-value of the gene-based association test with the lead glycan trait in the cohort. GWAS adj p – p-value of association test when conditioning on the significant variants from the GWAS analysis; ** no significant GWAS variants were found.. ExWAS adj p – p-value of association test when conditioning on								

the significant variants from the ExWAS analysis; ** no significant ExWAS variants were found. Results for transferri glycome are reported at the top of the table, while results for IgG glycome are reported at the bottom of the table. P-value significance threshold is 8.06×10^{-8} for transferrin and 1.19×10^{-7} for IgG glycans.

233 **Links to health-related traits**

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

240 We performed exome-wide “gene-level PheWAS” with 116 quantitative health-related traits
241 measured in the ORCADES, CROATIA-Korcula and VIKING cohorts, limited to the genes
242 containing pLoF and missense variants that were associated with transferrin or IgG glycome
243 variation (Table 2). When possible, we sought to perform the analysis in the same cohort where
244 the glycan-gene association was discovered. The only significant ($p\text{-value} < 5.4 \times 10^{-5}$)
245 association was with transferrin glycosylation-associated rare variants from the *MSRI* gene and
246 blood levels of HbA1c in the VIKING cohort (Supplementary Table 12). However, the
247 association with HbA1c levels is not significant in CROATIA-Korcula, the cohort where we
248 discovered the connection between *MSRI* and transferrin glycosylation, and it also does not
249 replicate in ORCADES, suggesting that it might be a false positive association. We next
250 checked whether any of the glycome-associated genes were significantly associated with
251 health-related traits in UK Biobank. We used two repositories of aggregated rare-variants
252 associations: Genebass²⁹ and the AstraZeneca PheWAS portal³⁰. Missense variants from the
253 *MSRI* gene were significantly associated with insulin-like growth factor 1 levels (*IGF1*) in
254 both Genebass (SKAT-O $p\text{-value} = 4.6 \times 10^{-10}$) and the PheWAS portal ($p\text{-value} = 1.6 \times 10^{-24}$, for
255 the “ptv5pct” collapsing model).

256 Discussion

257 Statistical power to detect associations with rare genetic variants can be increased by
258 aggregating the association signals across multiple rare variants in a gene³¹, or by using
259 genetically isolated populations where, due to genetic drift, some variants are increased in
260 frequency compared to a general population³². Further, intermediate phenotypes, more
261 proximal to the genes and consequently more strongly influenced by them, can be used as
262 “proxies” of complex diseases to boost power. Glycosylation, one of the most common post-
263 translational modifications, is one such intermediate phenotype and has been implicated in
264 many diseases^{10,13,14}. Here, we utilised the power of genetic isolates, aggregation of multiple
265 rare variants and intermediate phenotypes to study the effect of rare variants on glycosylation
266 of two proteins and their effect on disease.

267 We performed multiple gene-based aggregation tests to assess associations with transferrin (N
268 = 1907) and IgG (N = 4912) glycan traits in three isolated cohorts of European descent, testing
269 rare (MAF<5%) pLoF and missense variants from whole exome sequencing. We found rare
270 variants from 8 genes contributing to glycan levels of either IgG or transferrin. As previously
271 observed in GWAS using imputed genotypes, transferrin and IgG glycans showed mostly
272 protein-specific gene-based associations¹⁸, including genes encoding known glycosylation
273 enzymes (transferrin - *TIRAP*, a gene in the proximity of *ST3GAL4*; IgG - *ST6GAL1* and
274 *MGAT3*), transcription factors (transferrin - *FOXII*), as well as other genes (transferrin - *MSRI*;
275 IgG - *RFXAP*). On the other hand, rare variants in *FUT8* and *FUT6*, genes encoding
276 fucosyltransferase enzymes adding core and antennary fucose structures to the synthesised
277 glycan, were associated with glycosylation of both proteins. Previously we showed that, while
278 glycosylation of both transferrin and IgG proteins is associated with genes encoding *FUT6* and
279 *FUT8* fucosylation enzymes, these associations are driven by independent, protein-specific
280 variants mapped to the regulatory region of the two genes¹⁸. Accordingly, here we identified
281 rare variants in the exonic portions of *FUT8* and *FUT6*, acting independently or in concert with
282 GWAS-identifiable variants.

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

296

297 Except for *RFXAP* and *TIRAP*, all of 8 identified genes have already been associated with IgG
298 and transferrin glycosylation in previous GWAS studies^{17,18,20,22,23}. The novel gene *TIRAP* is
299 located in close proximity to *ST3GAL4*, another glycosyltransferase-coding gene known to be
300 associated with transferrin glycosylation. *TIRAP* has a function in the innate immune system,
301 where it is involved in cytokine secretion and the inflammatory response^{33,34}. The lead rare
302 variant in the mask, rs8177399 (Supplementary Table 3), in addition to being an expression
303 QTL (eQTL) for *TIRAP* and several other genes, is also a splicing QTL (sQTL) for *ST3GAL4*
304 in whole blood (GTE_x³⁵, p-value = 1.9×10^{-8}). The regulatory factor X-associated protein
305 encoded by *RFXAP* gene, whose variants are associated with IgG glycans, is part of a
306 multimeric complex, called the RFX DNA-binding complex, that binds to certain major
307 histocompatibility (MHC) class II gene promoters and activates their transcription. MHC-II
308 molecules are transmembrane proteins, found on the surface of professional antigen-presenting
309 cells (including B cells)³⁶, which have a central role in development and control of the immune
310 response. While the mechanism of *TIRAP*'s influence on the glycome could be through
311 controlling the splicing of the known glycosyltransferase enzyme *ST3GAL4*, the precise role
312 of *RFXAP* in protein glycosylation still needs to be established.

313 Changes in the glycosylation patterns are often observed in a wide range of pathological states,
314 such as cancer, inflammatory, autoimmune, neurodegenerative and cardiovascular diseases^{37–}
315 ⁴⁰. We thus assessed the potential involvement of glycome-associated genes in health, by
316 performing, in the same three cohorts, gene-based association tests of 116 quantitative health-
317 related traits, limited to genes whose rare variants we found associated with the protein
318 glycomes. However, given the likely small effect-size of variants on complex diseases, we did
319 not find any significant associations. On the other hand, using publicly available repositories
320 of gene-based associations in the UK Biobank data, we found that rare missense variants from
321 *MSRI* (associated with transferrin glycosylation) were also associated with blood levels of
322 insulin-like growth factor 1 (IGF1). IGF1 is a hormone with significant structural and
323 functional similarities to insulin: lower levels of IGF1 are associated with higher risk of Type
324 1 and 2 diabetes mellitus^{41,42}. Recently, a rare deleterious missense variant in *IGF1* receptor
325 (*IGF1R*) was found to be significantly associated with Type 2 diabetes in UK Biobank, further
326 corroborating the link between IGF1 and diabetes⁴³. In addition, genetic variants in *MSRI* have
327 been previously associated with plasma levels of the galectin-3-binding protein⁴⁴. Similarly to
328 IGF1, galectin-3 has been identified as a marker and a pathogenic factor in type 2 diabetes,
329 with the serum protein levels increased in type 2 diabetes patients^{47–51}. An important part of
330 iron delivery depends on recycling transferrin via clathrin-mediated endocytosis.
331 Interestingly, binding of galectin-3 to transferrin can affect its intracellular trafficking^{45,46}.
332 Based on the glycosylation profile, galectin-3 was found bound only to a select, minor fraction
333 (~5%) of transferrin, while interestingly none or little was bound to IgG⁴⁶. Overall, variants
334 from the *MSRI* gene seem to have a pleiotropic effect on transferrin glycosylation, galectin-3
335 and IGF1. In turn, both galectin-3 and IGF1 are associated with type 2 diabetes. The potential
336 role of glycosylation of transferrin in these processes still needs to be established.

337

338 In conclusion, we identified rare pLoF and missense variants associated with transferrin and
339 IgG N-glycome, in both known and not previously reported genes (*TIRAP*, *RFXAP*). By
340 utilising the power of genetic isolates and aggregated effects of rare variants, we discovered
341 biologically relevant associations with a 300-fold up-drifted variant in the ORCADES cohort
342 (in the sialyltransferase gene, *ST6GALI*, affecting levels of sialylation of IgG) and associations
343 independent of single-point GWAS and ExWAS analyses (in glycosyltransferase genes *FUT8*
344 and *MGAT3*). Interestingly, many of glycan traits are influenced both by common and rare
345 variants, revealing a complex genetic architecture of these intermediate phenotypes. While we
346 did not find any robust links between glycome-associated genes and diseases in studied cohorts,
347 we discover a potential link between transferrin glycosylation, galectin-3, IGF1 and diabetes.
348 The exact mechanism behind these connections still needs to be confirmed and further
349 explored. This study shows that, utilising the power of genetic isolates, gene-based aggregation
350 tests and intermediate phenotypes such as glycosylation, rare variant associations are detectable
351 even in relatively small sample sizes (low thousands). However, larger cohorts would be
352 required to identify the contribution of rare variants to multifactorial, complex diseases.

353 **Methods**

354

355 **Genotypic data**

356 Exome sequencing

357 The “Goldilocks” exome sequence data for ORCADES, CROATIA-Korcula and VIKING
358 cohorts was prepared at the Regeneron Genetics Center, following the protocol detailed in Van
359 Hout *et al.*² for the UK Biobank whole-exome sequencing project. In summary, the multiplexed
360 samples were sequenced on the Illumina NovaSeq 6000 platform using S2 flow cells. The raw
361 sequencing data was processed by automated analysis using the DNAnexus platform⁵², where
362 files were converted to FASTQ format, and then aligned to GRCh38 genome reference using
363 the BWA-mem⁵³. Duplicated reads were identified and flagged by the Picard tool⁵⁴. Genotypes
364 for each individual sample were called using the WeCall variant caller⁵⁵. During quality
365 control, samples genetically identified as duplicates, showing disagreement between
366 genetically determined and reported sex, high rates of heterozygosity or contamination, low
367 sequence coverage (less than 80% of targeted bases achieving 20X coverage) or discordant
368 with genotyping chip were excluded. The number of samples removed after quality control are
369 listed in Supplementary Table 13 for each cohort. Finally, the “Goldilocks” dataset was
370 generated by (i) filtering out genotypes with read depth lower than 7 reads, (ii) keeping variants
371 having at least one heterozygous variant genotype with allele balance ratio greater than or equal
372 to 15% ($AB \geq 0.15$) or at least one homozygous variant genotype, and (iii) filtering out variants
373 with more than 10% of missingness and HWE $p < 10^{-6}$. Overall, a total of 2,090 ORCADES
374 (820 male and 1,270 female), 2,872 CROATIA-Korcula (1,065 male and 1,807 female) and
375 2,108 VIKING (843 male and 1,265 female) participants passed all exome sequence and
376 genotype quality control thresholds. A pVCF file containing all samples passing quality control
377 was then created using the GLnexus joint genotyping tool.⁵⁶

378

379 Variant annotation

380 Exome sequencing variants were annotated as described in Van Hout, *et al.*² In brief, each
381 variant was labelled with the most severe consequence across all protein-coding transcripts,
382 implemented using SnpEff⁵⁷. Gene regions were defined according to Ensembl release 85.
383 Variants annotated as stop gained, start lost, splice donor, splice acceptor, stop lost and
384 frameshift were considered as predicted LOF variants. The deleteriousness of missense variants
385 was assessed using the following algorithms and classifications (based on dbNSFP 3.2): (1)
386 SIFT: “D” (Damaging), (2) Polyphen2_HDIV: “D” (Damaging) or “P” (Possibly damaging),
387 (3) Polyphen2_HVAR: “D” (Damaging) or “P” (Possibly damaging), (4) LRT⁵⁸: “D”
388 (Deleterious) and (5) MutationTaster⁵⁹: “A” (Disease causing automatic) or “D” (Disease
389 causing). Missense variants were considered “likely deleterious” if predicted as deleterious by

390 all five algorithms, “possibly deleterious” if predicted as deleterious by at least one of the
391 algorithms and “likely benign” if not predicted as deleterious by any of the algorithms.

392

393 Generation of gene burden masks

394 For each gene, we grouped the variants in the gene in four categories (masks), based on severity
395 of their functional consequence. Mask 1 included only predicted loss-of-function (pLoFs)
396 variants, mask 2 consisted of pLoF variants and all missense variants, and masks 3 and 4
397 contained pLoF and predicted deleterious missense variants (“possibly deleterious” and “likely
398 deleterious” for mask 3 and mask 4, respectively). We considered two separate variations of
399 each mask based on the frequency of the minor allele of the variants that were screened in that
400 group: $MAF \leq 5\%$ and $MAF \leq 1\%$. Overall, up to 8 burden tests were performed for each gene
401 (Supplementary Table 14). Consequently, the masks are not independent - certain masks will
402 include the variants listed in a different mask and additional, less severe or more frequent
403 variants.

404

405 **Phenotypic data**

406 Transferrin and IgG N-glycome quantification

407 Transferrin and total IgG N-glycome quantification for ORCADES, VIKING and CROATIA-
408 Korcula samples was performed at Genos Glycobiology Laboratory, following the protocol
409 described in Trbojević-Akmačić *et al.*⁶⁰ for transferrin, in Pučić *et al.*⁶¹ for IgG in ORCADES
410 cohort and batch 1 of CROATIA-Korcula cohort, in Trbojević-Akmačić *et al.*⁶² for IgG in
411 VIKING cohort and batch 2 of CROATIA-Korcula cohort. In summary, proteins of interest
412 were first isolated from blood plasma (IgG depleted blood plasma, in the case of transferrin)
413 using affinity chromatography binding to anti-transferrin antibodies plates for transferrin and
414 protein G plates for IgG. The protein isolation step was followed by release and labelling of N-
415 glycans and clean-up procedure.. IgG N-glycans have been released from total IgG (all
416 subclasses). N-glycans were then separated and quantified by hydrophilic interaction ultra-
417 high-performance liquid chromatography (HILIC-UHPLC). As a result, transferrin and total
418 IgG samples were separated into 35 (transferrin: TfGP1 – TfGP35) and 24 (IgG: GP1 – GP24)
419 chromatographic peaks. It is worth noting that there is no correspondence structure-wise
420 between transferrin TfGP and IgG GP traits labelled with the same number.

421

422 Normalisation and batch correction

423 Prior to genetic analysis, raw N-glycan UHPLC data was normalised and batch corrected to
424 reduce the experimental variation in measurements. Total area normalisation was performed
425 by dividing the area of each chromatographic peak (35 for transferrin, 24 for IgG) by the total

426 area of the corresponding chromatogram. Due to the multiplicative nature of measurement error
427 and right-skewness of glycan data, normalised glycan measurements were log10-transformed.
428 Batch correction was then performed using the empirical Bayes approach implemented in the
429 “ComBat” function of the “sva” R package⁶³, modelling the technical source of variation (96-
430 well plate number) as batch covariate. Batch corrected measurements were then exponentiated
431 back to the original scale. Prior to further analysis, each glycan trait was rank transformed to
432 normal distribution using the “rntransform” function from the “GenABEL” R package⁶⁴.

433

434 Derived glycan traits

435 IgG derived traits analysed included those defined by Huffman *et al.*⁶⁵, and were calculated
436 using the glycanr R package. In addition, new derived traits were calculated for both transferrin
437 and IgG, representing the overall presence of a certain sugar structure on the totality of
438 transferrin/IgG N-glycan traits measured (e.g. percentage of fucosylation). These newly
439 generated traits are expected to give a direct insight in the biological pathway involved in the
440 addition of the sugar moiety to glycan structures. Exact formulas used for defining transferrin
441 and IgG newly derived traits can be found in Supplementary Tables 15 and 16 respectively.

442

443 Health-related quantitative traits

444 To evaluate the potential effect of rare variants affecting glycome on health-related phenotypes,
445 in the same cohorts we collected 148 health-related, quantitative traits (e.g. anthropological
446 measurements, blood levels of proteins, metabolites and biomarkers). Excluding traits with
447 fewer than 800 samples, a total of 116 traits were considered for analysis (75 traits for
448 ORCADES, 79 for VIKING and 47 for CROATIA-Korcula cohort). Each health-related trait
449 was rank transformed to normal distribution using the “rntransform” function from the
450 “GenABEL” R package⁶⁴, followed by applying the rare-variants association pipeline
451 described below.

452

453 **Gene-based aggregation analysis**

454 We performed variant Set Mixed Model Association Tests (SMMAT)⁶⁶ on rank-transformed
455 glycan traits, fitting a GLMM adjusting for age, sex, sampling batch in the case of CROATIA-
456 Korcula IgG glycan traits, and familial or cryptic relatedness by kinship matrix. The kinship
457 matrix was estimated from the genotyped data using the ‘ibs’ function from GenABEL R
458 package⁶⁴. The SMMAT framework includes 4 variant aggregate tests: burden test, sequence
459 kernel association test (SKAT), SKAT-O and SMMAT-E, a hybrid test combining the burden
460 test and SKAT. The 4 variant aggregate tests were performed on 8 different pools of genetic
461 variants, called “masks”, described above (Supplementary Table 14).

462 Discovery significance threshold was Bonferroni corrected for the approximate number of
463 genes in the human genome, 20,000, and the number of independent glycan traits, 21 for IgG
464 and 31 for transferrin ($0.05/20000/31 = 8.06 \times 10^{-8}$ for transferrin, $0.05/20000/21 = 1.19 \times 10^{-7}$ for
465 IgG). The number of independent glycan traits was estimated as the number of principal
466 components that jointly explained 99% of the total variance of transferrin/IgG glycan traits in
467 each cohort (Supplementary Tables 17 and 18). PCA was calculated on rank-transformed
468 glycan traits, separately for each cohort, using the “prcomp” function from “factoextra” R
469 package⁶⁷. A gene association was considered significant if it passed the above-described
470 Bonferroni corrected significance threshold in at least one of the 4 performed variant aggregate
471 tests and if the cumulative allele count of the variants included in the gene was equal or higher
472 than 10. Replication significance threshold was defined as $P = 0.05$ divided by the number of
473 genes and independent glycans to be replicated. For IgG glycans, this threshold was $P =$
474 5.95×10^{-4} ($P = 0.05/4$ genes/21 glycans) and for transferrin glycans, this threshold was $P =$
475 3.23×10^{-4} ($P = 0.05/5$ genes/31 glycans).

476 A similar analysis plan was applied to the health-related phenotypes analysed. Variant Set
477 Mixed Model Association Tests (SMMAT)⁶⁶ was performed on rank-transformed traits, fitting
478 a GLMM adjusting for age, sex, first 20 ancestral principal components (PCs), batch covariates
479 when available (e.g. season, time of the day and batch/subcohort) and familial or cryptic
480 relatedness.

481

482 **IgG glycome gene-based aggregation meta-analysis**

483 Gene-based aggregation analysis of IgG glycan traits for ORCADES and VIKING cohorts was
484 repeated following the same approach as previously described, except for the restriction that
485 masks included only variants present in both cohorts. Since IgG GP3 was not quantified in
486 ORCADES cohort, this glycan was excluded from the meta-analysis, bringing the total number
487 of IgG glycan traits considered to 93. We then used the “SMMAT.meta” function of
488 “SMMAT” R package⁶⁶ to meta-analyse, for each trait, the two studies. To identify significant
489 results we filter results by the previously described Bonferroni-corrected significance threshold
490 of 1.19×10^{-7} and by the cumulative allele count of variants included in the gene equal or higher
491 than 10.

492

493 **Genome-wide association analysis**

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

500 glycan measurements were not available in ORCADES. Rank-transformed glycan traits were
501 adjusted for age and sex, as fixed effects, and relatedness (estimated as the kinship matrix
502 calculated from genotyped data) as random effect in a linear mixed model, calculated using the
503 “polygenic” function from the “GenABEL” R package⁶⁴. Since IgG N-glycan traits for the
504 CROATIA-Korcula cohort were measured at two separate occasions, the two were considered
505 as separate cohorts. Therefore, for CROATIA-Korcula, rank transformation was performed
506 separately in each subcohort. Samples were then merged together for GWAS, but adding batch
507 (subcohort number - 1 or 2) as fixed effect covariate. Residuals of covariate and relatedness
508 correction were tested for association with Haplotype Reference Consortium (HRC) r1.1-
509 imputed SNP dosages using the RegScan v. 0.5 software, applying an additive genetic model
510 of association.

511 The genomic control inflation factor (λ_{GC}) was calculated for each glycan and health-related
512 trait. The mean genomic control inflation factor (λ_{GC}) for IgG glycan traits was 1.002 (0.982-
513 1.026) in ORCADES, 1 in CROATIA-Korcula (0.971-1.031) and 0.993 in VIKING cohort
514 (0.972-1.017) cohort; for transferrin glycan traits λ_{GC} was 1.002 in CROATIA-Korcula (0.982-
515 1.026) and 0.998 in VIKING (0.974-1.021) cohort. Overall, the confounding effects of the
516 family structure were correctly accounted for in our analyses.

517

518 **Identification of rare variant associations independent of GWAS and ExWAS signals**

519 To ensure that the rare variant associations identified were independent of associations with
520 variants discoverable by a GWAS or single-point exome-wide (ExWAS) analysis, we repeated
521 the aggregate analysis while conditioning on the sentinel SNPs from the single-variant genome-
522 wide or exome-wide analysis. First, we performed GWAS of glycan traits using the same
523 individuals as in the analysis of the exome-sequencing data, but using as genotypes SNP
524 dosages imputed from the HRC imputation panel, as described above. For each glycan trait we
525 defined the sentinel SNPs as the variants having the lowest significant p-value ($p < 5 \times 10^{-8}$) in
526 a 1Mb window, and MAF > 1%. Then we also performed the exome-wide association analysis
527 (ExWAS), following exactly the same protocol, but with exome sequencing data used for
528 genotypes. We then re-run variant aggregate analysis as previously described, but with
529 adjusting the glycan traits for the genotype of the sentinel SNPs from the GWAS/ExWAS
530 significant loci, in addition to the other covariates listed above. The statistical significance level
531 was determined in the same way as outlined in the main analysis above.

532

533 **Replication of glycome rare associations in different cohorts and associations with health- 534 related traits**

535 To investigate whether glycome rare-variants associations were cohort specific, each
536 significant gene-glycan trait pair from the cohort-level discovery analysis was tested for
537 associations in the remaining cohorts. The p-value threshold for replication was set to $3.23 \times 10^{-}$

538 ⁴ for transferrin (0.05/31/5) and 5.95×10^{-4} for IgG (0.05/21/4) glycans, correcting for the
539 number of independent glycan traits (i.e. 31 for transferrin and 21 for IgG) and the number of
540 discovered glycome-gene pairs (i.e. 5 for transferrin and 4 for IgG in gene-based aggregation
541 analysis).

542 To investigate whether the glycome associated rare-variants may also affect health-related
543 phenotypes, we tested for association each glycome-associated gene and 116 health-related
544 traits. The significance threshold was set to 5.43×10^{-5} , correcting for the number of health-
545 related traits (116), and the number of discovered glycome-gene pairs and number of glycome-
546 associated genes (8).

547

548 **Code availability**

549 We used publicly available software tools for all analyses. These software tools are listed in
550 the main text and in the Methods.

551

552 **Data availability**

553 There is neither Research Ethics Committee approval, nor consent from individual participants,
554 to permit open release of the individual level research data underlying this study. The datasets
555 generated and analysed during the current study are therefore not publicly available. Instead,
556 the research data and/or DNA samples are available from accessQTL@ed.ac.uk on reasonable
557 request, following approval by the QTL Data Access Committee and in line with the consent
558 given by participants. Each approved project is subject to a data or materials transfer agreement
559 (D/MTA) or commercial contract. The UK Biobank genotypic data used in this study were
560 approved under application 19655, 48511 and 19655 are available to qualified researchers via
561 the UK Biobank data access process. The expression data used for the analyses described in
562 this manuscript were obtained from the GTEx Portal on 25/11/2022. Genebass
563 (<https://app.genebass.org/>) and AstraZeneca PheWAS portal (<https://azphewas.com/>) were
564 accessed on 06/12/2022.

565 References

- 566 1. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation
567 in regulatory DNA. *Science* (80-.). **337**, 1190–1195 (2012).
- 568 2. Van Hout, C. V *et al.* Exome sequencing and characterization of 49,960 individuals in
569 the UK Biobank. *Nature* **586**, 749–756 (2020).
- 570 3. Bomba, L. *et al.* Whole-exome sequencing identifies rare genetic variants associated
571 with human plasma metabolites. *Am. J. Hum. Genet.* (2022).
572 doi:10.1016/J.AJHG.2022.04.009
- 573 4. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440
574 controls. *Nature* **570**, 71–76 (2019).
- 575 5. Jurgens, S. J. *et al.* Analysis of rare genetic variation underlying cardiometabolic
576 diseases and traits among 200,000 individuals in the UK Biobank. *Nat. Genet.* 2022
577 **54**, 240–250 (2022).
- 578 6. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies
579 for association studies involving rare variants. *Nat Rev Genet* **11**, 773–785 (2010).
- 580 7. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association
581 studies. doi:10.1073/pnas.1322563111
- 582 8. Vanhooren, V. *et al.* Serum N-glycan profile shift during human ageing. *Exp Gerontol*
583 **45**, 738–743 (2010).
- 584 9. Vilaj, M., Gudelj, I., Trbojević-Akmačić, I., Lauc, G. & Pezer, M. IgG Glycans as a
585 Biomarker of Biological Age. in *Biomarkers of Human Aging* 81–99 doi:10.1007/978-
586 3-030-24970-0_7
- 587 10. Ząbczyńska, M., Link-Lenczowski, P. & Pocheć, E. Glycosylation in Autoimmune
588 Diseases. in *The Role of Glycosylation in Health and Disease* (eds. Lauc, G. &
589 Trbojević-Akmačić, I.) 205–218 (Springer International Publishing, 2021).
590 doi:10.1007/978-3-030-70115-4_10
- 591 11. Rudman, N., Gornik, O. & Lauc, G. Altered N-glycosylation profiles as potential
592 biomarkers and drug targets in diabetes. *FEBS Letters* **593**, 1598–1615 (2019).
- 593 12. Gudelj, I. & Lauc, G. Protein N-Glycosylation in Cardiovascular Diseases and Related
594 Risk Factors. *Curr. Cardiovasc. Risk Rep.* **12**, (2018).
- 595 13. Rebelo, A. L., Chevalier, M. T., Russo, L. & Pandit, A. Role and therapeutic
596 implications of protein glycosylation in neuroinflammation. *Trends Mol Med* **28**, 270–
597 289 (2022).
- 598 14. Costa, A. F., Campos, D., Reis, C. A. & Gomes, C. Targeting Glycosylation: A New
599 Road for Cancer Drug Discovery. *Trends Cancer* **6**, 757–766 (2020).
- 600 15. Peng, W. *et al.* Clinical application of quantitative glycomics. *Expert Rev. Proteomics*
601 **15**, 1007–1031 (2018).
- 602 16. Huffman, J. E. *et al.* Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated
603 with variation within the human plasma N-glycome of 3533 European adults. *Hum.*
604 *Mol. Genet.* **20**, 5000–5011 (2011).
- 605 17. Klarić, L. *et al.* Glycosylation of immunoglobulin G is regulated by a large network of
606 genes pleiotropic with inflammatory diseases. *Sci. Adv.* **6**, eaax0301 (2020).
- 607 18. Landini, A. *et al.* Genetic regulation of post-translational modification of two distinct
608 proteins. *Nat. Commun.* 2022 **131** **13**, 1–13 (2022).
- 609 19. Lauc, G. *et al.* Genomics Meets Glycomics—The First GWAS Study of Human N-
610 Glycome Identifies HNF1 α as a Master Regulator of Plasma Protein Fucosylation.
611 *PLoS Genet.* **6**, e1001256 (2010).
- 612 20. Lauc, G. *et al.* Loci Associated with N-Glycosylation of Human Immunoglobulin G
613 Show Pleiotropy with Autoimmune Diseases and Haematological Cancers. *PLoS*

- 614 *Genet.* **9**, (2013).
- 615 21. Sharapov, S. Z. *et al.* Defining the genetic control of human blood plasma N-glycome
616 using genome-wide association study. *Hum. Mol. Genet.* **28**, 2062–2077 (2019).
- 617 22. Shen, X. *et al.* Multivariate discovery and replication of five novel loci associated with
618 Immunoglobulin G N-glycosylation. *Nat. Commun.* **8**, 447 (2017).
- 619 23. Wahl, A. *et al.* Genome-wide association study on immunoglobulin G glycosylation
620 patterns. *Front. Immunol.* **9**, 277 (2018).
- 621 24. Bondt, A. *et al.* Immunoglobulin G (IgG) Fab glycosylation analysis using a new mass
622 spectrometric high-throughput profiling method reveals pregnancy-associated changes.
623 *Mol. Cell. proteomics* **13**, 3029–3039 (2014).
- 624 25. Wuhler, M. *et al.* Glycosylation profiling of immunoglobulin G (IgG) subclasses from
625 human serum. *Proteomics* **7**, 4070–4081 (2007).
- 626 26. Karlsson, I., Ndreu, L., Quaranta, A. & Thorsén, G. Glycosylation patterns of selected
627 proteins in individual serum and cerebrospinal fluid samples. *J. Pharm. Biomed. Anal.*
628 **145**, 431–439 (2017).
- 629 27. Spik, G. *et al.* Studies on glycoconjugates. LXIV. Complete structure of two
630 carbohydrate units of human serotransferrin. *FEBS Lett.* **50**, 296–299 (1975).
- 631 28. Vösa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic
632 loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **2021** 539
633 **53**, 1300–1310 (2021).
- 634 29. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of
635 thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**,
- 636 30. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank
637 exomes. *Nature* **597**, 527–532 (2021).
- 638 31. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and
639 applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
- 640 32. Zuk, O. *et al.* Searching for missing heritability : Designing rare variant association
641 studies. (2014). doi:10.1073/pnas.1322563111
- 642 33. Fitzgerald, K. A. *et al.* Mal (MyD88-adaptor-like) is required for Toll-like receptor-4
643 signal transduction. *Nature* **413**, 78–83 (2001).
- 644 34. Horng, T., Barton, G. M. & Medzhitov, R. TIRAP: an adapter molecule in the Toll
645 signaling pathway. *Nat Immunol* **2**, 835–841 (2001).
- 646 35. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**,
647 204–213 (2017).
- 648 36. Jones, E. Y., Fugger, L., Strominger, J. L. & Siebold, C. MHC class II proteins and
649 disease: a structural perspective. *Nat Rev Immunol* **6**, 271–282 (2006).
- 650 37. Juszcak, A. *et al.* Plasma fucosylated glycans and C-reactive protein as biomarkers of
651 HNF1A-MODY in young adult-onset nonautoimmune diabetes. *Diabetes Care* **42**,
652 17–26 (2019).
- 653 38. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature*
654 **591**, 92–98 (2021).
- 655 39. Verhelst, X. *et al.* Protein Glycosylation as a Diagnostic and Prognostic Marker of
656 Chronic Inflammatory Gastrointestinal and Liver Diseases. *Gastroenterology* **158**, 95–
657 110 (2020).
- 658 40. Wittenbecher, C. *et al.* Plasma N-Glycans as Emerging Biomarkers of
659 Cardiometabolic Risk: A Prospective Investigation in the EPIC-Potsdam Cohort Study.
660 *Diabetes Care* **43**, 661–668 (2020).
- 661 41. Meyer, N. M. T. *et al.* Low IGF1 and high IGFBP1 predict diabetes onset in
662 prediabetic patients. *Eur J Endocrinol* **187**, 555–565 (2022).
- 663 42. Segev, Y. *et al.* Systemic and renal growth hormone-IGF1 axis involvement in a

- 664 mouse model of type 2 diabetes. *Diabetologia* **50**, 1327–1334 (2007).
- 665 43. Gardner, E. J. *et al.* Damaging missense variants in IGF1R implicate a role for IGF-1
666 resistance in the etiology of type 2 diabetes. *Cell Genomics*
667 doi:10.1016/j.xgen.2022.100208
- 668 44. Pietzner, M. *et al.* Genetic architecture of host proteins interacting with SARS-CoV-2.
669 *bioRxiv Prepr. Serv. Biol.* (2020). doi:10.1101/2020.07.01.182709
- 670 45. Carlsson, M. C., Bengtson, P., Cucak, H. & Leffler, H. Galectin-3 guides intracellular
671 trafficking of some human serotransferrin glycoforms. *J Biol Chem* **288**, 28398–28408
672 (2013).
- 673 46. Cederfur, C. *et al.* Different affinity of galectins for human serum glycoproteins:
674 galectin-3 binds many protease inhibitors and acute phase proteins. *Glycobiology* **18**,
675 384–394 (2008).
- 676 47. Atalar, M. N. *et al.* Assessment of serum galectin-3, methylated arginine and Hs-CRP
677 levels in type 2 diabetes and prediabetes. *Life Sci* **231**, 116577 (2019).
- 678 48. Lin, D. *et al.* Galectin-3/adiponectin as a new biological indicator for assessing the risk
679 of type 2 diabetes: a cross-sectional study in a community population. *Aging (Albany*
680 *NY)* **13**, 15433–15443 (2021).
- 681 49. Ohkura, T. *et al.* Low serum galectin-3 concentrations are associated with insulin
682 resistance in patients with type 2 diabetes mellitus. *Diabetol Metab Syndr* **6**, 106
683 (2014).
- 684 50. Vora, A., de Lemos, J. A., Ayers, C., Grodin, J. L. & Lingvay, I. Association of
685 Galectin-3 With Diabetes Mellitus in the Dallas Heart Study. *J Clin Endocrinol Metab*
686 **104**, 4449–4458 (2019).
- 687 51. Weigert, J. *et al.* Serum galectin-3 is elevated in obesity and negatively correlates with
688 glycosylated hemoglobin in type 2 diabetes. *J Clin Endocrinol Metab* **95**, 1404–1411
689 (2010).
- 690 52. Reid, J. G. *et al.* Launching genomics into the cloud: deployment of Mercury, a next
691 generation sequence analysis pipeline. *BMC Bioinformatics* **15**, 30 (2014).
- 692 53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
693 transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 694 54. Institute, B. Picard Tools. (2018).
- 695 55. PLC, G. weCall. (2018).
- 696 56. Lin, M. F. *et al.* GLnexus: joint variant calling for large cohort sequencing. *bioRxiv*
697 (2018). doi:10.1101/343970
- 698 57. Cingolani, P. *et al.* A program for annotating and predicting the effects of single
699 nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster*
700 strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
- 701 58. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human
702 genomes. *Genome Res* **19**, 1553–1561 (2009).
- 703 59. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster
704 evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575–576
705 (2010).
- 706 60. Trbojević-Akmačić, I. *et al.* Chromatographic monoliths for high-throughput
707 immunoaffinity isolation of transferrin from human plasma. *Croat. Chem. Acta* **89**,
708 203–211 (2016).
- 709 61. Pucić, M. *et al.* High throughput isolation and glycosylation analysis of IgG-variability
710 and heritability of the IgG glycome in three isolated human populations. *Mol. Cell.*
711 *Proteomics* **10**, M111.010090-M111.010090 (2011).
- 712 62. Trbojević Akmačić, I., Ugrina, I. & Lauc, G. *Methods in Enzymology, Volume 586:*
713 *Chapter Three - Comparative Analysis and Validation of Different Steps in Glycomics*

- 714 *Studies*. (2017). doi:<https://doi.org/10.1016/bs.mie.2016.09.027>
- 715 63. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray
716 expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
- 717 64. Karssen, L. C., van Duijn, C. M. & Aulchenko, Y. S. The GenABEL Project for
718 statistical genomics. *F1000Research* **5**, 914 (2016).
- 719 65. Huffman, J. E. *et al.* Comparative Performance of Four Methods for High-throughput
720 Glycosylation Analysis of Immunoglobulin G in Genetic and Epidemiological
721 Research. *Mol. Cell. Proteomics* **13**, 1598–1610 (2014).
- 722 66. Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous
723 and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *Am J Hum*
724 *Genet* **104**, 260–274 (2019).
- 725 67. Kassambara, A. and Mundt, F. Factoextra: Extract and Visualize the Results of
726 Multivariate Data Analyses. R Package Version 1.0.7. [https://CRAN.R-](https://CRAN.R-project.org/package=factoextra)
727 [project.org/package=factoextra](https://CRAN.R-project.org/package=factoextra) (2020).
- 728

729 **Acknowledgements**

730

731 We thank Dr Nicola Pirastu for his help and advice regarding the statistical methods. The
732 Orkney Complex Disease Study (ORCADES) was supported by the Chief Scientist Office of
733 the Scottish Government (CZB/4/276 and CZB/4/710), the Royal Society, the MRC Human
734 Genetics Unit, Arthritis Research UK and the European Union framework program 6
735 EUROSPAN project (contract number LSHG-CT-2006-018947). ORCADES DNA
736 extractions and genotyping were performed at the Genetics Core of the Clinical Research
737 Facility, University of Edinburgh. We would like to acknowledge the invaluable contributions
738 of the research nurses in Orkney, the administrative team in Edinburgh, and the people of
739 Orkney. The CROATIA-Korcula study was funded by grants from the MRC (United
740 Kingdom), European Commission Framework 6 project EUROSPAN (contract number
741 LSHG-CT-2006-018947), Croatian Science Foundation (grant 8875) and the Republic of
742 Croatia Ministry of Science, Education and Sports (216-1080315-0302). Genotyping was
743 performed in the Genetics Core of the Clinical Research Facility, University of Edinburgh. We
744 would like to acknowledge all the staff of several institutions in Croatia that supported the
745 CROATIA-Korcula fieldwork, including, but not limited to, the University of Split and Zagreb
746 Medical Schools, Institute for Anthropological Research in Zagreb, and the Croatian Institute
747 for Public Health in Split. The Viking Health Study-Shetland (VIKING) was supported by the
748 MRC Human Genetics Unit quinquennial programme grant ‘QTL in Health and Disease’. DNA
749 extractions and genotyping were performed at the Edinburgh Clinical Research Facility,
750 University of Edinburgh. We would like to acknowledge the invaluable contributions of the
751 research nurses in Shetland, the administrative team in Edinburgh and the people of Shetland.
752 We acknowledge support from the European Union’s Horizon 2020 research and innovation
753 programme IMforFUTURE (A.L. and A.F.-H.: H2020-MSCA-ITN/721815); the RCUK
754 Innovation Fellowship from the National Productivity Investment Fund (L.K.: MR/R026408/1)
755 and the MRC Human Genetics Unit programme grant, ‘QTL in Health and Disease’ (J.F.W.
756 and C.H.: MC_UU_00007/10). Finally, this research has been conducted using data from the
757 UK Biobank Resource (under application 26041, 48511 and 19655). For the purpose of open
758 access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author
759 Accepted Manuscript version arising from this submission.

760

761 **Ethics**

762 All studies were approved by local research ethics committees and all participants have given
763 written informed consent. The ORCADES study was approved by the NHS Orkney Research
764 Ethics Committee and the North of Scotland REC. The CROATIA-Korcula study was
765 approved by the Ethics Committee of the Medical School, University of Split (approval ID:
766 2181-198-03-04/10-11-0008). The VIKING study was approved by the South East Scotland
767 Research Ethics Committee, NHS Lothian (reference: 12/SS/0151).

768

769 **Author contributions**

770 A.L.: Data analysis and interpretation, visualisation, writing—original draft preparation,
771 writing—review and editing. P.R.H.J.T.: preparation of pipeline for gene-based aggregation
772 test of rare variants, writing—review and editing. A.F.-H.: computation of new derived IgG
773 glycan traits, data interpretation. I.T.-A.: Quantification of transferrin and IgG N-glycans,
774 computation of derived transferrin glycan traits, writing—review and editing. F.V.: Glycan
775 data quality control. T.P.: Quantification of transferrin and IgG N-glycans. G.T.: preparation,
776 quality control and annotation of whole-exome sequencing data, writing—review and editing.
777 A.R.S.: Funding. O.P.: Genomic and demographic data provider for CROATIA-Korcula
778 cohort. C.H.: Genomic and demographic data provider for CROATIA-Korcula cohort. G.L.:
779 Conceptualisation, glycan data provider, writing—review and editing. J.F.W.: Funding,
780 conceptualisation, genomic and demographic data provider for ORCADES and VIKING
781 cohort, supervision, data interpretation, writing—original draft preparation, writing—review
782 and editing. L.K.: Conceptualisation, supervision, data interpretation, writing—original draft
783 preparation, writing—review and editing.
784

785 **Competing interests**

786 P.R.H.J.T. is an employee of BioAge Labs, Inc. G.T. and A.R.S. are full-time employees of
787 Regeneron Genetics Center and receive salary, stock and stock options as compensation. G.L.
788 is the founder and owner of Genos Ltd, a private research organisation that specialises in the
789 high-throughput glycomic analysis and has several patents in this field. A.F.-H., I.T.-A., F.V.,
790 and T.P. are employees of Genos Ltd. L.K. is an employee of Humanity Inc., a company
791 developing direct-to-consumer measures of biological ageing. All other authors declare no
792 competing interests.