Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers

Gordan Lauc1,2,3, Jennifer E. Huffman3,4, Maja Pučić1,4, Lina Zgaga4,5, Barbara Adamczyk2,6, Ana Mužinić1, Mislav Novokmet1, Ozren Polašek7, Olga Gornik8, Jasminka Krištic9, Toma Keser10, Veronique Vitart9, Blanca Scheijen8, Hae-Won Uh9,10, Mariam Molokhia8,10,11,12, Alan Leslie Patrick12, Paul McKeigue13, Ivana Kolčić14, Ivan Krešimir Lukić7, Olivia Swann8, Frank N. van Leeuwen8, L. Renee Ruhaak13, Jeanine J. Houwing-Duistermaat,9 P. Eline Slagboom10,14, Marian Beekman10,14, André Me Deelder15, Qiang Zeng17, Wei Wang18,19,20, Nicholas D. Hastie3, Ulf Gyllensten21, James F. Wilson4, Manfred Wuhrer16, Alan F. Wright3, Pauline M. Rudd4,6, Caroline Hayward23, Yuriu Aulchenko4,22, Harry Campbell45, Igor Rudan44*

1 Glycobiology Laboratory, Genos, Zagreb, Croatia, 2 Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb, Croatia, 3 Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom, 4 Centre for Population Health Sciences, The University of Edinburgh Medical School, Edinburgh, United Kingdom, 5 Faculty of Medicine, University of Zagreb, Zagreb, Croatia, 6 National Institute for Bioprocessing Research and Training, Dublin-Oxford Glycobiology Laboratory, Dublin, Ireland, 7 Faculty of Medicine, University of Split, Split, Croatia, 8 Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, 9 Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands, 10 Netherlands Consortium for Healthy Aging, Leiden, The Netherlands, 11 Department of Primary Care and Public Health Sciences, Kings College London, London, United Kingdom, 12 Kavli Institutet St. Medical Centre, Port of Spain, Trinidad and Tobago, 13 Department of Chemistry, University of California Davis, Davis, California, United States of America, 14 Department of Molecualr Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands, 15 Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands, 16 Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands, 17 Chinese PLA General Hospital, Beijing, China, 18 School of Public Health, Capital Medical University, Beijing, China, 19 Graduate University, Chinese Academy of Sciences, Beijing, China, 20 School of Medical Sciences, Edith Cowan University, Perth, Australia, 21 Department of Immunology, Genetics, and Pathology, SciLifeLab Uppsala, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden, 22 Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia

Abstract

Glycosylation of immunoglobulin G (IgG) influences IgG effector function by modulating binding to Fc receptors. To identify genetic loci associated with IgG glycosylation, we quantitated N-linked IgG glycans using two approaches. After isolating IgG from human plasma, we performed 77 quantitative measurements of N-glycosylation using ultra-performance liquid chromatography (UPLC) in 2,247 individuals from four European discovery populations. In parallel, we measured IgG N-glycans using MALDI-TOF mass spectrometry (MS) in a replication cohort of 1,848 Europeans. Meta-analysis of genome-wide association study (GWAS) results identified 9 genome-wide significant loci (P<2.27 x 10^-8) in the discovery analysis and two of the same loci (B4GALT1 and MGA73) in the replication cohort. Four loci contained genes encoding glycosyltransferases (ST6GAL1, B4GALT1, FUT8, and MGA73), while the remaining 5 contained genes that have not been previously implicated in protein glycosylation (IKZF1, IL6ST-ANKRD55, ABCF2, SMARC3, UV420H1, and SMARC11-DERL3). However, most of them have been strongly associated with autoimmune and inflammatory conditions (e.g., systemic lupus erythematosus, rheumatoid arthritis, ulcerative colitis, Crohn’s disease, diabetes type 1, multiple sclerosis, Graves’ disease, celiac disease, nodular sclerosis) and/or haematological cancers (acute lymphoblastic leukaemia, Hodgkin lymphoma, and multiple myeloma). Follow-up functional experiments in haplodeficient Ikzf1 knock-out mice showed the same general pattern of changes in IgG glycosylation as identified in the meta-analysis. As Ikzf1 was associated with multiple IgG-N-glycan traits, we explored biomarker potential of affected IgG glycans in 101 cases with SLE and 183 matched controls and demonstrated substantial discriminative power in a ROC-curve analysis (area under the curve = 0.842). Our study shows that it is possible to identify new loci that control glycosylation of a single plasma protein using GWAS. The results may also provide an explanation for the reported pleiotropy and antagonistic effects of loci involved in autoimmune diseases and haematological cancer.


Editor: Greg Gibson, Georgia Institute of Technology, United States of America

Received: August 30, 2012; Accepted: November 21, 2012; Published: January 31, 2013

Copyright: © 2013 Lauc et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The CROATIA-Vis study in the Croatian island of Vis was supported by grants from the Medical Research Council (UK); the Ministry of Science, Education, and Sport of the Republic of Croatia (grant number 108-108301-0302); and the European Union’s framework program 6 European Special Populations Research Network project (contract LSHG-CT-2006-018947). ORCADES was supported by the Chief Scientist Office of the Scottish Government, the Royal Society, and the European Union Framework Programme 6 European Special Populations Research Network project (contract LSHG-CT-2006-018947). Glycome analysis was supported by the Croatian Ministry of Science, Education, and Sport (grant numbers 399-061194-2023 and 216-108031-0302); the Croatian Science Foundation (grant number 04-47); the European Commission EuroGlycoarrays (contract #215536), GlycoBiOM (contract #259869), and HighGlycan (contract #278535) grants; and National Natural Science Foundation and Ministry of Science and Technology, China (grant numbers NSFC31070737 and 2012BAI37B03). The Leiden Longevity Study was supported by a grant from the Innovation-Oriented Research Program on Genomics (SenterNovem IGE05007), the Centre for Medical Systems Biology, and the Netherlands Consortium for Healthy Aging (grant 050-050-8110, all in the framework of the Netherlands Genomics Initiative, Netherlands Organization for Scientific Research (NWO). The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7-2007-2011) under grant agreement 259679. Research for the Systemic Lupus Erythmatosus (SLE) cases and controls series was supported with help from the ARUK (formerly ARC) grants M0600 and M0651 and by the National Institute for Health Research (NIHR) Biomolecular Research Centre at Guy’s and St. Thomas’ NHS Foundation Trust and King’s College London. The work of YA was supported by Helmholtz-RFBR JRG grant (12-04-91322-CN) and B5 acknowledges support from a grant from the Dutch childhood oncology foundation KiKa. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Introduction

Glycosylation is a ubiquitous post-translational protein modification that modulates the structure and function of polypeptide components of glycoproteins [1,2]. N-glycan structures are essential for multicellular life [3]. Mutations in genes involved in modification of glycan antennae are common and can lead to severe or fatal diseases [4]. Variation in protein glycosylation also has physiological significance, with immunoglobulin G (IgG) being a well-documented example. Each heavy chain of IgG carries a single covalently attached bi-antennary N-glycan at the highly conserved asparagine 297 residue in each of the CH2 domains of the Fc region of the molecule. The attached oligosaccharides are structurally important for the stability of the antibody and its effector functions [5]. In addition, some 15–20% of normal IgG molecules have complex bi-antennary oligosaccharides in the variable regions of light or heavy chains [6,7]. 36 different glycans (Figure 1) can be attached to the conserved Asn297 of the IgG heavy chain [8,9], leading to hundreds of different IgG isomers that can be generated from this single glycosylation site.

Glycosylation of IgG has important regulatory functions. The absence of galactose residues in association with rheumatoid arthritis was reported nearly 30 years ago [10]. The addition of sialic acid dramatically changes the physiological role of IgGs, converting them from pro-inflammatory to anti-inflammatory agents [11,12]. Addition of fucose to the glycan core interferes with the binding of IgG to FcyRIIIa and greatly diminishes its capacity for antibody dependent cell-mediated cytotoxicity (ADCC) [13,14]. Structural analysis of the IgG-Fc/FcγRIIIa complex has demonstrated that specific glycans on FcγRIIIa are also essential for this effect of core-fucose [15] and that removal of core fucose from IgG glycans increases clinical efficacy of monoclonal antibodies, enhancing their therapeutic effect through ADCC mediated killing [16–18].

New high-throughput technologies, such as high/ultra performance liquid chromatography (HPLC/UPLC), MALDI-TOF mass spectrometry (MS) and capillary electrophoresis (CE), allow us to determine N-linked glycans from individual human plasma proteins. Recently, we performed the first population-based study to demonstrate physiological variation in IgG glycosylation in three European founder populations [19]. Using UPLC, we showed exceptionally high individual variability in glycosylation of a single protein - human IgG - and substantial heritability of the observed measurements [19]. In parallel, we quantitated IgG N-glycans in another European population (Leiden Longevity Study – LLS) by mass spectrometry. In this study, we combined those high-throughput glycomics measurements with high-throughput genomics to perform the first genome wide association (GWA) study of the human IgG N-glycome.

Results

Genome-wide association study and meta-analysis

We separated a single protein (IgG) from human plasma and quantitated its N-linked glycans using two state-of-the-art technologies (UPLC and MALDI-TOF MS). Their comparative advantages in GWA studies were difficult to predict prior to the conducted analyses, so both were used - one in each available cohort. We performed 77 quantitative measurements of IgG N-glycosylation using ultra performance liquid chromatography (UPLC) in 2247 individuals from four European discovery populations (CROATIA-Vis, CROATIA-Korcula, ORCADES, NSPHS). In parallel, we measured IgG N-glycans using MALDI-TOF mass spectrometry (MS) in 1848 individuals from another European population (Leiden Longevity Study (LLS)). Descriptions of these population cohorts are found in Table S11. Aiming to identify genetic loci involved in IgG glycosylation, we performed a GWA study in both cohorts. Associations at 9 loci reached genome-wide significance (p<2.27×10−8) in the discovery meta-analysis and at two loci in the replication cohort. The two loci identified in the latter cohort were associated with the analogous glycan traits in the former cohort as detailed in the subsection “Replication of our findings”. Both UPLC and MS methods for quantitation of N-glycans were found to be amenable to GWA studies. Since our UPLC study gave a considerably greater yield of significant findings in comparison to MS study, the majority of our results section focuses on the findings from the discovery population cohort, which was used studying the UPLC method.

Among the nine loci that passed the genome-wide significance threshold, four contained genes encoding glycosyltransferases (ST6GAL1, B4GALT1, FUT8 and MALT3), while the remaining five loci contained genes that have not been implicated in protein glycosylation previously (IKZF1, IL6ST-ANKRD53, ABCF2-SMARCD3, SUV420H1-CHKA and SMARCBI-DERL3). As a rule, the implicated genes were associated with several N-glycan traits. The explanation and notation of the 77 N-glycan measures is presented in Table S1. It comprises 23 directly measured quantitative IgG glycosylation traits (shown in Figure 1) and 54 derived traits. Descriptive statistics of these measures in the discovery cohorts are presented in Table S2. GWA analysis was performed in each of the populations separately and the results were combined in an inverse-variance weighted meta-analysis. Summary data for each gene region showing genome-wide association (p<2.27×10−8) or found to be strongly suggestive (2.27×10−5<p<5×10−6) are presented in Table 1. Summary data for all single-nucleotide polymorphisms (SNPs) and traits with suggestive associations (p<1×10−5) are presented in Table S3, with population-specific and pooled genomic control (GC) factors reported in Table S4.

The most statistically significant association was observed in a region on chromosome 3 containing the gene ST6GAL1 (Table 1, Figure S1A). ST6GAL1 codes for the enzyme sialyltransferase 6 which adds sialic acid to various glycoproteins including IgG glycans (Figure 2), and is therefore a highly biologically plausible candidate. In this region of about 70 kilobases (kb) we identified 37 genome-wide significant SNPs associated with 14 different IgG glycosylation traits, generally reflecting sialylation of different glycan structures (Table 1). The strongest association was observed for the percentage of monosialylation of fucosylated digalactosylated structures in total IgG glycans (IGP29, see Figure 1 and Table S1 for notation), for which a SNP rs11710456 explained 17%, 16%, 18% and 3% of the trait variation for CROATIA-Vis.
CROATIA-Korcula, ORCADES and NSPHS respectively (meta-analysis $p = 6.12 \times 10^{-7}$). NSPHS had a very small sample size in this analysis ($N = 179$) and may not provide an accurate portrayal of the variance explained in this particular population (estimated as 3%). Although the allele frequency is similar between all populations, in the forest plot (Figure S1A) although NSPHS does overlap with the other populations, the 95% CI is much larger. It is also possible that there are population-specific genetic and/or environmental differences in NSPHS that are affecting the amount of variance explained by this SNP. After analysis conditioning on the top SNP (rs11710456) in this region, the SNP rs7652995 still reached genome-wide significance ($p = 4.15 \times 10^{-13}$). After adjusting for this additional SNP, the association peak was completely removed. This suggests that there are several genetic factors underlying this association. Conditional analysis of all other significant and suggestive regions resulted in the complete removal of the association peak.

We also identified 28 SNPs showing genome-wide significant associations with 11 IgG glycosylation traits ($2.70 \times 10^{-11} < p < 4.73 \times 10^{-8}$) at a locus on chromosome 9 spanning over 60 kb (Figure S1B). This region includes $B4GALT1$, which codes for the galactosyltransferase responsible for the addition of galactose to IgG glycans (Figure 2). The glycan traits showing genome-wide association included the percentage of FA2G2S1 in the total fraction (IGP17), the percentage of FA2G2 in the total and neutral fraction (IGP13, IGP53), the percentage of sialylation of fucosylated structures without bisecting GlcNAc (IGP24, IGP26), the percentage of digalactosylated structures in the total neutral fraction (IGP57) and, in the opposite direction, the percentage of bisecting GlcNAc in fucosylated sialylated structures (IGP36-IGP40).

**Figure 1. Structures of glycans separated by HILIC-UPLC analysis of the IgG glycome.**
doi:10.1371/journal.pgen.1003225.g001
Table 1. A complete list of genetic markers that showed genome-wide significant (P<2.27E-9) or strongly suggestive (P≤5E-08) association with glycosylation of Immunoglobulin G analysed by UPLC in the discovery meta-analysis.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>SNP with lowest P-value</th>
<th>Lowest P-value</th>
<th>Effect size* (s.e.)</th>
<th>MAF</th>
<th>Interval size, kb</th>
<th>nHits</th>
<th>nTraits</th>
<th>Genes in the interval</th>
<th>Trait with lowest P-value*</th>
<th>Other Associated Traits</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>rs11710456</td>
<td>6.12E-75</td>
<td>0.64 (0.04)</td>
<td>0.30</td>
<td>14.2</td>
<td>20</td>
<td>14</td>
<td>ST6GAL1</td>
<td>IGP29</td>
<td>IGP14, IGP15, IGP17, IGP23, IGP24, IGP26, IGP28, IGP30, IGP31, IGP32, IGP35, IGP37, IGP38</td>
</tr>
<tr>
<td>5</td>
<td>rs17348299</td>
<td>6.88E-11</td>
<td>0.29 (0.04)</td>
<td>0.16</td>
<td>16.1</td>
<td>4</td>
<td>6</td>
<td>IL6ST-ANKRD55</td>
<td>IGP53</td>
<td>IGP3, IGP13, IGP43, IGP55, IGP57</td>
</tr>
<tr>
<td>7</td>
<td>rs6421315</td>
<td>1.87E-13</td>
<td>0.23 (0.03)</td>
<td>0.37</td>
<td>21.4</td>
<td>11</td>
<td>13</td>
<td>IKZF1</td>
<td>IGP63</td>
<td>IGP2, IGP6, IGP42, IGP45, IGP6, IGP67, IGP70, IGP71, IGP72</td>
</tr>
<tr>
<td>7</td>
<td>rs1122979</td>
<td>2.10E-10</td>
<td>0.31 (0.05)</td>
<td>0.12</td>
<td>62.3</td>
<td>3</td>
<td>4</td>
<td>ABCF2-SMARCD3</td>
<td>IGP2</td>
<td>IGP5, IGP42, IGP45</td>
</tr>
<tr>
<td>9</td>
<td>rs12342831</td>
<td>2.70E-11</td>
<td>0.24 (0.04)</td>
<td>0.26</td>
<td>60.1</td>
<td>28</td>
<td>11</td>
<td>B4GALT1</td>
<td>IGP17</td>
<td>IGP13, IGP24, IGP26, IGP36, IGP37, IGP38, IGP39, IGP40, IGP53, IGP57</td>
</tr>
<tr>
<td>11</td>
<td>rs4930561</td>
<td>8.88E-10</td>
<td>0.19 (0.03)</td>
<td>0.49</td>
<td>58.7</td>
<td>5</td>
<td>2</td>
<td>SUV420H1</td>
<td>IGP41</td>
<td>IGP12</td>
</tr>
<tr>
<td>14</td>
<td>rs11847263</td>
<td>1.08E-22</td>
<td>0.31 (0.03)</td>
<td>0.39</td>
<td>17.1</td>
<td>167</td>
<td>12</td>
<td>FUT8</td>
<td>IGP59</td>
<td>IGP2, IGP6, IGP11, IGP42, IGP46, IGP51, IGP58, IGP60, IGP61, IGP63, IGP65</td>
</tr>
<tr>
<td>22</td>
<td>rs2186369</td>
<td>8.63E-17</td>
<td>0.35 (0.04)</td>
<td>0.19</td>
<td>49.4</td>
<td>10</td>
<td>20</td>
<td>SMARC81-DERL3</td>
<td>IGP72</td>
<td>IGP9, IGP10, IGP14, IGP39, IGP40, IGP49, IGP50, IGP52, IGP62, IGP63, IGP64, IGP66, IGP67, IGP68, IGP69, IGP70, IGP71, IGP74, IGP75, IGP76</td>
</tr>
<tr>
<td>22</td>
<td>rs909674</td>
<td>9.66E-25</td>
<td>0.34 (0.03)</td>
<td>0.30</td>
<td>27.9</td>
<td>60</td>
<td>17</td>
<td>SYNGR1-TAB1-MGAT3-CACNA1I</td>
<td>IGP40</td>
<td>IGP5, IGP9, IGP22, IGP34, IGP39, IGP45, IGP49, IGP62, IGP63, IGP64, IGP66, IGP67, IGP68, IGP70, IGP71, IGP72</td>
</tr>
</tbody>
</table>

Interval: size (kb) of the genomic interval containing SNPs with R^2>=0.6 with top associated SNP; nHits: number of SNPs with GW-significant association; nTraits: number of IgG glycosylation traits associated with the region at GW-significant level.

*effect size is in z-score units after adjustment for sex, age and first 3 principal components.

+Description of the traits provided in Table S1;

The SNP effect in opposite direction to most significant trait.

doi:10.1371/journal.pgen.1003225.t001
A large (541 kb) region on chromosome 14 harbouring the \textit{FUT8} gene contained 167 SNPs showing significant associations with 12 IgG glycosylation traits reflecting fucosylation of IgG glycans (Figure S1C). \textit{FUT8} codes for fucosyltransferase 8, an enzyme responsible for the addition of fucose to IgG glycans (Figure 2). The strongest association ($1.08 \times 10^{-22}$, $p = 2.60 \times 10^{-17}$) was observed with the percentage of A2 glycans in total and neutral fractions (IGP2, IGP42) and for derived traits related to the proportion of fucosylation (IGP58, IGP59 and IGP61; all in the opposite direction). In summary, SNPs at the \textit{FUT8} locus influence the proportion of fucosylated glycans, and, in the opposite direction, the percentages of A2, A2G1 and A2G2 glycans which are not fucosylated.

On chromosome 22, two loci were associated with IgG glycosylation. The first region, containing \textit{SYNGR1}-\textit{TAB1}-\textit{MGAT3}-\textit{CACNA1I} genes, spans over 233 kb. This region harboured 60 SNPs showing genome-wide significant association with 17 IgG glycosylation traits (Figure S1D). Association was strongest between SNP rs909674 and the incidence of bisecting GlcNAc in all fucosylated disialylated structures (IGP40, $p = 9.66 \times 10^{-25}$) and the related ratio IGP39 ($p = 8.87 \times 10^{-24}$). In summary, this locus contained variants influencing levels of fucosylated species and the ratio between fucosylated (especially disialylated) structures with and without bisecting GlcNAc (Figure 2). Since \textit{MGAT3} codes for the enzyme N-acetylglucosaminyltransferase III (beta-1,4-mannosyl-glycoprotein-4-beta-N-acetylglucosaminyltransferase), which is responsible for the addition of bisecting GlcNAc to IgG glycans, this gene is the most biologically plausible candidate.

Bioinformatic analysis of known and predicted protein-protein interactions using String 9.0 software (http://string-db.org/) showed that interactions between the clusters of \textit{FUT8}-\textit{B4GALT1}-\textit{MGAT3} and \textit{ST6GAL1}-\textit{B4GALT1}-\textit{MGAT3} genes had high confidence score: \textit{FUT8}-\textit{B4GALT1} of 0.90; \textit{FUT8}-\textit{MGAT3} of 0.95; \textit{ST6GAL1}-\textit{B4GALT1} of 0.90; and \textit{ST6GAL1}-
Located on chromosome 11, the SNP rs4930561 located in the SUV420H1 gene was associated with percentage of FA1 in neutral IgG glycans (p = 8.49 × 10^{-4}), SUV420H1 codes for histone-lysine N-methyltransferase which specifically trimethylates lysine 20 of histone H4 and could therefore affect activity of many different genes; it is thought to be involved in proviral silencing in somatic and germ line cells through epigenetic mechanisms [60]. **CHKA** has a key role in phospholipid biosynthesis and may contribute to tumour cell growth. We recently reported a number of strong associations between lipiddomics and glycomics traits in human plasma [61]. Thus, an enzyme involved in phospholipid synthesis is also a possible candidate because the lipid environment is known to affect glycosyltransferases activity [61].

Three further loci were identified as strongly suggestive through GWAS and deserve attention for their possible pleiotropic effects. SNP rs9296009 in **PRTT1** (proline-rich transmembrane protein 1) was associated with IgG2 (p = 3.79 × 10^{-8}) and total (IGP1; p = 1.30 × 10^{-8}) fractions of IgG glycans. **SUV420H1** codes for histone-lysine N-methyltransferase which specifically trimethylates lysine 20 of histone H4 and could therefore affect activity of many different genes; it is thought to be involved in proviral silencing in somatic and germ line cells through epigenetic mechanisms [60]. **CHKA** has a key role in phospholipid biosynthesis and may contribute to tumour cell growth. We recently reported a number of strong associations between lipiddomics and glycomics traits in human plasma [61]. Thus, an enzyme involved in phospholipid synthesis is also a possible candidate because the lipid environment is known to affect glycosyltransferases activity [61].

Three further loci were identified as strongly suggestive through GWAS and deserve attention for their possible pleiotropic effects. SNP rs9296009 in **PRTT1** (proline-rich transmembrane protein 1) was associated with IgG2 (p = 3.79 × 10^{-8}) and total (IGP1; p = 1.30 × 10^{-8}) fractions of IgG glycans. **SUV420H1** codes for histone-lysine N-methyltransferase which specifically trimethylates lysine 20 of histone H4 and could therefore affect activity of many different genes; it is thought to be involved in proviral silencing in somatic and germ line cells through epigenetic mechanisms [60]. **CHKA** has a key role in phospholipid biosynthesis and may contribute to tumour cell growth. We recently reported a number of strong associations between lipiddomics and glycomics traits in human plasma [61]. Thus, an enzyme involved in phospholipid synthesis is also a possible candidate because the lipid environment is known to affect glycosyltransferases activity [61].

Three further loci were identified as strongly suggestive through GWAS and deserve attention for their possible pleiotropic effects. SNP rs9296009 in **PRTT1** (proline-rich transmembrane protein 1) was associated with IgG2 (p = 3.79 × 10^{-8}) and total (IGP1; p = 1.30 × 10^{-8}) fractions of IgG glycans. **SUV420H1** codes for histone-lysine N-methyltransferase which specifically trimethylates lysine 20 of histone H4 and could therefore affect activity of many different genes; it is thought to be involved in proviral silencing in somatic and germ line cells through epigenetic mechanisms [60]. **CHKA** has a key role in phospholipid biosynthesis and may contribute to tumour cell growth. We recently reported a number of strong associations between lipiddomics and glycomics traits in human plasma [61]. Thus, an enzyme involved in phospholipid synthesis is also a possible candidate because the lipid environment is known to affect glycosyltransferases activity [61].
Table 2. An analysis of pleiotropy between loci associated with IgG glycans and previously reported disease/trait susceptibility loci, with linkage disequilibrium computed between the most significantly associated SNPs.

<table>
<thead>
<tr>
<th>Gene</th>
<th>IgG Glycan Top SNP</th>
<th>Disease</th>
<th>Top Disease SNP</th>
<th>Risk Allele</th>
<th>P-Value</th>
<th>Reference</th>
<th>Ancestry</th>
<th>HapMap 2</th>
<th>1000G Pilot 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R²</td>
<td>D'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>D</td>
<td>R²</td>
</tr>
<tr>
<td>IKZF1</td>
<td>rs6421315</td>
<td>SLE</td>
<td>rs921916</td>
<td>C</td>
<td>2.00E-06</td>
<td>Gateva et al Nat Genet 2009</td>
<td>European</td>
<td>0.021</td>
<td>0.388</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs2366293</td>
<td>G</td>
<td>2.33E-09</td>
<td>Cunninghame Graham et al PLoS Genet 2011</td>
<td>European</td>
<td>0.030</td>
<td>0.484</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs4917014</td>
<td>A</td>
<td>3.00E-23</td>
<td>Han et al Nat Genet 2009</td>
<td>Han Chinese</td>
<td>0.001</td>
<td>0.040</td>
</tr>
<tr>
<td>ALL</td>
<td>rs11978267</td>
<td>G</td>
<td>8.00E-11</td>
<td></td>
<td></td>
<td>Trevino et al Nat Genet 2009</td>
<td>European</td>
<td>0.002</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs1432601</td>
<td>C</td>
<td>1.00E-19</td>
<td>Papeammanuil et al Nat Genet 2009</td>
<td>European</td>
<td>0.002</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs10276619</td>
<td>–</td>
<td>3.00E-06</td>
<td>Potkin et al PLoS One 2009</td>
<td>European</td>
<td>0.005</td>
<td>0</td>
</tr>
<tr>
<td>Hippocampal atrophy (AD qt)</td>
<td>rs7805803</td>
<td>–</td>
<td>9.00E-06</td>
<td></td>
<td></td>
<td>Furney et al Mol Psychiatry 2010</td>
<td>European</td>
<td>0.071</td>
<td>0.280</td>
</tr>
<tr>
<td>Total ventricular volume (AD qt)</td>
<td>rs1456893</td>
<td>A</td>
<td>5.00E-09</td>
<td></td>
<td></td>
<td>Barrett et al Nat Genet 2008</td>
<td>European</td>
<td>0.011</td>
<td>0.117</td>
</tr>
<tr>
<td>Mean corpuscular volume</td>
<td>rs1271859</td>
<td>A</td>
<td>5.00E-13</td>
<td></td>
<td></td>
<td>Ganesh et al Nat Genet 2009</td>
<td>European</td>
<td>0.018</td>
<td>0.181</td>
</tr>
<tr>
<td>Malaria</td>
<td>rs1451375</td>
<td>–</td>
<td>6.00E-06</td>
<td></td>
<td></td>
<td>Jallow et al Nat Genet 2009</td>
<td>Gambian</td>
<td>0.014</td>
<td>0.204</td>
</tr>
<tr>
<td>Systemic sclerosis</td>
<td>rs1240874</td>
<td>–</td>
<td>1.00E-06</td>
<td></td>
<td></td>
<td>Gorlova et al PLoS Genet 2011</td>
<td>European</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>T1D</td>
<td>rs10272274</td>
<td>C</td>
<td>1.10E-11</td>
<td></td>
<td></td>
<td>Swafford et al Diabetes 2011</td>
<td>European</td>
<td>0.002</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>rs11710456</td>
<td>Drug-induced liver injury (flucloxacillin)</td>
<td>rs10937275</td>
<td>–</td>
<td>1.00E-08</td>
<td>Daly et al Nat Genet 2009</td>
<td>European</td>
<td>0</td>
<td>0.048</td>
</tr>
<tr>
<td>T2D</td>
<td>rs16816329</td>
<td>G</td>
<td>3.00E-08</td>
<td></td>
<td></td>
<td>Kooner et al Nat Genet 2011</td>
<td>South Asian</td>
<td>0.011</td>
<td>0.221</td>
</tr>
<tr>
<td>IL6ST-ANKRDSS</td>
<td>rs17348299</td>
<td>Rheumatoid arthritis</td>
<td>rs6859219</td>
<td>C</td>
<td>1.00E-11</td>
<td>Stahl et al Nat Genet 2010</td>
<td>European</td>
<td>0.012</td>
<td>0.487</td>
</tr>
<tr>
<td>LAMBI</td>
<td>rs2072209</td>
<td>Ulcerative colitis</td>
<td>rs2158836</td>
<td>A</td>
<td>7.00E-06</td>
<td>Silverberg et al Nat Genet 2009</td>
<td>European</td>
<td>0.027</td>
<td>0.716</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ulcerative colitis</td>
<td>rs4591895</td>
<td>A</td>
<td>8.00E-08</td>
<td>McGovern et al Nat Genet 2010</td>
<td>European</td>
<td>0.071</td>
<td>0.807</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ulcerative colitis</td>
<td>rs868774</td>
<td>G</td>
<td>3.00E-08</td>
<td>Barrett et al Nat Genet 2009</td>
<td>European</td>
<td>0.031</td>
<td>0.733</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ulcerative colitis</td>
<td>rs4510766</td>
<td>A</td>
<td>2.00E-16</td>
<td>Anderson et al Nat Genet 2011</td>
<td>European</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ulcerative colitis</td>
<td>rs4730276</td>
<td>–</td>
<td>9.00E-06</td>
<td>Silverberg et al Nat Genet 2009</td>
<td>European</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ulcerative colitis</td>
<td>rs4730273</td>
<td>–</td>
<td>5.00E-06</td>
<td>Silverberg et al Nat Genet 2009</td>
<td>European</td>
<td>0.032</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ulcerative colitis</td>
<td>rs2108225</td>
<td>A</td>
<td>1.00E-07</td>
<td>Asano et al Nat Genet 2011</td>
<td>Japanese</td>
<td>0.024</td>
<td>0.548</td>
</tr>
<tr>
<td>FUT8</td>
<td>rs11847263</td>
<td>N-Glycans (DGd)</td>
<td>rs10483776</td>
<td>G</td>
<td>1.00E-08</td>
<td>Lauc et al PLoS Genet 2010</td>
<td>European</td>
<td>0.369</td>
<td>0.864</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N-Glycans (DG1)</td>
<td>rs7159888</td>
<td>A</td>
<td>3.00E-18</td>
<td>Lauc et al PLoS Genet 2010</td>
<td>European</td>
<td>0.714</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Conduct disorder (symptom count)</td>
<td>rs1256531</td>
<td>–</td>
<td>4.00E-06</td>
<td>Dick et al Mol Psychiatry 2010</td>
<td>European, African, other</td>
<td>0.092</td>
<td>0.796</td>
</tr>
<tr>
<td>Waist Circumference</td>
<td>rs7158173</td>
<td>–</td>
<td>4.00E-06</td>
<td></td>
<td></td>
<td>Polasek et al Croat Med J 2009</td>
<td>European</td>
<td>0.011</td>
<td>0.182</td>
</tr>
<tr>
<td>Multiple Sclerosis - brain glutamate levels</td>
<td>rs8007846</td>
<td>–</td>
<td>9.00E-06</td>
<td></td>
<td></td>
<td>Baranzini et al Brain 2010</td>
<td>American</td>
<td>0.196</td>
<td>0.571</td>
</tr>
<tr>
<td>Gene</td>
<td>IgG Glycan</td>
<td>Top SNP</td>
<td>Top Disease SNP</td>
<td>Risk Allele</td>
<td>P-Value</td>
<td>Reference</td>
<td>Ancestry</td>
<td>R²</td>
<td>D¹</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>--------</td>
<td>-----------------</td>
<td>-------------</td>
<td>---------</td>
<td>-----------</td>
<td>----------</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>SYNGR1-TAB1-MGAT3-CACNA11</td>
<td>Sudden cardiac arrest</td>
<td>rs909674</td>
<td>rs54211</td>
<td>–</td>
<td>8.00E-07</td>
<td>Aouizerat et al BMC Car Diso 2011</td>
<td>European</td>
<td>0.053</td>
<td>0.362</td>
</tr>
<tr>
<td>Primary biliary cirrhosis</td>
<td>rs968451</td>
<td>T</td>
<td>1.00E-09</td>
<td>Mells et al Nat Genet 2011</td>
<td>European</td>
<td>0.041</td>
<td>0.682</td>
<td>0.080</td>
<td>1</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>rs2413583</td>
<td>C</td>
<td>1.00E-26</td>
<td>Franke et al Nat Genet 2010</td>
<td>European</td>
<td>0.043</td>
<td>0.313</td>
<td>0.053</td>
<td>0.292</td>
</tr>
<tr>
<td>SMARC81-DERL3</td>
<td>GGT</td>
<td>rs2186369</td>
<td>T</td>
<td>2.00E-09</td>
<td>Chambers et al Nat Genet 2011</td>
<td>European</td>
<td>0.009</td>
<td>0.255</td>
<td>0</td>
</tr>
<tr>
<td>PRRT1</td>
<td>rs9296009</td>
<td>Nodular sclerosis Hodgkin lymphoma</td>
<td>rs204999</td>
<td>–</td>
<td>8.00E-18</td>
<td>Cozen et al Blood 2012</td>
<td>European</td>
<td>0.125</td>
<td>1</td>
</tr>
<tr>
<td>HLA-DQA2 - HLA-DQB2</td>
<td>SLE</td>
<td>rs1049110</td>
<td>rs2301271</td>
<td>T</td>
<td>2.00E-12</td>
<td>Chung et al PLoS Genet 2011</td>
<td>European</td>
<td>0.967</td>
<td>1</td>
</tr>
<tr>
<td>Hepatitis B</td>
<td>rs743920</td>
<td>G</td>
<td>6.00E-28</td>
<td>Mbarek et al Hum Mol Genet 2011</td>
<td>Japanese</td>
<td>0.967</td>
<td>1</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Narcolepsy</td>
<td>rs2858884</td>
<td>A</td>
<td>3.00E-08</td>
<td>Hor et al Nat Genet 2010</td>
<td>European</td>
<td>0.193</td>
<td>1</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>BACH2</td>
<td>rs404256</td>
<td>Graves’ disease</td>
<td>rs370409</td>
<td>T</td>
<td>2.00E-06</td>
<td>Chu et al Nat Genet 2011</td>
<td>Chinese</td>
<td>0.010</td>
<td>0.187</td>
</tr>
<tr>
<td>Celiac disease</td>
<td>rs10806425</td>
<td>A</td>
<td>4.00E-10</td>
<td>Dubois et al Nat Genet 2010</td>
<td>European</td>
<td>0.005</td>
<td>0.103</td>
<td>0.006</td>
<td>0.09</td>
</tr>
<tr>
<td>T1D</td>
<td>rs3575247</td>
<td>A</td>
<td>1.00E-06</td>
<td>Grant et al Diabetes 2009</td>
<td>European</td>
<td>0.039</td>
<td>0.196</td>
<td>0.042</td>
<td>0.204</td>
</tr>
<tr>
<td>T1D</td>
<td>rs1175527</td>
<td>G</td>
<td>3.00E-08</td>
<td>Plagnol et al PloSGenet 2011</td>
<td>European</td>
<td>0.031</td>
<td>0.186</td>
<td>0.031</td>
<td>0.179</td>
</tr>
<tr>
<td>T1D</td>
<td>rs1175527</td>
<td>–</td>
<td>5.00E-08</td>
<td>Barnett et al Nat Genet 2009</td>
<td>European</td>
<td>0.031</td>
<td>0.186</td>
<td>0.031</td>
<td>0.179</td>
</tr>
<tr>
<td>T1D</td>
<td>rs1175527</td>
<td>G</td>
<td>5.00E-12</td>
<td>Cooper et al Nat Genet 2008</td>
<td>European</td>
<td>0.031</td>
<td>0.186</td>
<td>0.031</td>
<td>0.179</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>rs1847472</td>
<td>G</td>
<td>5.00E-09</td>
<td>Franke et al Nat Genet 2010</td>
<td>European</td>
<td>0</td>
<td>0.011</td>
<td>0.009</td>
<td>0.124</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>rs12212193</td>
<td>G</td>
<td>4.00E-08</td>
<td>Sawcer et al Nature 2011</td>
<td>European</td>
<td>0.001</td>
<td>0.036</td>
<td>0.027</td>
<td>0.166</td>
</tr>
<tr>
<td>SLC38A10</td>
<td>rs7224668</td>
<td>Longevity</td>
<td>rs10445407</td>
<td>T</td>
<td>1.00E-06</td>
<td>Yashin et al Aging 2010</td>
<td>European</td>
<td>0.714</td>
<td>1</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pgen.1003225.t002
IGP31 \( (p = 3.33 \times 10^{-05}) \). Although the function of this gene is not understood, it has been associated with autism and longevity [65,66].

The remaining three signals implicated ABCF2-SMARCD3 region (rs1122959) was associated with IGP2, 5, 42, 45, with \( p \)-value ranging between \( 2.10 \times 10^{-05} \) to \( 1.80 \times 10^{-05} \). RECK (rs14878639) was suggestively associated with IGP17; \( p = 5.31 \times 10^{-05} \) and PEX5 (rs12092421) suggestively associated with IGP41; \( p = 4.40 \times 10^{-05} \). The function of ABCF2 (ATP-binding cassette, sub-family F, member 2) is not well understood. SMARCD3 stimulates nuclear receptor mediated transcription; it belongs to the neural progenitors-specific chromatin remodelling complex (npBAF complex) and the neuron-specific chromatin-remodelling complex (nBAF complex). RECK is known to be a strong suppressor of tumour invasion and metastasis, regulating metalloproteinases which are involved in cancer progression. PEX 5 binds to the C-terminal PTS1-type tripeptide peroxisomal targeting signal and plays an essential role in peroxisomal protein import (www.genecards.org).

Results from an independent cohort using MS quantitation method

The parallel effort in the outbred Leiden Longevity Study (LLS) was based on a different \( N \)-glycan quantitation method (MS). While UPLC groups glycans according to structural similarities, MS groups them by mass. Furthermore, MS analysis focused on Fc glycans while UPLC measures both Fc and Fab glycans, thus traits measured by the two methods could not have been directly compared. Glycosylation patterns of IgG1 and IgG2 were investigated by analysis of tryptic glycopeptides, with six glycoforms per IgG subclass measured. The intensities of all glycoforms were related to the monogalactosylated, core-fucosylated biantennary species, providing five relative intensities registered per IgG subclass (Tables S5 and S6). The analysis identified two loci as genome-wide significant - implicating MGTAT3 \( (p = 1.6 \times 10^{-10}) \) for G1F, analogous to UPLC IGP5; \( p = 3.12 \times 10^{-08} \) for G0FN, analogous to UPLC IGP9, and B4GALT1 \( (p = 5.4 \times 10^{-08}) \) for G2F, analogous to UPLC IGP13) confirming GWAS signals in the discovery meta-analysis.

Replication of our findings

We then sought a separate independent replication of the other 14 genome-wide significant and strongly suggestive signals identified in the discovery analysis, which was performed in the LLS cohort, appreciating that the quantitated \( N \)-glycan traits do not exactly match between the two cohorts. SNPs were chosen for replication based on initial meta-analysis results of genotype data prior to imputed analysis. All five traits measured in LLS cohort were tested for association with all the selected SNPs (Table S6). We were able to reproduce association to ST6GAL1 \( (p = 0.1 \times 10^{-07}) \) for G2F, substrate for sialyltransferase) and SMARCB1-DERL3 \( (p = 1.6 \times 10^{-07}) \) (G1N, analogous to UPLC IGP9). Weaker, though nominally significant associations were confirmed at IKZF1 \( (p = 2.3 \times 10^{-05}) \) for G1N, SLC38A10 \( (p = 4.3 \times 10^{-05}) \) for G2N, ILST-ANKRD53 \( (p = 1.3 \times 10^{-07}) \) for G0N and ABCF2-SMARCD3 \( (p = 2.7 \times 10^{-5}) \) for G2N. The fact that we did not replicate associations at the other 8 loci was not unexpected, because those 8 loci showed association with UPLC-measured \( N \)-glycan traits that do not compare to any of the traits measured by MS (see Table S5 for comparison of MS and UPLC traits).

Functional experiment: Ikzf1 haplodeficiency results in altered \( N \)-glycosylation of IgG

IKZF1 is considered to be the important regulator governing differentiation of T cells into CD4+ and CD8+ T cells [67]. Since glycan traits associated with IKZF1 were related to the presence and absence of core-fucose and bisecting GlcNAc, we analysed the promoter region of MGTAT3 (codes for enzyme that adds bisecting GlcNAc to IgG glycans) in silico and identified two binding sites for IKZF1 that were conserved between humans and mice, while recognition sites for IKZF1 were not found in the promoter region of PGT1 (which codes for an enzyme that adds core-fucose to IgG glycans). Since the promoter regions of MGTAT3 were conserved between humans and mice, we used Ikzf1 knockout mice [68] as a model to study the effects of IKZF1 deficiency on IgG glycosylation. IgG was isolated from the plasma of 5 heterozygous knockout mice and 5 wild-type controls. The summary of the results of IgG glycosylation analysis is presented in Table 3, while complete results are presented in Table S7. We observed a number of alterations in glycome composition that were all consistent with the role of IKZF1 in the down-regulation of fucosylation and up-regulation of the addition of bisecting GlcNAc to IgG glycans; 12 out of 77 IgG N-glycans measured showed statistically significant difference \( (p<0.05) \) between wild type and heterozygous Ikzf1 knock-outs, where 5 mice from each group were compared (Table 3).

Investigating the biomarker potential of IgG N-glycans in Systemic Lupus Erythematosus (SLE)

Given that IKZF1 has been convincingly associated with SLE in previous studies [35–37], and that functional studies in heterozygous knock-out mice in our study showed clear differences in profiles of several IgG \( N \)-glycan traits, we explored an intriguing hypothesis: whether the same IgG \( N \)-glycan traits that were significantly affected in Ikzf1 knock-out mice could be demonstrated to differ between human SLE cases and controls. If this were true, then pleiotropy between the effects of IKZF1 on SLE and on IgG \( N \)-glycans in human plasma, revealed by independent GWA studies, would lead to a discovery of a novel class of biomarkers of SLE – IgG \( N \)-glycans – which could possibly extend their usefulness in prediction of other autoimmune disorders, cancer and neuropsychiatric disorders, through the same mechanism.

To test this hypothesis, we measured IgG \( N \)-glycans in 101 SLE cases and 183 matched controls (typically two controls per case), recruited in Trinidad (see materials and methods for further details). Table 4 shows the results of the measurements: for 10 of 12 \( N \)-glycan traits chosen on a basis of the experiments in mice (Table 3). The entire dataset for all glycans can be found in Table S8. There was a statistically significant difference \( (p<0.05) \) between SLE cases and controls, which was generally not the case with other groups of \( N \)-glycans (data not shown). Moreover,
the significance of the difference was striking in some cases, e.g. $p < 10^{-14}$ for IGP48, $p < 10^{-13}$ for IGP8, and $p < 10^{-9}$ for IGP64. Furthermore, the differences in the direction of effect in mice were strikingly preserved in humans (Table 4). The most significant differences observed across all 77 IgG N-glycans measurements between SLE cases and controls (Table 4) were overlapping well with the 12 N-glycan groups that were significantly changed in functional experiments in $Ikzf1$ knock-out mice.

To strengthen our findings and control for possible bias, we repeated the analysis excluding all the cases on corticosteroid treatment at the time of interview (77/101) and subsequently all the cases that were not on corticosteroid treatment at the time of interview (24/101). Although the power of the analysis decreased due to reduced number of cases, the results did not change and they remained highly statistically significant. We also hypothesized that the observed glycan changes may not be specific to SLE, but...
may be caused by corticosteroid treatment, or secondary to any inflammatory process. For this reason, and in SLE cases only, we investigated whether corticosteroid treatments and/or CRP measurements, were associated with IgG N-glycan traits. Analysis for CRP was repeated with CRP treated as a binary variable (with cut-off value at 10 mg/L). In all these analyses, the initial results held and were not changed: the association of IgG N-glycans and SLE remained striking, while the association with corticosteroid treatment and CRP was not (Table S9). Finally, we also repeated the analysis adjusting for percent African admixture, as it has been reported that SLE in Afro-Caribbean population is associated with African admixture [69]. However, this adjustment only had a minor and non-systemic effect on the previous results, and the reported observations remained.

We then validated biomarker potential of IGP48, the IgG N-glycan trait most significantly associated with SLE status, in prediction of SLE in 101 cases and 183 matched controls. We used the PredictABEL package for R (see materials and methods) [70]. As shown in Figure 3, age, sex and African admixture did not have any predictive power for this disease, but addition of IGP48 substantially increased sensitivity and specificity of prediction, with area under receiver-operator curve (AUC) increasing from 0.515 to 0.842 (0.791–0.842; 95% confidence interval (CI)). This result, we split our dataset with SLE cases and controls into a "training set" (2/3; 67 cases and 122 controls) and "test set" (1/3; 34 cases and 61 controls). Area under ROC-curve (AUC) was calculated for the test dataset. The whole process was repeated 1000 times, to allow computation of the mean AUC (and 95% CI) in the test datasets. Mean AUC was virtually unchanged compared to AUC obtained when using the complete dataset and no training, which suggests that the predictive power of IGP48 on SLE is very robust.

**Discussion**

This study clearly demonstrates that the recent developments in high-throughput glycomics and genomics now allow identification of genetic loci that control N-glycosylation of a single plasma protein using a GWAS approach. This progress should allow many similar follow-up studies of genetic regulation of N-glycosylation of other important plasma proteins, thus bringing unprecedented insights into the role of protein glycosylation in systems biology. As a prelude to this discovery, we recently reported the results of the first GWA study of the overall human plasma N-glycome using the HPLC method. Although the study was of a comparable sample size (N=2000), it only identified genome-wide associations with two glycosyltransferases and one transcription factor (HNF1α) [71]. We believe that the power of our initial study was reduced because N-glycans in human plasma originate from different glycoproteins where they have different functions and undergo protein-specific, or tissue-specific glycosylation. In this study the largest percentage of variance explained by a single association was 16–18% whereas as in the N-glycan study this was 1–6%. Furthermore, concentrations of individual glycoproteins in plasma vary in many physiological processes, introducing substantial "noise" to the quantitation of the whole-plasma N-glycome.

In this study we avoided both problems by isolating a single protein from plasma (IgG), which is produced by a single cell type (B lymphocytes), thus effectively excluding differential regulation of gene expression in different tissues, and the “noise” introduced by variation in plasma IgG concentration and by N-glycans on other plasma proteins. The only remaining “noise” in our system was the incomplete separation of some glycan structures (which co-eluted from the UPLC column) and the presence of Fab glycans on a subset of IgG molecules, but for the majority of glycan structures this "noise" was well below 10% [19]. We expected that the specificity of our phenotype and precision of the measurement provided by novel UPLC and MS methods should substantially increase the power of the study to detect genome-wide associations. Prior to analysis we could not predict which quantitation method would work better in GWA study design (UPLC vs. MS), so we used them both, each in one separate cohort of comparable sample size (N~2000).

The UPLC method yielded many more, and much stronger, genome-wide association signals in comparison to our previous study of the total plasma N-glycome in virtually same sample set of examinees [27,71]. Sixteen loci were identified in association with glycan traits with p-values<5×10^-8 and nine reached the strict genome wide threshold of 2.27×10^-9. The parallel study in the LLS cohort using MS quantitation has independently identified two of those 16 loci, showing genome-wide association with N-glycan traits. MS quantitation also allowed us to replicate 6 further loci identified in the discovery analysis, using comparable N-glycan traits measured by the two methods. However, in this follow-up analysis we were unable to replicate associations for the remaining 8 loci. This was not unexpected, because those glycosylation traits correspond to different fucosylated glycans; since fucosylation was not quantified by MS, the association between glycans measured by MS and those regions should not be expected.

Among the nine loci that reached genome-wide statistical significance, four involved genes encoding glycosyltransferases known to glycosylate IgG (ST6GALI, B4GALT1, FUT8, MGAT3). The enzyme beta1,4-galactosyltransferase 1 is responsible for the

---

**Figure 3. Validation of biomarker potential of IGP48 IgG N-glycan percentage in prediction of Systemic Lupus Erythematosus (SLE) in 101 Afro-Caribbean cases and 183 matched controls.** As shown in the graph, age and sex do not have any predictive power for this disease, but addition of IGP48 substantially increases sensitivity and specificity of prediction, with area under receiver-operator curve increased to 0.828. doi:10.1371/journal.pgen.1003225.g003
addition of galactose to IgG glycans. Interestingly, variants in \textit{B4GALT1} gene did not affect the main measures of IgG galactosylation, but rather differences in sialylation and the percentage of bisecting GlcNAc. These associations are still biologically plausible, because galactosylation is a prerequisite for sialylation, and enzymes which add galactose and bisecting GlcNAc compete for the same substrate \cite{72}. A potential candidate for \textit{B4GALT1} regulator is \textit{IL6ST}, which codes for interleukin 6 signal transducer, because it showed stronger associations with the main measures of IgG galactosylation than \textit{B4GALT1} itself. Molecular mechanisms behind this association remain elusive, but early work on \textit{IL6} (then called \textit{PHGF}) suggested that it may be relevant for glycosylation pathways in B lymphocytes \cite{73}.

Core-fucosylation of IgG has been intensively studied due to its role in antibody-dependent cell-mediated cytotoxicity (ADCC). This mechanism of killing is considered to be one of the major mechanisms of antibody-based therapies against tumours. Core-fucose is critically important in this process, because IgGs without core fucose on the Fc glycan have been found to have ADCC activity enhanced by up to 100-fold \cite{74}. Alpha-(1,6)-fucosyltransferase (fucosyltransferase 8) catalyses the transfer of fucose from GDP-fucose to N-linked type complex glycopeptides, and is encoded by the \textit{FUT8} gene. We found that SNPs located near this gene influenced overall levels of fucosylation. The directly measured IgG glycome traits most strongly associated with SNPs in the \textit{FUT8} region consisted of A2, and, less strongly, A2G1 and A2G2. These associations are biologically plausible as these glycans serve as substrates for fucosyltransferase 8. Interestingly, SNPs located near the \textit{IKZF1} gene influenced fucosylation of a specific subset of glycans, especially those without bisecting GlcNAc, and were also related to the ratio of fucosylated structures with and without bisecting GlcNAc. This suggests the \textit{IKZF1} gene encoding Ikaros as a potential indirect regulator of fucosylation in B-lymphocytes by promoting the addition of bisecting GlcNAc, which then inhibits fucosylation. The analysis of IgG glycosylation in \textit{Ikzf1} haplodeficient mice confirmed the postulated role of Ikaros in the regulation of IgG glycosylation (Table 3). The effect of \textit{Ikzf1} haplodeficiency on IgG glycans manifested mainly in the decrease in bisecting GlcNAc on different glycan structures. The increase in fucose was observed only in a subset of structures, but since very high level of fucosylation was present in the wild type mouse (up to 99.8%), a further increase could not have been demonstrated.

Nearly all genome-wide significant loci in our study have already been clearly demonstrated to be associated with autoimmune diseases, haematologic cancers, and some of them also with chronic inflammation and/or neuropsychiatric disorders. Although the literature on those associations is extensive, we tried to highlight only those associations that were identified using genome-wide association studies in datasets independent from our study. We gave prominence to associations arising from GWA studies because they are typically replicable; GWA studies have sufficient power to detect true associations, and require stringent statistical testing and replication to avoid false positive results. They have been reviewed and summarized in Table 2. The table implies abundant pleiotropy between loci that control N-glycosylation (in this case, of IgG protein) and loci that have been implicated in many human diseases. Autoimmune diseases (including SLE, RA, UC and over 80 others) are generally thought to be triggered by aggressive responses of the adaptive immune system to self antigens, resulting in tissue damage and pathological sequelae \cite{38}. Among other mechanisms, IgG autoantibodies are responsible for the chronic inflammation and destruction of healthy tissues by cross-linking Fc receptors on innate immune effector cells \cite{75}. Class and glycosylation of IgG are important for pathogenicity of autoantibodies in autoimmune diseases (reviewed in \cite{76}). Removal of IgG glycans leads to the loss of the proinflammatory activity, suggesting that in vivo modulation of antibody glycosylation might be a strategy to interfere with autoimmune processes \cite{75}. Indeed, the removal of IgG glycans by injections of Endo8 in vivo interfered with autoimmune-mediated proinflammatory processes in a variety of autoimmune models \cite{75}.

Results from our study suggest that IgG N-glycome composition is regulated through a complex interplay between loci affecting an overlapping spectrum of glycome measurements, and through interaction of genes directly involved in glycosylation and those that presumably have a “higher-level” regulatory function. SNPs at several different loci in this GWA study showed genome-wide significant associations with the same or similar IgG glycosylation traits. For example, SNPs at loci on chromosomes 9 (\textit{B4GALT1} region) and 3 (\textit{ST6GAL1} region) both influenced the percentage of sialylation of galactosylated fucosylated structures (without bisecting GlcNAc) in the same direction. SNPs at these loci also influenced the ratio of fucosylated monosialylated structures (with and without bisecting GlcNAc) in the opposite direction. SNPs at the locus on chromosome 9 (\textit{B4GALT1}), and two loci on chromosome 22 (\textit{MGAT3} and \textit{SMARCBI-DERL3} region) simultaneously influenced the ratio of fucosylated disialylated structures with and without bisecting GlcNAc. SNPs at loci on chromosome 7 (\textit{IKZF1} region) and 14 (\textit{FUT8} region) influenced an overlapping range of traits: percentage of A2 and A2G1 glycans, and, in the opposite direction, the percentage of fucosylation of agalactosylated structures.

Finally, this study demonstrated that findings from “hypothesis-free” GWA studies, when targeted at a well defined biological phenotype of unknown relevance to human health and disease (such as N-glycans of a single plasma protein), can implicate genomic loci that were not thought to influence protein glycosylation. Moreover, unexpected pleiotropy of the implicated loci that linked them to diseases has changed this study from “hypothesis-free” to “hypothesis-driven” \cite{77}, and led us to explore biomarker potential of a very specific IgG N-glycan trait in prediction of a specific disease (SLE) with considerable success. To our knowledge, this is one of the first convincing demonstrations that GWA studies can lead to biomarker discovery for human disease. This study offers many additional opportunities to validate the role of further N-glycan biomarkers for other diseases implicated through pleiotropy.

Conclusions

A new understanding of the genetic regulation of IgG N-glycan synthesis is emerging from this study. Enzymes directly responsible for the addition of galactose, fucose and bisecting GlcNAc may not have primary responsibility for the final IgG N-glycan structures. For all three processes, genes that are not directly involved in glycosylation showed the most significant associations: \textit{IL6ST-ANKRD55} for galactosylation; \textit{IKZF1} for fucosylation; and \textit{SMARCBI-DERL3} for the addition of bisecting GlcNAc. The suggested higher-level regulation is also apparent from the differences in IgG Fab and Fc glycosylation, observed in human IgG \cite{78,79} and different myeloma cell lines \cite{80}, and further supported by recent observation that various external factors exhibit specific effects on glycosylation of IgG produced in cultured B lymphocytes \cite{81}.

Moreover, this study showed that it is possible to identify loci that control glycosylation of a single plasma protein using a
GWAS approach, and to develop a novel class of disease biomarkers. This should lead to large advances in understanding of the role of protein glycosylation in the future. This study identified 16 genetic loci that are likely to be part of a much larger genetic network that regulates the complex process of IgG N-glycosylation and several further loci that show suggestive association with glycan traits and merit further study. Genetic variants in several of these genes were previously associated with a number of inflammatory, neoplastic and neuropsychiatric diseases across ethnically diverse populations, all of which could benefit from earlier and more accurate diagnosis based on molecular biomarkers. Variations in individual SNPs have relatively small effects, but when several polymorphisms are combined in a complex pathway like N-glycosylation, the final product of the pathway - in this case IgG N-glycan - can be significantly different, with consequences for IgG function and possibly also disease susceptibility. Our results may also provide an explanation for the reported pleiotropy and antagonistic genetic effects of loci involved in autoimmune diseases and hematologic cancers [39,77].

Materials and Methods

Ethics statement
All research in this study that involved human participants has been approved by the appropriate ethics committees: the Ethics Committee of the University of Split Medical School for all Croatian examinees from Vis and Korcula islands; the Local Research Ethics Committees in Orkney and Aberdeen for the Orkney Complex Disease Study (ORCADES); the University of Uppsala (Dnr 2005:325) for all examinees from Northern Sweden; the Leiden University Medical Centre Ethical Committee for all participants in the Leiden Longevity Study (LLS); and the Ethics Committee of the London School of Hygiene and Tropical Medicine for all SLE cases and controls from Trinidad. All ethics approvals were given in compliance with the Declaration of Helsinki (World Medical Association, 2000). All human subjects included in this study have signed appropriate informed consent.

Study participants—discovery and replication cohorts
All population studies recruited adult individuals within a community irrespective of any specific phenotype. Fasting blood samples were collected, biochemical and physiological measurements taken and questionnaire data for medical history as well as lifestyle and environmental exposures were collected following similar protocols. Basic cohort descriptors are included in Table S1.

The CROATIA-Vis study included 1008 Croats, aged 18–93 years, who were recruited from the villages of Vis and Komiza on the Dalmatian island of Vis during 2003 and 2004 within a larger genetic epidemiology program [82]. The CROATIA-Korcula study includes 969 Croats between the ages of 18 and 90 [83]. The field work was performed in 2007 and 2008 in the eastern part of the island, targeting healthy volunteers from the town of Korcula and the villages of Umljuba, Žrnovo and Racisce.

The Orkney Complex Disease Study (ORCADES) was performed in the Scottish archipelago of Orkney and collected data between 2005 and 2011 [84]. Data for 689 participants aged 18 to 100 years from a subgroup of ten islands, were used for this analysis.

The Northern Swedish Population Health Study (NSPHS) is a family-based population study including a comprehensive health investigation and collection of data on family structure, lifestyle, diet, medical history and samples for laboratory analyses from 2247 individuals from all three founder populations.

Genotype and phenotype quality control
Genotyping quality control was performed using the same procedures for all four discovery populations (CROATIA-Vis, CROATIA-Korcula, ORCADES, and NSPHS). Individuals with a call rate less than 97% were removed as well as SNPs with a call rate less than 98% (95% for CROATIA-Vis), minor allele frequency less than 0.02 or Hardy-Weinberg equilibrium p-value less than $1 \times 10^{-10}$. 924 individuals passed all quality control thresholds from CROATIA-Vis, 898 from CROATIA-Korcula, 889 from ORCADES and 700 from NSPHS, providing a platform for genome-wide association study of multiple quantitative traits in these founder populations.

Isolation of IgG and glycan analysis
In the discovery population cohorts (CROATIA-Vis, CROATIA-Korcula, ORCADES, and NSPHS), the IgG was isolated using protein G plates and its glycans analysed by UPLC in 2247 individuals, as reported previously [19]. Briefly, IgG glycans were labelled with 2-AB fluorescent dye and separated by hydrophilic interaction ultra-performance liquid chromatography (UPLC). Glycans were separated into 24 chromatographic peaks and quantified as relative contributions of individual peaks to the total IgG glycome. The majority of peaks contained individual glycan structures, while some contained more structures. Relative intensities of each glycan structure in each UPLC peak were determined by mass spectrometry as reported previously [19]. On the basis of these 24 directly measured “glycan traits”, additional 54 “derived traits” were calculated. These include the percentage of galactosylation, fucosylation, sialylation, etc. described in the Table S1. When UPLC peaks containing multiple traits were used to calculate derived traits, only glycans with major contribution to fluorescence intensity were used.

In the replication population cohort (Leiden Longevity Study), the IgG was isolated from plasma samples of 1848 participants. Glycosylation patterns of IgG1 and IgG2 were investigated by analysis of tryptic glycopeptides using MALDI-TOF MS. Six glycoforms per IgG subclass were determined by MALDI-TOFMS. Since the intensities of all glycoforms were related to the monogalactosylated, core-fucosylated biantennary species (glycoform B), five relative intensities were registered per IgG subclass [86].
account for errors in quantification and to remove individuals not representative of normal variation within the population. After phenotype quality control the number of individuals with complete phenotype and covariate information for the meta-analysis was 2247, consisting of 906 men and 1341 women (802 from CROATIA-Vis, 651 from CROATIA-Korcula, 415 from ORCADIES, 179 from NSPHS).

In Leiden Longevity Study, GenomeStudio was used for genotyping calling algorithm. Sample call rate was >95%, and SNP exclusions criteria were Hardy-Weinberg equilibrium p-value<10^{-4}, SNP call rate<95%, and minor allele frequency <1%. The number of the overlapping SNPs that passed quality controls in both samples was 296,619.

To combine the data from the different array sets and to increase the overall coverage of the genome up to 2.5 million SNPs, we imputed autosomal SNPs reported in the Haptype Mapping Project (release #22, http://hapmap.ncbi.nlm.nih.gov) CEU sample. Based on the SNPs that were genotyped in all arrays and passed quality control, the imputation programmes MACH (http://www.sph.umich.edu/csg/abecasis/MACH/) or IMPUTE2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) were used to obtain ca. 2.5 million SNPs for further analysis.

For replication of genome-wide significant hits identified in the discovery meta-analysis, all SNPs listed in Table S6 were used and looked up in LLS. The only exception was rs11621121, which had low imputation accuracy and did not pass quality control criteria. For this SNP, a set of 11 proxy SNPs from HapMap r.22 (all with R^2>0.85) was studied. All studied SNPs had imputation quality of 0.3 or greater.

**Genome-wide association analysis**

In the discovery populations, genome-wide association analysis was firstly performed for each population and then combined using an inverse-variance weighted meta-analysis for all traits. Each trait was adjusted for sex, age and the first 3 principal components obtained from the population-specific identity-by-state (IBS) derived distances matrix. The residuals were transformed to ensure their normal distribution using quantile normalisation. Sex-specific analyses were adjusted for age and principal components only. The residuals expressed as z-scores were used for association analysis. The “mmscore” function of ProbaBEL [87] was used for the association test under an additive model. This score test for family based association takes into account relationship structure and allowed unbiased estimations of SNP allelic effect when relatedness is present between examinees. The relationship matrix used in this analysis was generated by the “ibs” function of GenABEL (using weight = “freq” option), which uses genomic data to estimate the realized pair-wise kinship coefficient. All lambda values for the population-specific analyses were below 1.05 (Table S4), showing that this method efficiently accounts for family structure.

Inverse-variance weighted meta-analysis was performed using the MetaABEL package [http://www.genabel.org] for R. SNPs with poor imputation quality (R^2<0.8) were excluded prior to meta-analysis. Principal component analysis was performed using R to determine the number of independent traits used for these analyses (Table S10). 21 principal components explained 99% of the variance so an association was considered statistically significant at the genome-wide level if the p-value for an individual SNP was less than 2.27 x 10^{-8} and 2.27 x 10^{-8}.

Regions of association were visualized using the web-based software LocusZoom [89] to display the linkage disequilibrium (LD) of the region based on hg18/1000 Genomes June 2010 CEU data. The effect of the most significant SNP in each gene region expressed as percentage of the variance explained was calculated for each glycan trait adjusted for sex, age and first 3 principal components in each cohort individually using the “polygenic” function of the GenABEL package for R. Conditional analysis was undertaken for all significant and suggestive regions. GWAS was performed as described above with the additional adjustment for the dosage of the top SNP in the region for only the chromosome containing the association. Subsequent meta-analysis was performed as described previously and the results visualised using LocusZoom to ensure that the association peak have been removed.

In LLS, all IgG measurements were log-transformed. The score statistic for testing for an additive effect of a diallelic locus on quantitative phenotype was used. To account for relatedness in offspring data we used the kinship coefficients matrix when computing the variance of the score statistic. Imputation was dealt with by accounting for loss of information due to genotype uncertainty [90]. For the association analysis of the GWAS data, we applied the score test for the quantitative trait correcting for sex and age using an executable C++ program QTassoc [http://www.lumc.nl/uh, under GWAS Software]. For further details we refer to supplementary online information.

**Experiments in Ikzf1 knockout mice**

The Ikzf1^-/- mice harbouring the Neo-PAX5-ires-GFP knock in allele were obtained from Meinrad Busslinger (IMP, Vienna) and backcrossed to C57BL/6 mice. Both wild-type and Ikzf1^-/- animals at the age of about 8 months were subjected to retro-orbital puncture to collect blood in the presence of EDTA. Samples were centrifuged for 10 minutes at room temperature and plasma was harvested. IgG was isolated and subjected to glycan analyses.

Statistical significance of the difference in distributions of IgG glycome between wild type and the Ikzf1^-/- mice was assessed using empirical version of the Hotelling’s test. In brief, the empirical distribution of the Hotelling’s T^2 statistics was worked out by permuting the group status of the animals at random without replacement 10,000 times. This empirical distribution was then contrasted with the original value of T^2, with the proportion of empirically observed T^2 values greater than or equal to the original T^2 regarded as the empirical p-value.

**Dataset with SLE cases and matched controls**

A total of 101 SLE cases and 183 controls from Trinidad were studied. The inclusion criteria for cases and controls in Trinidad were designed to restrict the sample to individuals without Indian or Chinese ancestry. Cases and controls were eligible to be included if they were resident in northern Trinidad (excluding the southern part of the island where Indians are in the majority) and they had Christian (rather than Hindu, Muslim or Chinese) first names. Identification of cases was carried out by contacting all physicians specializing in rheumatology, nephrology and dermatology at the two main public hospitals in northern Trinidad and asking for a list of all SLE patients from their out-patient clinics. At the main dermatology clinic a register of cases since 1992 was available. Furthermore, a systematic search of: (a) outpatient records at the two hospitals, (b) hospital laboratory test results positive for auto-antibodies (anti-nuclear or anti-double-stranded DNA antibody titre >1:256) and (c) histological reports of skin biopsy examination consistent with SLE was performed. Lastly, SLE cases were also identified through the Lupus Society of Trinidad and Tobago (90% of those patients were also identified through one of the two main public hospitals). For each case, randomly chosen households in the same neighbourhood were sampled by the field team to obtain (where possible) two controls, matched with the case for sex and for 20-year age group. Cases
and controls were interviewed at home or in the project office by using a custom made questionnaire.

The case definition of SLE was based on American Rheumatism Association (ARA) criteria [91], applied to medical records (available for more than 90% of cases), and to the medical history given by the patient. Informed consent for blood sampling and the use of the sample for genetic studies including estimation of admixture was obtained from each participant. Initial case ascertainment identified 264 possible cases of SLE. Of these, 72 (27%) were excluded either on the basis of their names or because their medical history did not meet ARA criteria for the diagnosis of SLE. Of the remaining 192 individuals, 54 had incomplete addresses or were not resident in northern Trinidad, four were too ill to be interviewed, eight were aged less than 18 years and two refused to participate. For 80% (99/124) of cases, two matched controls were obtained; the response rate from those invited to participate as controls was 70%. The total sample consisted of 124 cases and 219 controls aged over 20 years who completed the questionnaire. Blood samples were obtained from 122 cases and 219 controls and DNA was successfully extracted from 93% (317/341) of these. IgG glycans were successfully measured in 303 individuals. Age at sampling was not available for 17 individuals and 2 individuals were lost due to the ID mismatch.

To test predictive power of selected glycan trait, we fitted logistic regression models (including and excluding the glycan) and used predRisk function of PredictABEL package for R to evaluate the predictive ability.

**Supporting Information**

**Figure S1** Forrest plots for associations of glycan traits measured by UPLC and genetic polymorphisms.

**Table S1** The description of 23 quantitative IgG glycosylation traits measured by UPLC and 54 derived traits.

**Table S2** Descriptive statistics of glycan traits in discovery cohorts.

**Table S3** Summary data for all single-nucleotide polymorphisms and traits with suggestive associations (p < 1 x 10^-5) with glycans measured by UPLC.

**Table S4** Population-specific and pooled genomic control (GC) factors for associations with UPLC glycan traits.

**Table S5** Description of five glycan traits measured by MS and their descriptive statistic in the replication cohort.

**Table S6** Summary data for all single-nucleotide polymorphisms with replicated in the LLS cohort.

**Table S7** IgG glycans in 5 heterozygous Ikzf1 knockout mice and 5 wild-type controls.

**Table S8** Data for all IgG N-glycans measured in 101 Afro-Caribbean cases with SLE and 183 controls (Extended Table 4 from the main manuscript).

**Table S9** Effects of corticosteroids on IgG glycans.

**Table S10** Principal component analysis of IgG glycosylation traits.

**Table S11** Description of the analysed populations.

**Acknowledgments**

The CROATIA-Vis and CROATIA-Korcula studies would like to acknowledge the invaluable contributions of the recruitment team (including those from the Institute of Anthropological Research in Zagreb) in Vis and Korcula, the administrative teams in Croatia and Edinburgh, and the people of Vis and Korcula. ORCADES would like to acknowledge the invaluable contributions of Lorraine Anderson, the research nurses in Orkney, and the administrative team in Edinburgh. SNP genotyping of the CROATIA-Vis samples and DNA extraction for ORCADES was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, WGH, Edinburgh, Scotland. SNP genotyping for CROATIA-Korcula and ORCADES was performed by Helmholtz Zentrum München, Gmbh, Neuherberg, Germany. The authors are grateful to all patients and staff who contributed to the collection in Trinidad and the UK.

**Author Contributions**

Conceived and designed the experiments: GL MW AFW PMR CH YA HC IR. Performed the experiments: JEH MP BA AM MN JK TK BS LRR. Analyzed the data: LZ OP GV VH-WU MM OS JHH-D PES MB YA. Contributed reagents/materials/analysis tools: ALP FM IK IKL FNC L AqMfC AMD QZ WW NDH UG JFW PMR. Wrote the paper: GL IR.

**References**

35. Cunninghame Graham DS, Morris DL, Bhangale TR, Criswell LA, Syvanen
66. Yashin AI, Wu D, Arbeev KG, Ukraintseva SV (2010) Joint influence of small-
35. Cunninghame Graham DS, Morris DL, Bhangale TR, Criswell LA, Syvanen


