Association between human gut microbiome and N-glycan composition of total plasma proteome

Vyacheslav A. Petrov[1, 2], Sodbo Z. Sharapov[3], Lev Shagam[1], Arina V. Nostaeva[4], Marija Pezer[5], Dalin Li[6], **Maja Hanić**[7], Dermot McGovern[6], Edouard Louis[8], Souad Rahmouni[1], Gordan Lauc[7, 9], Michel Georges[1], Yurii S. Aulchenko[3*]

[1]Unit of Animal Genomics, GIGA Research Center, University of Liège, Belgium, [2]Siberian State Medical University, Central Research Laboratory, Russia, [3]Laboratory of Glycogenomics, Institute of Cytology and Genetics, Prospekt Akademika Lavrent'yeva, 10, Novosibirsk, 630090, Russia, Russia, [4]Laboratory of Theoretical and Applied Functional Genomics, Novosibirsk State University, Novosibirsk, 630090, Russia, Russia, [5]Genos Ltd., Croatia, [6]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, United States, [7]Genos Glycoscience Research Laboratory, Croatia, [8]Laboratory of Translational Gastroenterology, GIGA Institute, University of Liège, Belgium, [9]Department of Biochemistry and Molecular Biology, Faculty of Pharmacy and Biochemistry, University of Zagreb, Croatia

Submitted to Journal:
Frontiers in Microbiology

Specialty Section:
Microbial Immunology

Article type:
Original Research Article

Manuscript ID:
811922

Received on:
09 Nov 2021

Revised on:
22 Mar 2022

Journal website link:
www.frontiersin.org

Conflict of interest statement

The authors declare a potential conflict of interest and state it below

YSA is a co-founder of PolyOmica and PolyKnomics. GL is the founder and owner of Genos Ltd. - a private research organization that specializes in high-throughput glycomic analyses and has several patents in this field, and of Genos Glycoscience Ltd. - a spin-off of Genos Ltd. that commercializes its scientific discoveries. MH and MP are employees of Genos Ltd. MP are also employed by Genos Glycoscience Ltd.

Author contribution statement

YA, MG, SR, and GL contributed to the conception and design of the study. EL, MP, MH contributed to the methodology of the study. VP, SSh, AN performed the statistical analysis. LSh and VP performed the bioinformatical analysis. DL and DMcG performed validation of the result. VP wrote the first draft of the manuscript. VP, SSh, LSh, YA, SR, and MH wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version

Keywords

Mucosal microbiome, plasma N-glycome, 16s sequencing, IgG N-glycome, Bilophila

Abstract

Word count: 154

Being one of the most dynamic entities in the human body, glycosylation of proteins fine-tunes the activity of organismal machinery, including the immune system, and mediates interaction with the human microbial consortium, typically represented by the gut microbiome. Using data from 194 healthy people, we conducted an associational study to uncover potential relations between gut microbiome and blood plasma N-glycome, including N-glycome of immunoglobulin G. While lacking strong linkages on the multivariate level, we were able to identify associations between alpha and beta microbiome diversity and blood plasma N-glycome profile. Moreover, for two bacterial genera, Bilophila and Clostridium innocuum group, significant associations with specific glycans were also shown. Our results suggest a non-trivial, possibly weak link between total plasma N-glycome and gut microbiome, predominantly involving glycans related to the immune system proteins, including immunoglobulin G. Lager studies of glycans linked to microbiome-related proteins in well-selected patient groups are required to conclusively establish specific associations.

Contribution to the field

Assemblages of host glycans play a role of an additional line of defense, protecting the host cells from the binding of pathogens and giving benefits to symbiotic bacteria. However, the glycome profile isn't stable - some microorganisms, mostly pathogenic, could recognize host glycans and invade the cell leading to masking/modification of own glycans by the host. This dynamic process could be orchestrated by the gut microbiome, potentially allowing intervention by modification of microbiota. The significance of our research is in the identification of links between microbiome and glycome and discussing the opportunity of potential glycome-modifying probiotics invention.

Funding statement

The work of SZhS and YSA was supported by a grant from the Russian Science Foundation (RSF) No. 19-15-00115; The work of AN was supported by the Ministry of Education and Science of the Russian Federation via the state assignment of the Novosibirsk State University (project "Graduates 2020"); the work of SR and MG was supported by grants from Horizon 2020 (SYSCID), EOS grant N° O018118F, PDR (FNRS): N° T.0190.19 and N° T.0096.19; the work of VAP and MP was supported by a grant from Horizon 2020 (SYSCID). The IgG N-glycome quantification was funded by a grant from the Russian Science Foundation (RSF) No. 19-15-00115.

Ethics statements

Studies involving animal subjects

Generated Statement: No animal studies are presented in this manuscript.

Studies involving human subjects

Generated Statement: The studies involving human participants were reviewed and approved by Ethics committee of the University of Liège Academic Hospital. The patients/participants provided their written informed consent to participate in this study.

Inclusion of identifiable human data

Generated Statement: No potentially identifiable human images or data is presented in this study.

Data availability statement

Generated Statement: The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA814419.

Association between human gut microbiome and N-glycan composition of total plasma proteome

Vyacheslav A. Petrov[1,2], Sodbo Zh. Sharapov[3], Lev Shagam[2], Arina V. Nostaeva[4], Marija Pezer[5], Dalin Li[6], Maja Hanić[7], Dermot McGovern[6], Edouard Louis[8], Souad Rahmouni[2], Gordan Lauc[7,9], Michel Georges[2], Yurii S. Aulchenko[3]

[1]Unit of Animal Genomics, GIGA-Institute, University of Liège, Liège, Belgium

[2]Federal State Budget Educational Institution of Higher Education, Siberian State Medical University, Ministry of Healthcare of Russian Federation, Central Research Laboratory, Tomsk, Russian Federation

[3]Laboratory of Glycogenomics, Institute of Cytology and Genetics, Prospekt Akademika Lavrent'yeva, 10, Novosibirsk, 630090, Russia

[4]Laboratory of Theoretical and Applied Functional Genomics, Novosibirsk State University, Novosibirsk, 630090, Russia

[5]Genos Ltd., Zagreb, Croatia

[6]The F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[7]Glycoscience Research Laboratory, Genos Ltd., Zagreb, Croatia

[8]Laboratory of Translational Gastroenterology, GIGA-Institute, University of Liège, Liège, Belgium.

[9]Department of Biochemistry and Molecular Biology, Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb, Croatia

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

Keywords: Mucosal microbiome, plasma N-glycome, 16S sequencing, IgG N-glycome, *Bilophila*.

Abstract

Being one of the most dynamic entities in the human body, glycosylation of proteins fine-tunes the activity of organismal machinery, including the immune system, and mediates interaction with the human microbial consortium, typically represented by the gut microbiome. Using data from 194 healthy people, we conducted an associational study to uncover potential relations between gut microbiome and blood plasma N-glycome, including N-glycome of immunoglobulin G. While lacking strong linkages on the multivariate level, we were able to identify associations between alpha and beta microbiome diversity and blood plasma N-glycome profile. Moreover, for two bacterial genera, *Bilophila* and *Clostridium innocuum group*, significant associations with specific glycans were also

shown. Our results suggest a non-trivial, possibly weak link between total plasma N-glycome and gut microbiome, predominantly involving glycans related to the immune system proteins, including immunoglobulin G. Lager studies of glycans linked to microbiome-related proteins in well-selected patient groups are required to conclusively establish specific associations.

1 Introduction

Protein glycosylation is a post-translational modification that consists of the binding of carbohydrate chains, or glycans, to the polypeptide backbone. Such modifications regulate protein activity, their half-life, and even serve as a form of cellular memory, reflecting the past and current processes in a cell, in both physiological and pathological conditions (Lauc et al. 2016). Changes in the plasma glycome profile are evident for a number of diseases, including congenital and multifactorial (Dotz and Wuhrer 2019). By affecting the activity of immunoglobulins and immune receptors(Cambay et al. 2020; Wolfert and Boons 2013), glycosylation potentially exerts influence on the interaction between the host organism and its microbiome. It was shown that the gut microbial community can itself manipulate the glycosylation profile of the enteral epithelium, co-regulating the gut homeostasis along with the host, but whether these effects remain local or extend organismal-wide is unknown (see (Kudelka et al. 2020) for a review). The present work aims to study for the first time potential links between total plasma N-glycome profile and gut mucosal microbiome composition. For that, we performed analysis of association between the gut microbiome and relative abundance of different glycans attached to blood plasma proteins (including immunoglobulin G) in a group of individuals from the CEDAR (Correlated Expression and Disease Association Research) cohort consisting in 323 well-characterized healthy individuals with intestinal biopsies (ileum, transverse colon, rectum) available (Momozawa et al. 2018).

2 Material and methods

3.1 Studied population

The analyzed population sample included 194 healthy Europeans visiting the Academic Hospital of the University of Liège as part of a national screening campaign for colon cancer. Enrolled individuals were not suffering from any autoimmune or inflammatory disease and were not taking corticosteroids or non-steroid anti-inflammatory drugs with the exception of low doses of aspirin to prevent thrombosis (Momozawa et al. 2018).

3.2 16S rRNA gene sequencing

DNA was extracted from intestinal biopsies of the ileum, transverse colon, and rectum using QIAamp DNA Stool Mini Kit (QIAgen, Germany). Three fragments of the 16S rRNA gene representing variable regions V1-V2, V3-V4, and V5-V6 were amplified independently (primer sequences are shown in Supplementary Table 4). For library preparation, a protocol of two PCR strategies for locus-specific deep sequencing was used (Jervis-Bardy et al. 2015). Sequencing of the paired-end libraries was performed on the Illumina MiSeq instrument with a 2×300bp read length.

3.3 Microbiome data processing

The reads were QV 20 trimmed from 3' end, demultiplexed, primer sequences were removed, then reads mapping to the human genome were eliminated using the BBTools suite (Bushnell B. 2014). The pipeline was constructed using Snakemake (Köster and Rahmann 2012). Further analysis was performed by QIIME 2 2018.11 (Bolyen et al. 2019). As a result, 180.5 mln paired-end reads were produced, and 156.8 mln reads were retained after a quality check. The paired-end reads were denoised and joined by the DADA2 plugin (Callahan et al. 2016) using batch-specific trimming length parameters yielding 9.1±2.0K amplicon sequence variants (ASVs) per run for V1V2, 4.5±1.6K for V3V4, and 6.8±0.67K for V5V6 amplicon. Taxonomy was assigned at a genus level to all ASVs using the q2-feature-classifier (Bokulich et al. 2018) classify-sklearn naïve Bayes taxonomy classifier against the SILVA ribosomal RNA database release 132 (Quast et al. 2013). Accordingly, we obtained 3 microbiota profiles for each of the intestinal locations.

Further analysis was performed in R language, version 3.6.1 (R Core Team 2019). Given the fact that the contamination from reagents can significantly distort the observed taxa abundance distributions as described elsewhere (de Goffau et al. 2018; Eisenhofer et al. 2019; Salter et al. 2014), we aimed to identify taxa that demonstrate abnormal behavior characteristics for contaminants. The list of taxa determined in negative controls is shown in Supplementary Table 5. We modeled taxa abundance to reveal genera that behave as contaminants taking advantage of (i) presence of biological replicates for 25 sample-location combinations, (ii) dependence of taxon abundance on sample coverage depth for some taxa, and (iii) batch effects traceable due to the presence of 9 sequencing batches. On central log-ratio transformed data (zero read counts were imputed by a minimal fraction of the taxon across all samples and locations), we revealed genera that match either of the conditions: (i) have a significant ($p<0.05$ after Benjamini-Hochberg correction) negative correlation with coverage depth, (ii) have low consistency across biological replicates (Spearman correlation $r<0.3$), (iii) have

relatively low consistency across biological replicates (r <0.4) and are not characteristic for human gut microbiota, (iv) have significant run discordance (p<0.05 after Benjamini-Hochberg correction) and are not characteristic for human gut microbiota. Run discordance and correlation with coverage depth were calculated using ANOVA of a linear model with the following explanatory variables: patient age, sex, BMI, smoking status, sample collection batch, intestinal location as well as sequencer run batch crossed with 16S rRNA amplicon nested into location. On average across locations and amplicons, 2.5% of sequencing reads were mapped to contaminant taxa revealed above that were removed from further analysis.

Only the samples with at least 10,000x (for V1-V2 and V5-V6) or 5,000x (for V3-V4) coverage were subjected to further analysis. Taxa that had <0.01% average abundance in any location/amplicon combination were eliminated. For other taxa, zero read counts were imputed by a minimal fraction of the taxon across all samples and locations. After performing a centered log-ratio (CLR) transformation, the data were corrected for technical batch effects (sequencing batch effect, amplicon, and location) using a linear mixed model implemented in the lme4 package (Bates et al. 2015):

taxon abundance ~ (Run:Amplicon)%in%Location + (1|Date.collection) + Location + Amplicon

Then, the 9 available taxa abundance distributions per sample were averaged to get one more precise measurement for each individual. Patients' age, sex, age, body mass index, and smoking status, were treated as possible covariates. To additionally refine the data, we performed PCA (ade4 package (Bougeard and Dray 2018)) and added the values of the first 4 principal components (jointly explained 24.2% of the total variance) to the covariates list.

3.4 Plasma N-glycome quantification

Plasma N-glycome quantification of CEDAR samples was performed at Genos (https://genos-glyco.com) by applying the following protocol. Plasma N-glycans were enzymatically released from proteins by PNGase F, fluorescently labeled with 2-aminobenzamide, and cleaned up from the excess of reagents by hydrophilic interaction liquid chromatography solid phase extraction (HILIC-SPE), as previously described (Akmačić et al. 2015). Fluorescently labeled and purified N-glycans were separated by HILIC on a Waters BEH Glycan chromatography column, 150 × 2.1 mm, 1.7 µm BEH particles, installed on an Acquity UPLC instrument (Waters, Milford, MA, USA) consisting of a quaternary solvent manager, sample manager and a fluorescence detector set with excitation and emission wavelengths of 250 and 428 nm, respectively. Following chromatography conditions previously described in detail (Akmačić et al. 2015), total plasma N-glycans were separated into 39

peaks. The amount of N-glycans in each chromatographic peak was expressed as a percentage of the total integrated area. Glycan peaks (GPs) — quantitative measurements of glycan levels — were defined by the automatic integration of intensity peaks on a chromatogram. The composition of major N-glycan structures in chromatographic peaks has been assigned previously (Zaytseva et al. 2020).

3.5 IgG N-glycome quantification

IgG was isolated from 10 ul of human plasma per sample using a 96-well CIM® Protein G monolithic plate (BIA Separations, Ajdovščina, Slovenia). Subsequently, IgG N-glycans were enzymatically released by incubation with PNGase F, fluorescently labeled with 2-aminobenzamide, and cleaned up by HILIC-SPE as previously described (Trbojević-Akmačić, Ugrina, and Lauc 2017). Following previously established chromatographic parameters, fluorescently labeled and purified IgG N-glycans were separated into 24 glycan peaks by HILIC on a Waters BEH Glycan chromatography column (100 × 2.1 mm, 1.7 µm BEH particles), installed on an Acquity UPLC instrument (Waters, Milford, MA, USA) (Trbojević-Akmačić, Ugrina, and Lauc 2017). The amount of N-glycans in each chromatographic peak was expressed as a percentage of the total integrated area, and their N-glycan composition had been assigned previously (Pučić et al. 2011).

3.6 Harmonization of glycan peaks

It is known that the order of the glycan peaks (GPs) on a UPLC chromatogram is similar among studies (Sharapov et al. 2019). However, depending on the cohort some peaks located near one another might be indistinguishable. In order to make the protocol of our study applicable to other cohorts and promote replicational studies, we performed harmonization of total plasma N-glycome samples using a recently developed protocol (Sharapov et al. 2019). In brief, according to the major glycostructures within the GPs we manually created the table of correspondence between different GPs (or sets of GPs) across several cohorts, where plasma glycome was measured using UPLC technology. Then, based on this table of correspondence, we defined the list of 36 harmonized GPs (listed in Supplementary Table 6) and the harmonization algorithm for each cohort, including CEDAR. Using this algorithm, the total plasma N-glycome profile of each CEDAR sample was harmonized into 36 GPs.

3.7 Normalization, batch correction of GPs, and derived trait calculation.

Normalization and batch correction was performed on harmonized UPLC glycan data. We used total area normalization (the area of each GP was divided by the total area of the corresponding chromatogram). From the 36 directly measured glycan traits, 81 derived traits were calculated (see

Supplementary Table 6). These derived traits average glycosylation features such as branching, galactosylation, and sialylation across different individual glycan structures, and consequently, they may be more closely related to individual enzymatic activity. For the original traits, CLR transformation from the “compositions" R package (van den Boogaart and Tolosana-Delgado 2008) was implemented to account for the compositional nature of the data (Galligan et al. 2013). For the derived traits, different approaches of compositional transformations were used depending on the type of the features (Supplementary Table 6). Briefly, if a derived trait represented a relative concentration of the sum of some original traits (e.g. the sum of PGP1, PGP2, and PGP3 in all of 117 traits) in the whole composition, then the derived trait was computed as the sum of these original traits followed by CLR transformation (CLR(sum(PGP1..PGP3), other traits)). If a derived trait represented the sum of original traits in some repertoire of glycans (e.g. the sum of PGP1, PGP2, and PGP3 in the first 10 traits), then at the first stage the subcomposition of this repertoire was obtained (PGP1..PGP10/sum(PGP1..PGP10)), and the second stage is similar to the previous case. Finally, if a derived trait represents the ratio between two parts of the composition, the isometric log-ratio transformation was used (Greenacre 2018).

3.8 Polygenic score derivation and Mendelian randomization

A polygenic risk score, or PRS, aggregates the effects of many genetic variants into a single number which predicts genetic predisposition for the phenotype. In the standard approach, a PRS is a linear combination of linear regression effect size estimates and allele counts at single-nucleotide polymorphisms (SNPs).

We developed PRS models using the SBayesR (Lloyd-Jones et al. 2019) method that utilized summary statistics from a genome-wide association study (GWAS). This tool reweights the effect of each variant according to the marginal estimate of its effect size, statistical strength of association, the degree of correlation between the variant and other variants nearby, and tuning parameters. Also, the SBayesR method requires a GCTB (Lloyd-Jones et al. 2019) – compatible LD matrix file computed using individual-level data from a reference population. For these analyses, we used publicly available shrunk sparse GCTB LD matrix computed from a random set of 50,000 individuals of European ancestry from the UK Biobank data set (Bycroft et al. 2018; Lloyd-Jones et al. 2019). The models were validated in CEDAR dataset, which was not part of the samples used for GWAS. The prediction accuracy was defined as the proportion of the variance of a phenotype that is explained by PRS values (R2). To calculate PRS based on the PRS model we used PLINK2 software (Chang et al. 2015), where PRS values were calculated as a weighted sum of allele counts.

Association between PRS values, acting as an instrumental variable, and microbial genera abundance were checked in linear regression (Richardson et al. 2019).

3.9 Statistical analysis

Statistical analysis was conducted in R language, version 3.6.1 (R Core Team 2019). Principal component analysis of glycome data was conducted via standard prcomp function of stats R package. Associations were examined in a linear regression model. We separately tested associations between (i) total plasma N-glycome and gut microbiome composition; (ii) beta-diversity and total plasma N-glycome; alpha-diversity and total plasma N-glycome - both (iii) the glycan traits and (iv) the first ten microbial principal components. Patients' age, gender, body mass index (BMI), smoking status were used as covariates. For the first model, the values of the first four microbial principal components were used as additional covariates. Before regression modeling, bacterial abundances were quantile-normalized via qqnorm R function.

P-values were adjusted to multiple hypothesis testing with the Sidak correction procedure. Taking into consideration possible correlations between hypotheses, the number of effective tests for Sidak correction was computed both for glycome and microbiome data. For the estimation of the number of effective tests, the approach of Galwey (Galwey 2009) implemented in the poolR package (Cinar and Viechtbauer 2020) was used. Visualization was performed with the ggplot2 package (Wickham 2009).

3 Results

To access the gut mucosal microbiome composition, biopsies were collected from consented donors who attended the department of gastroenterology of Liege University hospital in the framework of the Belgian colon cancer prevention programme. Biopsies were collected from three different locations of the gut: ileum, transverse colon, and rectum. Study participants were selected based on their health records. Exclusion criteria included autoimmune diseases and any type of inflammatory bowel diseases, cancer or polyps found during colonoscopy, antibiotics and anti-inflammatory uptake at least three weeks prior to the biopsies collection, and absence of diarrhea. Biopsies were snap-frozen and kept at -80°C until DNA extraction. The three amplicons, V1-V2, V3-V4, and V5-V6, were used to amplify microbial 16S rRNA genes. In total, nine Illumina MiSeq runs (3 amplicons x 3 gut locations) were performed on 2012 samples collected from 336 patients and 40 negative controls for sequencing. DADA2 amplicon sequence variants were analyzed by the q2-feature-classifier trained on the Silva database to assign taxonomy at the genus level. Furthermore, we measured total plasma N-

glycome for 234 CEDAR samples and 15 standard samples. From these, 230 samples passed quality control. Chromatograms for each sample were separated into 39 peaks and harmonized into 36 glycome peaks for easier comparison with other published research. Additionally, based on shared structural features, 81 derived traits were calculated. Hereafter we use "PGP_number" to refer to the originally measured and derived glycan traits, but also provide description of the glycan structures along with their Oxford notation (Harvey et al. 2009).

Metagenomic and glycomic data were simultaneously available for 194 individuals (Table. 1), thus, allowing to investigate inter-omics relationships on different levels of detalization from diversity and multivariate associations to individual linkages.

The analysis was conducted on the level of genera. After the removing of contaminants and low-abundant microorganisms 145 microbial genera were retained and used for further analyses. Among them, *Bacteroides* (ileum 34.6 (standard deviation 17.5)%, transversum 33.7 (19.0)%, rectum 31.6 (17.3)%), *Prevotella 9* (ileum 8.3 (13.0)%, transversum 9.9 (14.8)%, rectum 8.6 (12.8)%), and *Faecalibacterium* (ileum 6.0 (3.5)%, transversum 5.0 (4.3)%, rectum 5.4 (3.3)%) dominated in the microbiome of studied individuals irrespective of localization. According to the results of permutational multivariate analysis of variance interindividual variation explains beta-diversity of microbiome better than the bioptate localization ($p = 0.0001$, Fig. 1), which motivates averaging of the microbiome to obtain more precise measurement for each individual.

Univariate associations between levels of individual glycan traits and microbial genera were studied using a linear model. Before the regression analysis, the number of effective statistical tests for total plasma N-glycome and gut microbiome data were calculated. According to the effective statistical tests estimation, there were 24 effective tests in the glycome data and 87 in the microbiome data, which gives a product of 2088 independent tests. Genera abundances were normalized, adjusted for technical batch effects, known covariates as age, sex, body mass index, smoking status, and the first four microbial PC were added to the model.

Microbiome alpha-diversity was calculated with the Shannon index (Shannon, 1948). Regression analysis was performed to identify possible links between plasma glycome profile and gut microbiome diversity. Significant negative associations were found between alpha-diversity and the percentage of sialylation of core-fucosylated galactosylated structures without bisecting GlcNAc (derived trait PGP37, FGS/(FG+FGS), $p=0.041$) and the percentage of disialylation of core-fucosylated digalactosylated structures without bisecting GlcNAc (derived trait PGP43, FG2S2/(FG2+FG2S1+FG2S2), $p=0.044$) (Table. 2).

We then computed the first ten glycan PCs on 117 traits. An association between alpha-diversity and the value of the fifth glycan principal component was identified (Table. 2). This principal component had a positive correlation with the abundance of FA2B (mostly linked to immunoglobulin G (Vučković et al. 2016)) and A2G2 (mostly linked to serotransferrin (Clerc et al. 2016)) (Supplementary Table. 1). At the same time, this component was negatively correlated with glycan traits representing abundances of FA2BG2S2 (mostly attached to immunoglobulins M and A, respectively) (Clerc et al. 2016) and FA2G2S2 (attached to various N-glycoproteins, mostly secreted to the bloodstream by the liver) (Supplementary Table. 1).

To check the interplay between microbial communities and plasma glycome profile, Mantel correlation and Procrustes analysis with 9999 permutations were used. The result did not support a strong interrelation between studied omics (Mantel R=-0.014, p=0.63; Procrustes correlation=0.22, p=0.16). However, individual glycan traits associated with the microbiome of studied individuals: traits PGP43 and PGP37 were positively correlated with the microbiome-derived sixth principal component (Table. 3, Supplementary Table. 2).

In regression analysis, 981 bacterial-glycan pairs out of 16965 pairs tested, including all glycan traits and 117 out of 145 bacterial genera, had a nominal p-value below the 0.05 threshold (Supplementary Table. 3). This indicates an enrichment (p-value of binomial test = 0.0047) of the p-value distribution for significant p-values. Three bacterial-glycan pairs remained significant after adjustment for multiple testing. Specifically, we identified an association between the abundance of *Bilophila* genus and the level of FA2[3]G1 in total neutral plasma glycans (PGP62 trait, beta = 1.600 (0.278), nominal p = 4.24e-08, Sidak-corrected p = 0.00009, Fig. 2A) as well as the level of FA2[3]G1 in total plasma glycans (PGP5 trait, beta = 1.164 (0.246), nominal p = 4.44e-06, Sidak-corrected p = 0.009, Fig. 2B). The abundance of *Clostridium innocuum* group (an ASV defined on genus level) demonstrated a negative association with the ratio of disialylated and trisialylated trigalactosylated structures in total plasma N-glycans (PGP109, G3S2/G3S3, beta = -1.460 (0.331), nominal p = 1.74e-05, Sidak-corrected p = 0.036, Fig. 2C).

Additionally, univariate association analysis was performed on levels of microbial phyla and families. We estimated the number of effective statistical tests as 11 at the phylum level and 69 at the family level, which, together with the genus-level, resulted in 167 tests for microbiome data. Given 24 effective tests for the glycomic data provides an estimate of 4008 of independent tests in total. In this additional analysis, we did not identify significant associations on phylum level. However, abundance of bacterial family Tannerellaceae was shown to be negatively associated with the levels of FA2[3]G1 in total plasma glycans, percentage of neutral glycan structures and monogalactosylated structures in

total plasma glycome (Supplementary Table. 7, Supplementary Fig. 1). Identified genus-level associations remains significant after correction for additional statistical tests.

Validation of univariate findings on genus level was performed in two steps. First, as N-glycosylation of immunoglobulin G (IgG) is the main source of FA2[3]G1 in total plasma N-glycome (Clerc et al. 2016), we measured plasma IgG-glycome profiles for 192 out of 194 individuals for the technical validation of association with FA2[3]G1. Using this data, we were able to validate the association between the abundance of FA2[3]G1 in IgG-glycome and the abundance of *Bilophila* genus (beta = 1.899 (0.306), nominal p = 3.62e-09, Fig. 2D).

As an external validation dataset, microbiome and total plasma N-glycome profiles from McHardy et al. were used (McHardy et al. 2013). Given the differences in taxonomical databases used, metadata availability, and protocols of glycome and microbiome analysis between studies, it was possible to only study the association between the level of FA2[3]G1 in total plasma N-glycome and abundance of the *Bilophila*. The 47 samples for which microbiome and total plasma N-glycome were available had an expected 56% power to replicate the results. We were unable to validate this association (beta = -109.192 (174.668), nominal p = 0.53), although the sign of association was consistent.

The fact that strong and robust genetic instruments are becoming available both for total plasma (Sharapov et al. 2019; 2021) as well as for IgG (Klarić et al. 2020) N-glycomic traits open up an opportunity to investigate causal relations between plasma N-glycans and microbiome using Mendelian randomization. As instrumental variables for Mendelian randomization, we used polygenic scores computed for glycans traits that showed significant association with individual genera abundances. As a result, we found that the abundance of *Bilophila* genus was associated with polygenic score for FA2[3]G1 in total plasma glycans (PGP5 trait, beta = 0.987 (0.429), nominal p = 0.0226) as well as suggestively associated with the polygenic score for FA2[3]G1 in total neutral plasma glycans (PGP5 trait, beta = 0.025 (0.137), nominal p = 0.0663). This suggests a potentially causative link between FA2[3]G1 and the abundance of *Bilophila* genus.

4 Discussion

Overall, while our results suggest presence of association between the gut microbiota and total plasma N-glycome, this interrelation seems to be relatively weak, with the largest proportion of variance explained equal to 14.7%. The strongest associations we showed were predominantly for N-glycans (FA2B, FA2[3]G1, and FA2BG2S2) linked to immunoglobulins. Both FA2G1 and *Bilophila* abundances showed a negative correlation with the risk of UC (Hirano et al. 2018; Trbojević Akmačić

et al. 2015), which is consistent with the observed positive correlation between FA2[3]G1 glycan and *Bilophilia*.

Clostridium innocuum group showed an inverse association with the ratio of disialylated and trisialylated trigalactosylated structures in total plasma glycans. This ratio was reported to be negatively correlated with the blood level of C reactive protein - a known biomarker of inflammation (Suhre et al. 2019). *Clostridium innocuum*, the type species of the genus, treated as an unusual nosocomial agent mainly caused infections in immunodeficient patients (Crum-Cianflone 2009). It is shown that this microorganism could be linked with antibiotic-associated diarrhea and may cause colitis (Chia et al. 2018).

In conclusion, in this study of 194 healthy people, we observed several associations between plasma N-glycome and the gut microbiome. We were able to perform technical validation of our strongest finding. However, we were not able to replicate our finding in an independent dataset, perhaps due to its limited sample size (n=47, expected power 56%). Taken together, our results suggest the weak link between the gut microbiome and composition of total plasma N-glycome. Obtained results may suggest that a study of glycosylation of specific proteins, potentially connected with the microbiome, could be a more fruitful approach than an untargeted analysis performed here. One could also consider taking into account additional covariates – such as blood groups status, – that may influence both microbiome and glycome.

5 Conflict of Interest

YSA is a co-founder of PolyOmica and PolyKnomics. GL is the founder and owner of Genos Ltd. - a private research organization that specializes in high-throughput glycomic analyses and has several patents in this field, and of Genos Glycoscience Ltd. - a spin-off of Genos Ltd. that commercializes its scientific discoveries. MH and MP are employees of Genos Ltd. MP are also employed by Genos Glycoscience Ltd.

6 Funding

The work of SZhS and YSA was supported by a grant from the Russian Science Foundation (RSF) No. 19-15-00115; The work of AN was supported by the Ministry of Education and Science of the Russian Federation via the state assignment of the Novosibirsk State University (project "Graduates 2020”); the work of SR and MG was supported by grants from Horizon 2020 (SYSCID), EOS grant N° O018118F, PDR (FNRS): N° T.0190.19 and N° T.0096.19; the work of VAP and MP was supported

by a grant from Horizon 2020 (SYSCID). The IgG N-glycome quantification was funded by a grant from the Russian Science Foundation (RSF) No. 19-15-00115.

7 Data Availability Statement

The microbiome dataset generated during the current study are available in NCBI Sequence Read Archive, reference number PRJNA814419, the link to the dataset http://www.ncbi.nlm.nih.gov/bioproject/814419.

8 References

Akmačić, I. Trbojević, I. Ugrina, J. Štambuk, I. Gudelj, F. Vučković, G. Lauc, and M. Pučić-Baković. 2015. "High-Throughput Glycomics: Optimization of Sample Preparation." *Biochemistry. Biokhimiia* 80 (7): 934–42. https://doi.org/10.1134/S0006297915070123.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Bokulich, Nicholas A., Benjamin D. Kaehler, Jai Ram Rideout, Matthew Dillon, Evan Bolyen, Rob Knight, Gavin A. Huttley, and J. Gregory Caporaso. 2018. "Optimizing Taxonomic Classification of Marker-Gene Amplicon Sequences with QIIME 2's Q2-Feature-Classifier Plugin." *Microbiome* 6 (May). https://doi.org/10.1186/s40168-018-0470-z.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37 (8): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Boogaart, K. Gerald van den, and R. Tolosana-Delgado. 2008. "'compositions': A Unified R Package to Analyze Compositional Data." *Computers & Geosciences* 34 (4): 320–38. https://doi.org/10.1016/j.cageo.2006.11.017.

Bougeard, Stéphanie, and Stéphane Dray. 2018. "Supervised Multiblock Analysis in R with the Ade4 Package." *Journal of Statistical Software* 86 (1): 1–17. https://doi.org/10.18637/jss.v086.i01.

Bushnell B. 2014. "BBTools." https://sourceforge.net/projects/bbmap/.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. "The UK Biobank Resource with Deep Phenotyping and Genomic Data." *Nature* 562 (7726): 203–9. https://doi.org/10.1038/s41586-018-0579-z.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Cambay, Florian, Catherine Forest-Nault, Lea Dumoulin, Alexis Seguin, Olivier Henry, Yves Durocher, and Gregory De Crescenzo. 2020. "Glycosylation of Fcγ Receptors Influences Their Interaction with Various IgG1 Glycoforms." *Molecular Immunology* 121 (May): 144–58. https://doi.org/10.1016/j.molimm.2020.03.010.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (February): 7. https://doi.org/10.1186/s13742-015-0047-8.

Chia, J.-H., T.-S. Wu, T.-L. Wu, C.-L. Chen, C.-H. Chuang, L.-H. Su, H.-J. Chang, et al. 2018. "Clostridium Innocuum Is a Vancomycin-Resistant Pathogen That May Cause Antibiotic-Associated Diarrhoea." *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 24 (11): 1195–99. https://doi.org/10.1016/j.cmi.2018.02.015.

Cinar, Ozan, and Wolfgang Viechtbauer. 2020. *Poolr: Package for Pooling the Results from (Dependent) Tests*. https://ozancinar.github.io/poolr/.

Clerc, Florent, Karli R. Reiding, Bas C. Jansen, Guinevere S. M. Kammeijer, Albert Bondt, and Manfred Wuhrer. 2016. "Human Plasma Protein N-Glycosylation." *Glycoconjugate Journal* 33: 309–43. https://doi.org/10.1007/s10719-015-9626-2.

Crum-Cianflone, Nancy. 2009. "Clostridium Innocuum Bacteremia in an AIDS Patient: Case Report and Review of the Literature." *The American Journal of the Medical Sciences* 337 (6): 480–82. https://doi.org/10.1097/MAJ.0b013e31819f1e95.

Dotz, Viktoria, and Manfred Wuhrer. 2019. "N-Glycome Signatures in Human Plasma: Associations with Physiology and Major Diseases." *FEBS Letters* 593 (21): 2966–76. https://doi.org/10.1002/1873-3468.13598.

Eisenhofer, Raphael, Jeremiah J. Minich, Clarisse Marotz, Alan Cooper, Rob Knight, and Laura S. Weyrich. 2019. "Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations." *Trends in Microbiology* 27 (2): 105–17. https://doi.org/10.1016/j.tim.2018.11.003.

Galligan, Marie C., Radka Saldova, Matthew P. Campbell, Pauline M. Rudd, and Thomas B. Murphy. 2013. "Greedy Feature Selection for Glycan Chromatography Data with the Generalized Dirichlet Distribution." *BMC Bioinformatics* 14 (1): 155. https://doi.org/10.1186/1471-2105-14-155.

Galwey, Nicholas W. 2009. "A New Measure of the Effective Number of Tests, a Practical Tool for Comparing Families of Non-Independent Significance Tests." *Genetic Epidemiology* 33 (7): 559–68. https://doi.org/10.1002/gepi.20408.

Goffau, Marcus C. de, Susanne Lager, Susannah J. Salter, Josef Wagner, Andreas Kronbichler, D. Stephen Charnock-Jones, Sharon J. Peacock, Gordon C. S. Smith, and Julian Parkhill. 2018. "Recognizing the Reagent Microbiome." *Nature Microbiology* 3 (8): 851–53. https://doi.org/10.1038/s41564-018-0202-y.

Greenacre, Michael. 2018. *Compositional Data Analysis in Practice*. 1st edition. Boca Raton: Chapman and Hall/CRC.

Harvey, David J., Anthony H. Merry, Louise Royle, Matthew P. Campbell, Raymond A. Dwek, and Pauline M. Rudd. 2009. "Proposal for a Standard System for Drawing Structural Diagrams of N- and O-Linked Carbohydrates and Related Compounds." *PROTEOMICS* 9 (15): 3796–3801. https://doi.org/10.1002/pmic.200900096.

Hirano, Atsushi, Junji Umeno, Yasuharu Okamoto, Hiroki Shibata, Yoshitoshi Ogura, Tomohiko Moriyama, Takehiro Torisu, et al. 2018. "Comparison of the Microbial Community Structure

between Inflamed and Non-Inflamed Sites in Patients with Ulcerative Colitis." *Journal of Gastroenterology and Hepatology*, February. https://doi.org/10.1111/jgh.14129.

Jervis-Bardy, Jake, Lex E. X. Leong, Shashikanth Marri, Renee J. Smith, Jocelyn M. Choo, Heidi C. Smith-Vaughan, Elizabeth Nosworthy, et al. 2015. "Deriving Accurate Microbiota Profiles from Human Samples with Low Bacterial Content through Post-Sequencing Processing of Illumina MiSeq Data." *Microbiome* 3 (1): 19. https://doi.org/10.1186/s40168-015-0083-8.

Klarić, Lucija, Yakov A. Tsepilov, Chloe M. Stanton, Massimo Mangino, Timo Tõnis Sikka, Tõnu Esko, Eugene Pakhomov, et al. 2020. "Glycosylation of Immunoglobulin G Is Regulated by a Large Network of Genes Pleiotropic with Inflammatory Diseases." *Science Advances* 6 (8): eaax0301. https://doi.org/10.1126/sciadv.aax0301.

Köster, Johannes, and Sven Rahmann. 2012. "Snakemake—a Scalable Bioinformatics Workflow Engine." *Bioinformatics* 28 (19): 2520–22. https://doi.org/10.1093/bioinformatics/bts480.

Kudelka, Matthew R., Sean R. Stowell, Richard D. Cummings, and Andrew S. Neish. 2020. "Intestinal Epithelial Glycosylation in Homeostasis and Gut Microbiota Interactions in IBD." *Nature Reviews Gastroenterology & Hepatology* 17 (10): 597–617. https://doi.org/10.1038/s41575-020-0331-7.

Lauc, Gordan, Marija Pezer, Igor Rudan, and Harry Campbell. 2016. "Mechanisms of Disease: The Human N-Glycome." *Biochimica et Biophysica Acta (BBA) - General Subjects*, Glycans in personalised medicine, 1860 (8): 1574–82. https://doi.org/10.1016/j.bbagen.2015.10.016.

Lloyd-Jones, Luke R., Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E. Kemper, Huanwei Wang, et al. 2019. "Improved Polygenic Prediction by Bayesian Multiple Regression on Summary Statistics." *Nature Communications* 10 (1): 5086. https://doi.org/10.1038/s41467-019-12653-0.

McHardy, Ian H., Maryam Goudarzi, Maomeng Tong, Paul M. Ruegger, Emma Schwager, John R. Weger, Thomas G. Graeber, et al. 2013. "Integrative Analysis of the Microbiome and Metabolome of the Human Intestinal Mucosal Surface Reveals Exquisite Inter-Relationships." *Microbiome* 1 (1): 17. https://doi.org/10.1186/2049-2618-1-17.

Momozawa, Yukihide, Julia Dmitrieva, Emilie Théâtre, Valérie Deffontaine, Souad Rahmouni, Benoît Charloteaux, François Crins, et al. 2018. "IBD Risk Loci Are Enriched in Multigenic Regulatory Modules Encompassing Putative Causative Genes." *Nature Communications* 9 (1): 2427. https://doi.org/10.1038/s41467-018-04365-8.

Pučić, Maja, Ana Knežević, Jana Vidič, Barbara Adamczyk, Mislav Novokmet, Ozren Polašek, Olga Gornik, et al. 2011. "High Throughput Isolation and Glycosylation Analysis of IgG–Variability and Heritability of the IgG Glycome in Three Isolated Human Populations." *Molecular & Cellular Proteomics : MCP* 10 (10): M111.010090. https://doi.org/10.1074/mcp.M111.010090.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (Database issue): D590–96. https://doi.org/10.1093/nar/gks1219.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Richardson, Tom G, Sean Harrison, Gibran Hemani, and George Davey Smith. 2019. "An Atlas of Polygenic Risk Score Associations to Highlight Putative Causal Relationships across the Human Phenome." *ELife* 8 (March): e43657. https://doi.org/10.7554/eLife.43657.

Salter, Susannah J., Michael J. Cox, Elena M. Turek, Szymon T. Calus, William O. Cookson, Miriam F. Moffatt, Paul Turner, Julian Parkhill, Nicholas J. Loman, and Alan W. Walker. 2014. "Reagent and Laboratory Contamination Can Critically Impact Sequence-Based Microbiome Analyses." *BMC Biology* 12 (1): 87. https://doi.org/10.1186/s12915-014-0087-z.

Sharapov, Sodbo Zh, Alexandra S Shadrina, Yakov A Tsepilov, Elizaveta E Elgaeva, Evgeny S Tiys, Sofya G Feoktistova, Olga O Zaytseva, et al. 2021. "Replication of 15 Loci Involved in Human Plasma Protein N-Glycosylation in 4802 Samples from Four Cohorts." *Glycobiology* 31 (2): 82–88. https://doi.org/10.1093/glycob/cwaa053.

Sharapov, Sodbo Zh, Yakov A Tsepilov, Lucija Klaric, Massimo Mangino, Gaurav Thareja, Alexandra S Shadrina, Mirna Simurina, et al. 2019. "Defining the Genetic Control of Human Blood Plasma N-Glycome Using Genome-Wide Association Study." *Human Molecular Genetics* 28 (12): 2062–77. https://doi.org/10.1093/hmg/ddz054.

Suhre, Karsten, Irena Trbojević-Akmačić, Ivo Ugrina, Dennis O. Mook-Kanamori, Tim Spector, Johannes Graumann, Gordan Lauc, and Mario Falchi. 2019. "Fine-Mapping of the Human Blood Plasma N-Glycome onto Its Proteome." *Metabolites* 9 (7). https://doi.org/10.3390/metabo9070122.

Trbojević Akmačić, Irena, Nicholas T. Ventham, Evropi Theodoratou, Frano Vučković, Nicholas A. Kennedy, Jasminka Krištić, Elaine R. Nimmo, et al. 2015. "Inflammatory Bowel Disease Associates with Proinflammatory Potential of the Immunoglobulin G Glycome." *Inflammatory Bowel Diseases* 21 (6): 1237–47. https://doi.org/10.1097/MIB.0000000000000372.

Trbojević-Akmačić, I., I. Ugrina, and G. Lauc. 2017. "Comparative Analysis and Validation of Different Steps in Glycomics Studies." *Methods in Enzymology* 586: 37–55. https://doi.org/10.1016/bs.mie.2016.09.027.

Vučković, Frano, Evropi Theodoratou, Kujtim Thaçi, Maria Timofeeva, Aleksandar Vojta, Jerko Štambuk, Maja Pučić-Baković, et al. 2016. "IgG Glycome in Colorectal Cancer." *Clinical Cancer Research* 22 (12): 3078–86. https://doi.org/10.1158/1078-0432.ccr-15-1867.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Use R! New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-98141-3.

Wolfert, Margreet A., and Geert-Jan Boons. 2013. "Adaptive Immune Activation: Glycosylation Does Matter." *Nature Chemical Biology* 9 (12): 776–84. https://doi.org/10.1038/nchembio.1403.

Zaytseva, Olga O., Maxim B. Freidin, Toma Keser, Jerko Štambuk, Ivo Ugrina, Mirna Šimurina, Marija Vilaj, et al. 2020. "Heritability of Human Plasma N-Glycome." *Journal of Proteome Research* 19 (1): 85–91. https://doi.org/10.1021/acs.jproteome.9b00348.

Figures and tables

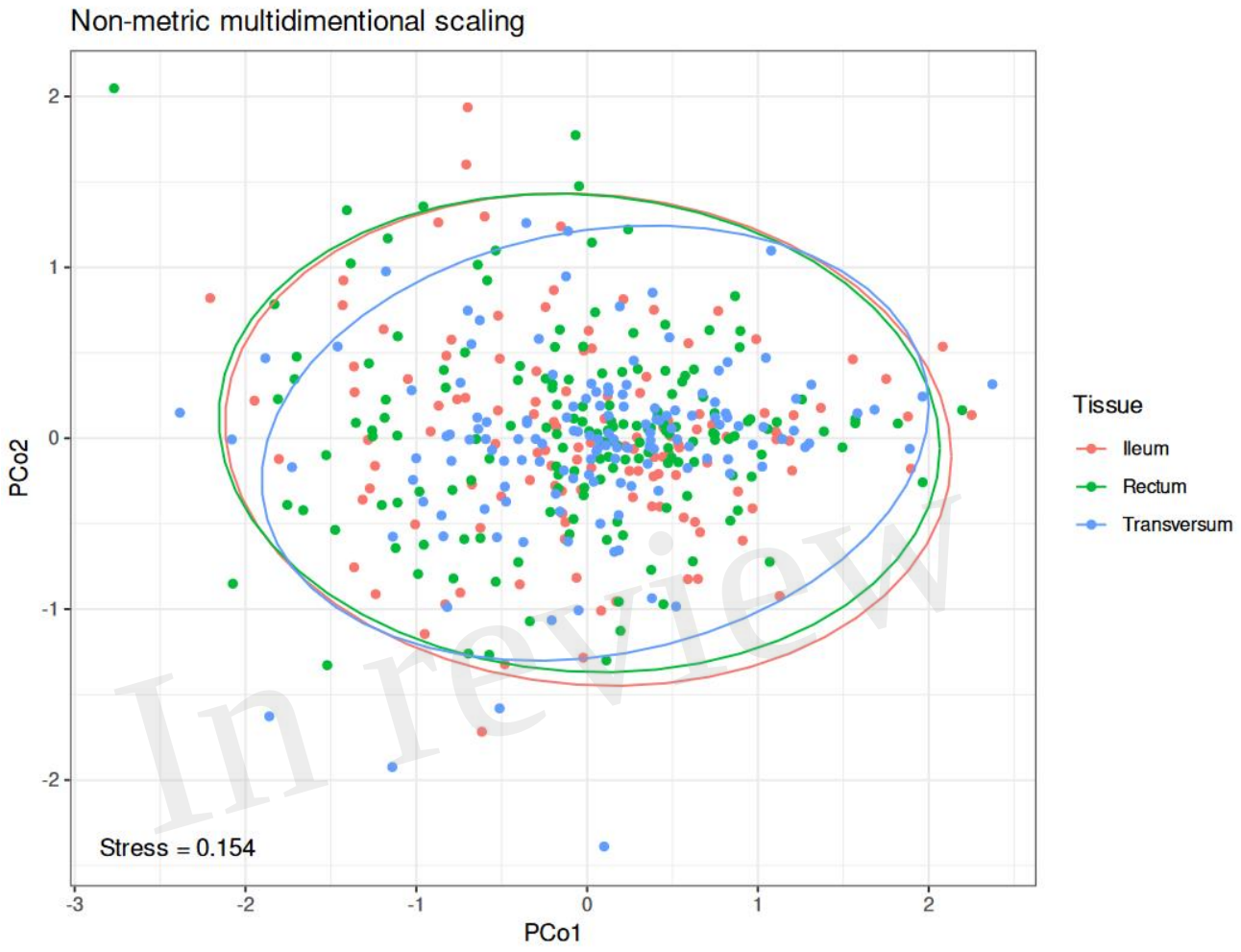


Figure 1 – non-metric multidimensional scaling of microbial abundances on genera level in Euclidean distance, the first two principal coordinates were shown. Color of dots represents microbiome samples from ileum (red), transversum (blue) and rectum (green) mucosa.

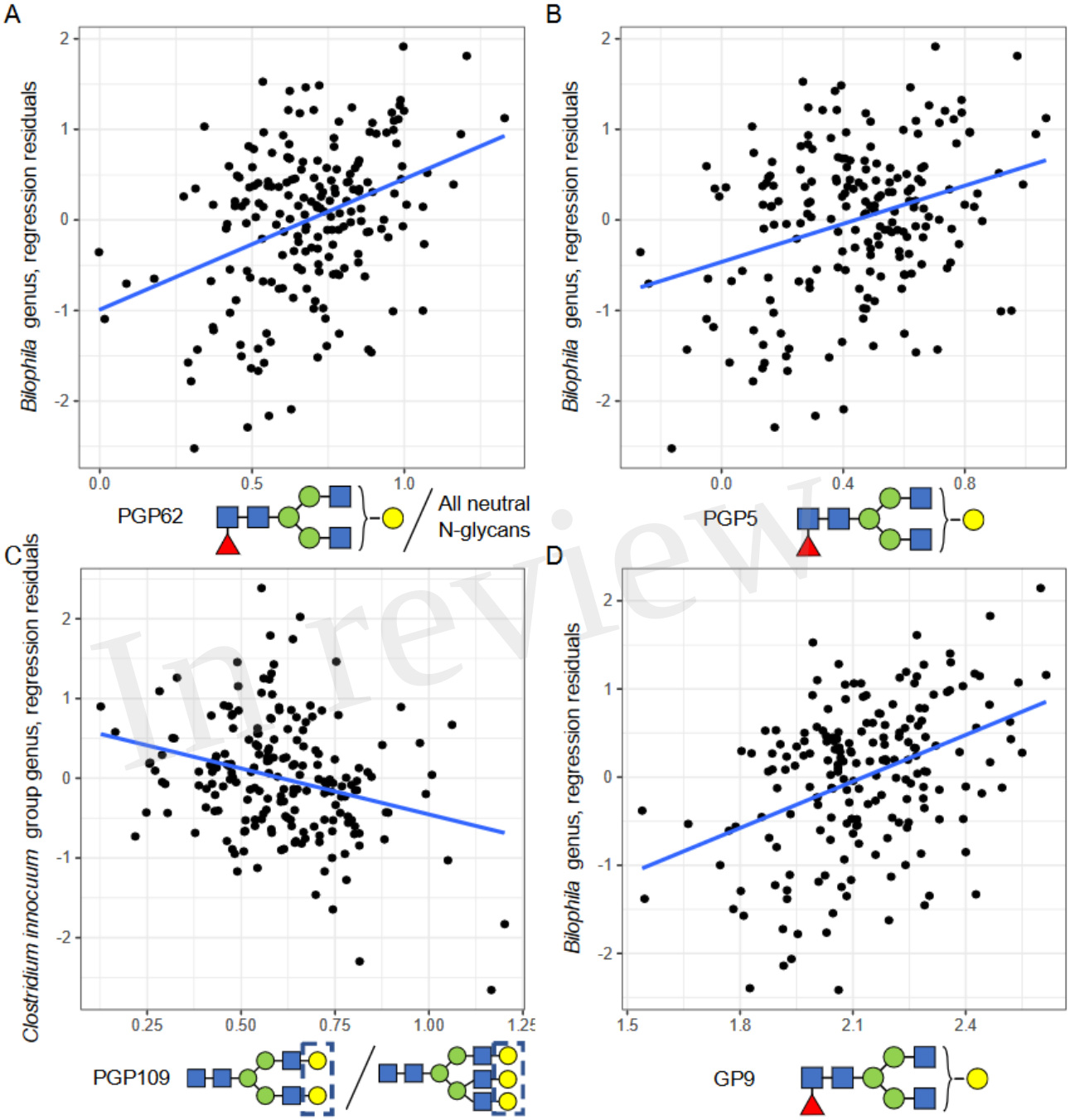
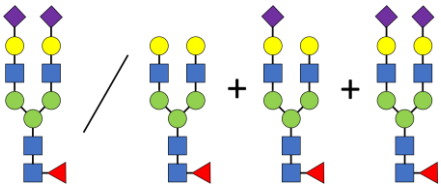
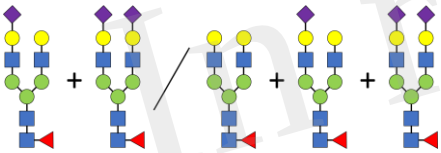


Figure 2 - univariate associations between microbial genera and glycan traits. On the plot, dots represent samples, a regression line shown in black. Panel A - association between the abundance of *Bilophila* genus and the level of FA2[3]G1 in total neutral plasma N-glycans; panel B - association between the abundance of *Bilophila* genus and the level of FA2[3]G1 in total plasma N-glycans; panel C - association between the abundance of *Clostridium innocuum* group genus and the ratio of disialylated and trisialylated trigalactosylated structures in total plasma N-glycans; panel D - technical validation of an association between IgG FA2[3]G1 glycan level and the abundance of *Bilophila* genus.

Table 1 - Demographic information of the cohort studied

Characteristic	Overall
Sample size	194
Age, mean (SD)	55.66 (13.05)
Body Mass Index, mean (SD)	26.37 (4.64)
Ethnicity, absolute n (%)	
Caucasian	159 (82.0)
Mediterranean	23 (11.9)
Mixed	12 (6.2)
Sex (males), absolute n (%)	82 (42.3)
Smoking status (smokers), absolute n (%)	45 (23.2)

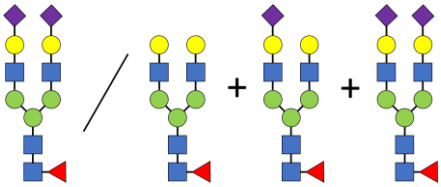
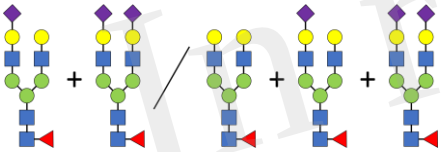
Table 2 - Association between microbiome alpha-diversity (Shannon index) and plasma total N-glycome profile

N-glycan trait	regression beta coefficient	beta standard error	nominal p-value	Sidak-corrected p-value
PGP43 (FG2S2/(FG2+FG2S1+FG2S2)) 	-1.213	0.385	0.0019*	0.0440
PGP37 (FGS/(FG+FGS)) 	-1.270	0.400	0.0018*	0.0410
Glycomic principal component 5	0.275	0.096	0.0045#	0.0440

* corrected for 24 tests that reflects the effective number of glycomic traits

corrected for 10 tests (the number of glycomic PCs tested)

Table 3 - Association between microbiome beta-diversity (Principal Component 6) and total plasma N-glycome profile

N-glycan trait	regression beta coefficient	beta standard error	nominal p-value	Sidak-corrected p-value*
PGP43 (FG2S2/(FG2+FG2S1+FG2S2)) 	2.992	0.734	6.8E-05	0.0161
PGP37 (FGS/(FG+FGS)) 	3.013	0.766	0.0001	0.0280

* The multiple testing correction was made accounting for 240 tests (24 x 10 where 24 is the effective number of glycomic traits and 10 is the number of Microbiome PCs)

Figure 1.TIFF

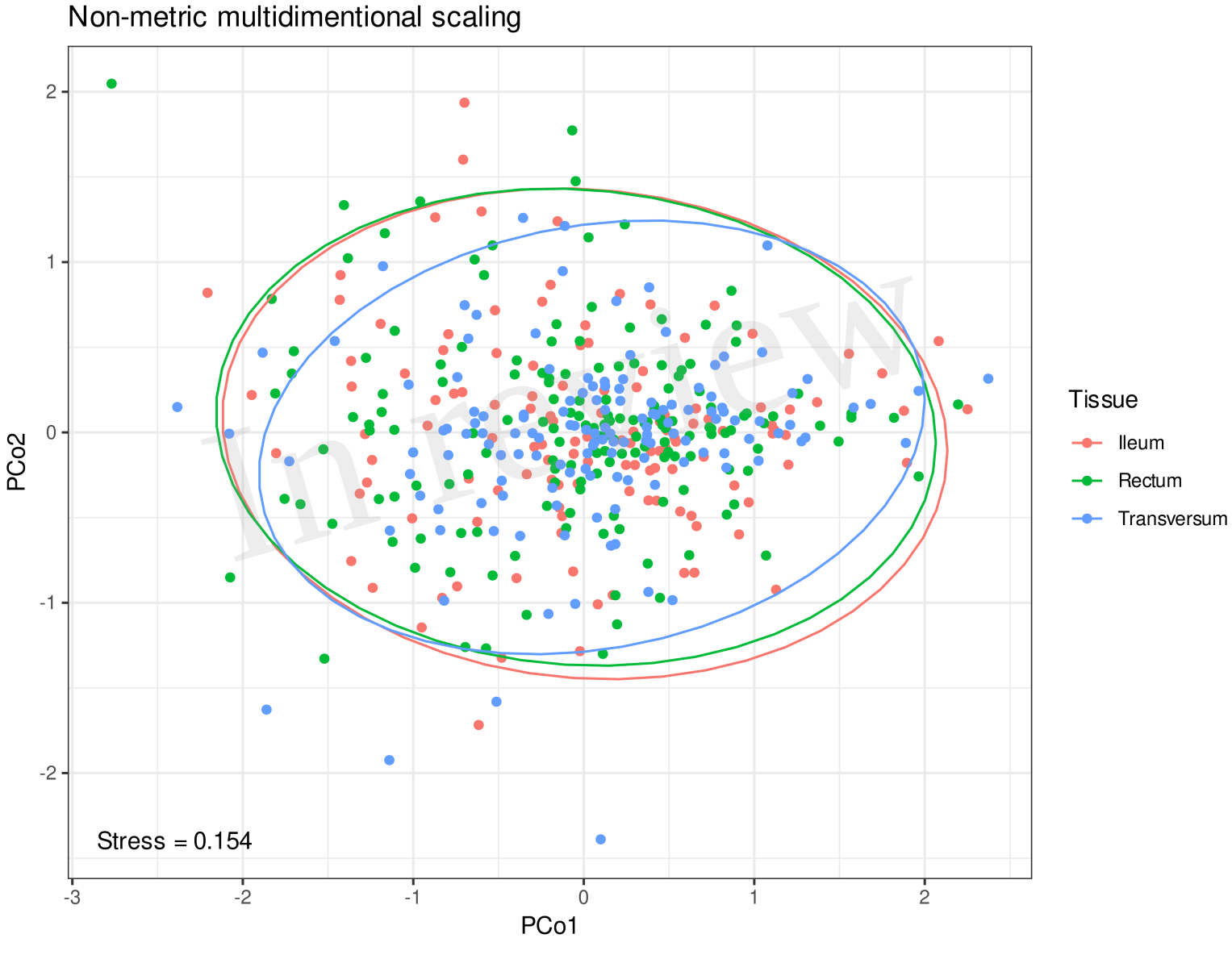


Figure 2.TIFF