Computational analysis of human plasma N-glycome and genotypes

Pedrosa Pinto, Ana Sofia

Doctoral thesis / Disertacija

2013

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Split, Faculty of Science / Sveučilište u Splitu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:166:821273

Rights / Prava: Attribution-NonCommercial-ShareAlike 4.0 International/Imenovanje-Nekomercijalno-Dijeli pod istim uvjetima 4.0 međunarodna

Download date / Datum preuzimanja: **2024-04-20**

Repository / Repozitorij:

Repository of Faculty of Science

UNIVERSITY OF SPLIT
FACULTY OF SCIENCE
PHD STUDY OF BIOPHYSICS

DOCTORAL THESIS

COMPUTATIONAL ANALYSIS OF HUMAN PLASMA N-GLYCOME AND GENOTYPES

Sofia Pinto

Zagreb, 2013

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Splitu
Prirodoslovno-matematički fakultet

Doktorski rad

RAČUNALNA ANALIZA GENOTIPOVA I N-GLIKOMA LJUDSKE PLAZME

Sofia Pinto

Rad je izrađen u:
Prirodoslovno – matematičkom fakultetu Sveučilišta u Zagrebu

Sažetak

Glikozilacija je jedna od najopsežnijih modifikacija proteina. Glikani utječu na strukturu i funkciju proteina na koje su vezani, a poznato je i da imaju važne uloge u fiziološkim i patološkim procesima. Nedostatak univerzalnog koda za sintezu glikana zajedno sa tehnološkim poteškoćama kvantifikacije glikana razlozi su ograničenom razumijevanju procesa koji reguliraju njihovu sintezu. Značajni napretci u analitičkim postupcima omogućili su pouzdane razvoj pouzdanih visoko-protočnih metoda za kvantifikaciju glikana, a time i prve studije plazma N-glikoma velikog broja ljudi.

Kako bi se istražila genomska i i okolišna regulacija glikozilacije, u ovome su radu glikanski, fiziološki i biokemijski podaci te genotipovi iz tri različite izolirane populacije analizirani različitim računalnim metodama. Identificirani su glikanski profili specifični za opću populaciju evaluiran je potencijal glikana kao biomarkera dijabetesa. Također, analizirane su asocijacije glikana i fenotipova te su istraženi glikanski, fenotipski i genotipski uzorci koji definiraju pojedine populacije. Analize polimorfizama povezanih sa glikozilacijom potvrdile su prethodna otkrića te su otkrivene nove potencijalne poveznice.

Broj stranica: 117
Broj slika: 31
Broj tablica: 8
Broj literaturnih navoda: 151
Broj priloga: 2
Jezik izvornika: Engleski
Rad je pohranjen u: Nacionalnoj sveučilišnoj knjižnici u Zagrebu, Sveučilišnoj knjižnici u Splitu, Knjižnici Prirodoslovno – matematičkog fakulteta (PMF) Sveučilišta u Splitu.

Ključne riječi: algoritmi strojnog učenja, cijelogenomske studije, dijabetes, glikomika, glikozilacija, N-glikani, selekcija SNP-ova

Mentor:
Prof.dr.sc. Kristian Vlahoviček, redoviti professor, Prirodoslovno–matematički fakultet, Sveučilište u Zagrebu

Ocjenjivači:
Prof.dr.sc. Igor Weber, znanstveni savjetnik, Institut Ruđer Bošković, Zagreb
Prof.dr.sc. Gordan Lauc, redoviti professor, Farmaceutsko-biokemijski fakultet, Sveučilište u Zagrebu
Prof.dr.sc. Andreja Ambriović-Ristov, viša znan. sur., Institut Ruđer Bošković, Zagreb

Rad prihvaćen: 27. studenog 2013

BASIC DOCUMENTATION CARD

University of Split
Faculty of Science

Doctoral thesis

COMPUTATIONAL ANALYSIS OF HUMAN PLASMA N-GLYCOME AND GENOTYPES

Sofia Pinto

Thesis performed at**:**
Faculty of Science, University of Zagreb

Abstract

Glycosylation is one of the most extensive protein modifications. Glycans influence both structure and function of the proteins and known to have important roles in physiological and pathological processes. The absence of a universal code for glycan synthesis combined with the technological challenges faced by glycan quantification analysis has hindered the knowledge about the processes regulating the assembly of glycans. Major breakthroughs in analytical procedures created the possibility to reliably quantify glycans in a high-throughput manner and allowed the first large-scale studies on human plasma N-glycome.

In order to explore the genomic and environmental regulation of glycosylation, different computational methods were employed to the integrated analysis of glycan, physiological/biochemical and genotype data in three isolated population cohorts. Specific glyco-phenotypes were identified in the general population and the potential use of glycan modifications as biomarkers was evaluated for the particular case of diabetes. General associations between glycans and phenotypes were observed and glycan, phenotypic and genotypic patterns capable of discriminating the populations were explored. The analysis of polymorphisms associated with glycosylation was addressed replicating previous findings and suggesting possible novel associations.

Number of pages: 117
Number of figures: 31
Number of tables: 8
Number of references: 151
Number of appendices: 2
Original in: English
Thesis deposited at: the National and University Library in Zagreb, the University Library in Split, the Library of the Faculty of Science of the University of Split.

Keywords: diabetes, genome-wide association studies, glycomics, glycosylation, machine learning algorithms, N-glycans, SNP selection

Supervisor:
Prof.dr.sc. Kristian Vlahoviček, Faculty of Science, University of Zagreb

Reviewers:
Prof.dr.sc. Igor Weber, Institute Ruđer Bošković, Zagreb
Prof.dr.sc. Gordan Lauc, Faculty of Pharmacy and Biochemistry, University of Zagreb
Prof.dr.sc. Andreja Ambriović-Ristov, Institute Ruđer Bošković, Zagreb

Thesis accepted: 27 November 2013

To my parents and grandparents

and

in memory of São.

ACKNOWLEDGMENTS

...to those people living in a small beautiful country by the Adriatic sea that I have been discovering for 6 long years:

I would like to express my gratitude to my academic advisor, *Professor Kristian Vlahoviček* for first accepting me in a three-month trial into his group and later giving me the opportunity to extend my stay and work. I greatly appreciate his guidance and wise advice during all these years. I would also like to thank him for giving me confidence (or at least trying to!) in my own abilities and for sharing his immense knowledge about the bioinformatics world.

I would like to thank my thesis committee members: *Professor Igor Weber* and *Professor Andreja Ambriović-Ristov* for all the help they provided during the course of my studies; and *Professor Gordan Lauc* for kindly providing the data analysed in this thesis.

The writing of this thesis has benefited from the comments, suggestions and proofreading of several people. I am deeply grateful to *Rosa Karlić* for her thorough reading of this thesis and all of the wise comments and constructive corrections which greatly improved the thesis. I am also indebted to *Lucija Klarić* who kindly translated the abstracts to Croatian language with exceptional competence. Appreciation also goes to *Tina Kokan* and *Ivan Franulović* for their suggestions and found typos. A special thanks to the research collaborators from the Genos group, *Maja Pučić*, *Frano Vučković* and *Lucija Klarić* for their ready and helpful answers and explanations to my glyco- and mathematical-related questions.

I would furthermore like to acknowledge *Petar Jager* for his assistance and incomparable proficient skills in solving unwanted and unexpected computer and technological-related issues (including the ones I sometimes imprudently created myself!) encountered while sitting in front of the computer.

A huge THANK YOU to the Bioinfo group for all the fun moments as well as the depressing ones spent together in and out of the office! To *Tina*, *Vedranex* and *Jager*, my sincere thanks for your friendly welcome when I first step into the office and your inestimable help ever since, it meant a lot to me.

To all friends I have made during my stay in Croatia, 'hvala' for the enjoyable and unforgettable time I have spent in Zagreb.

Last but not the least, to *Ivan*, thank you for going through all the ups and downs with me, for listening to my complaints and, most of all, for your endless patience and understanding. A special word of appreciation also goes to his family for their wonderful role as my Croatian family.

...to the people living in a small beautiful country by the Atlantic Ocean which I temporarly left during 6 long years but never forgot:

First and foremost, my sincere gratitude to my family for the unconditional trust and love and all the supportive words. I would especially like to thank my parents and grandparents for all they taught me and for believing in me even when I did not.

My greatest appreciation to those Friends that were always there for me no matter what. I am deeply thankful for their constant encouragement and, most of all, for their trustworthy and invaluable friendship in all times. Thank you for all the great and memorable moments which were a source of strength and laugh 2500km away! To all of you...SAUDADES...

To *Pipe*, *Nini*, *Pedrocas* and *Dinocas*, thank you for your warm smile, for making me smile and for the funny drawings!

TABLE OF CONTENTS

ACKNOWLEDGMENTS ..... **i**

LIST OF FIGURES ..... **vi**

LIST OF TABLES ..... **viii**

LIST OF ABBREVIATIONS ..... **ix**

1. INTRODUCTION **1**

1.1. From Glycans to the Glycome and Glycomics 1

1.1.1. Glycans 1

1.1.2. Glycome 1

1.1.3. Glycomics 2

1.2. Glycosylation 3

1.2.1. Principles overview 3

1.2.2. Protein glycosylation 4

1.2.3. Structural basis of glycan and glycoprotein diversity 4

1.3. N-glycosylation and N-glycans in Eukaryotes 5

1.3.1. Synthesis of N-glycans 5

1.3.2. Structure of N-glycans 8

1.3.3. Diversity of N-glycans and N-glycoproteins 9

1.3.4. Biological roles of N-glycans in health and disease 10

1.4. Human Plasma N-glycome 13

1.4.1. Challenges of structural analyses of N-glycans 13

1.4.2. Variability, heritability and stability of N-glycans 14

1.4.3. Potential diagnostic value of the N-glycome 16

1.4.4. Genome and Glycome-wide association studies 17

1.5. Immunoglobulin G: an N-linked glycoprotein 20

1.5.1. Structure and function of IgG 20

1.5.2. Glycosylation of IgG 22

1.5.3. Structural analyses of IgG 23

1.6. Aim and objectives of the thesis 24

2. MATERIALS & METHODS 26

2.1. Study Populations 26

2.2. N-glycan Quantification Analysis 26

2.2.1. Plasma N-glycans 26

2.2.2. IgG N-glycans 30

2.3. Feature Data Sets 31

2.3.1. Plasma N-glycan profile data 31

2.3.2. IgG N-glycan profile data 32

2.3.3. Phenotype data 32

2.3.4. Genotype data 33

2.4. Data Preprocessing 33

2.4.1. Data quality control 34

2.4.2. Data integration 35

2.4.3. Data normalization 36

2.4.4. Data correction 36

2.4.5. Data removal of outliers 38

2.4.6. Data imputation 38

2.4.7. Data comparison 38

2.5. Computational Tools 39

2.5.1. R statistical package 39

2.5.2. PLINK 39

2.5.3. Perl programming language 39

2.6. Computational Methods/Algorithms 40

2.6.1. Nearest neighbours computation 40

2.6.2. Clustering 40

2.6.3. Principal component analysis and partial least squares regression 43

2.6.4. Discriminant analysis of principal components 44

2.6.5. Random Forests and Random Jungle 46

2.6.6. Correlation adjusted scores 49

2.6.7. Genome-wide efficient mixed model association 50

2.7. Statistical Methods 50

3. RESULTS 52

3.1. Data Preprocessing/Analysis Pipeline 52

3.2. Common aberrations from the normal human plasma N-glycan profile 54

3.3. Analysis of clustering patterns inside populations 56

3.4. Correlation between N-glycome and phenotypic traits 60

3.5. Comparison of feature profiles from diabetes groups 66

3.6. Comparison of feature profiles from isolated populations 75

3.7. N-glycome association studies 89

4. DISCUSSION 92

4.1. Glyco-phenotypes in the general population 92

4.2. Internal clustering structure of isolated populations 93

4.3. Association between N-glycans and phenotypes 94

4.4. Diabetes: a case example 97

4.5. Population-specific patterns 100

4.6. Association between N-glycans and genotypes 102

5. REFERENCES 107

CURRICULUM VITAE 118

SAŽETAK 119

ABSTRACT 120

APPENDIX A. Supplementary Figures 121

APPENDIX B. Supplementary Tables 142

LIST OF FIGURES

Figure 1. Biosynthesis of N-glycans. 7

Figure 2. Major structural classes of mature N-glycans in eukaryotes. 9

Figure 3. Immunoglobulin G structure. 21

Figure 4. Example of a chromatographic peak division for glycan quantification. 28

Figure 5. Plasma N-glycome chromatographic and quantification analysis. 29

Figure 6. IgG N-glycome chromatographic and quantification analysis. 31

Figure 7. Principles of the affinity propagation algorithm. 42

Figure 8. Fundamental difference between PCA and DA. 45

Figure 9. Diagram of the Random Forest algorithm. 47

Figure 10. IgG glycan quantification measurements by solution method versus gel method. 54

Figure 11. Normal and aberrant plasma N-glycan profiles. 55

Figure 12. Affinity propagation clustering results based on phenotype data for the pooled data of populations. 58

Figure 13. Affinity propagation clustering results based on plasma glycan profiles for the pooled data of populations. 60

Figure 14. Statistically significant correlations between plasma glycans and phenotypes for all populations. 62

Figure 15. Statistically significant correlations between IgG glycans and phenotypes for all populations. 64

Figure 16. Correlation of 11 plasma peaks and their corresponding IgG peaks with phenotypes. 65

Figure 17. Parallel coordinates plots of plasma glycan, IgG glycan and phenotype profiles for non-diabetic, pre-diabetic and diabetic groups. 67

Figure 18. Wilcoxon sum rank test p-values for the pairwise comparison of the diabetes groups with regard to phenotype data. 68

Figure 19. PLS-DA and PCA analysis of the diabetes groups using phenotype data. 70

Figure 20. Random Forest variable importance for the classification of the diabetes groups. 72

Figure 21. Genetic context of polymorphisms possibly associated with the diabetes condition. 75

Figure 22. Parallel coordinates plots of plasma glycan, IgG glycan and phenotype profiles for Vis, Korčula and Orkney populations. 76

Figure 23. Wilcoxon sum rank test p-values for the pairwise comparison of the population cohorts with regard to plasma, IgG and phenotype data. 77

Figure 24. PLS-DA and PCA analysis of the population cohorts using plasma glycans data. 78

Figure 25. Random Forest variable importance for the classification of the population cohorts. 81

Figure 26. DAPC analysis of the population cohorts using genotype data. 84

Figure 27. Genotype frequencies of the 15 SNPs most contributing to the first discriminant component of the DAPC analysis of the population cohorts. 85

Figure 28. Ranking comparison of the most important SNPs consistently identified by the three investigated approaches in the analysis of the genetic structure of populations. 87

Figure 29. Genetic context of the most important SNPs consistently identified by the three investigated approaches in the analysis of the genetic structure of populations. 89

Figure 30. Histograms of SNPs ranking by method. 90

Figure 31. Proportion of the variance of all traits of the three feature data sets explained by genotype. 91

LIST OF TABLES

Table 1. Data summary for the study population cohorts. 35

Table 2. Data summary for the diabetes data set..... 36

Table 3. Random Jungle algorithm parameters..... 49

Table 4. Random Forest confusion matrices for the classification of the diabetes groups. 71

Table 5. Random Forest confusion matrices for the classification of the population cohorts. 80

Table 6. Random Jungle confusion matrices for the classification of population cohorts based on genotype data. 83

Table 7. Glycan-related SNPs present among the most contributing SNPs for the genetic structure of populations. 86

Table 8. Genetic variants implied to be associated with plasma glycan traits. 104

LIST OF ABBREVIATIONS

2-AB	2-aminobenzamide
Asn	Asparagine
BMI	Body mass index
BSLMM	Bayesian Aparse Linear Mixed Model
DA	Discriminant Analysis
DAPC	Discriminant Analysis of Principal Components
ER	Endoplasmic reticulum
Fab	Antigen-binding fragment of Immunoglobulin G
Fc	Crystallizable fragment of Immunoglobulin G
GlcNAc	N-Acetylglucosamine
GWAS	Genome-wide association studies
HbA1c	Glycated haemoglobin
HDL	High-density lipoprotein
HILIC	Hydrophilic interaction liquid chromatography
HPLC	High performance liquid chromatography
IgG	Immunoglobulin G
LDL	Low-density lipoprotein
LMM	Linear Mixed Models
PCA	Principal Component Analysis
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis
RF	Random Forests
RJ	Random Jungle
Ser	Serine
SNP	Single-nucleotide polymorphism
Thr	Threonine
UPLC	Ultra performance liquid chromatography
WAX	Weak anion exchange

“the beauty of a glycan is not the sugars that go into it,
but the way those sugars are put together”
(from GlycoBase)

1. INTRODUCTION

1.1. From Glycans to the Glycome and Glycomics

Investigating the structure, biosynthesis and biological function of glycans has been the focus of the field of glycobiology. Glycobiology owes its fast expansion and growth to the development and continuous improvement of technological approaches aimed to explore the structural complexity of glycans.

1.1.1. Glycans

Glycans are considered to be the most abundant and diverse biopolymers formed in nature and they constitute one of the four major building blocks of cells, together with proteins, nucleic acids and lipids. Through the years, advances in glycobiology have revealed several essential roles played by glycans at the molecular level, extending the initial view of glycans as simple structural components and sources of energy in a cell.

Glycans are chains of monosaccharides (or simple sugars) which have variable length from a few sugars to several hundred. Glycans are the product of a series of stepwise reactions involving the complex interaction of hundreds of enzymes and transcriptional factors and can be found in free form or as glycoconjugates when attached to another molecule, usually a protein or a lipid.

1.1.2. Glycome

The spectrum of all glycans and glycoconjugates in an organism forms the glycome which is estimated to be much larger than the proteome itself (Lee *et al.*, 2005). The interrelated and complicated pathways participating in the synthesis of glycans and the various linkages allowed between monosaccharides, together contribute to the structural and functional diversity presented by the glycome.

The glycome has a role in both normal physiology and disease and its importance in molecular biology should be regarded at the same level as the one of the proteome or transcriptome. The fact that glycans are an essential part in the function of physiologic systems highlights the need to include and take them into account in research analyses of biological systems and processes whenever possible. The diversity of the glycome might hold necessary information to link biological theories or uncover new findings.

1.1.3. Glycomics

Research in the fields of genomics, proteomics and transcriptomics has enabled unquestionably important discoveries. Nonetheless, the view of human physiology is far from being complete and many explanations regarding the different mechanisms and processes occurring in biological systems are still lacking. The new emerging field of glycomics is believed to contain a great deal of significant information that can help filling in the existing biological gaps as well as to give additional insights into the current view of biological processes.

The concept of glycomics has emerged in reference to the glycome and concerns the systematic study of genetic, physiologic and pathologic aspects of the glycome expressed by specific cells, tissues or organisms in order to elucidate the factors regulating the synthesis of glycans and the association of glycans with biological processes.

In comparison to the analogous terms of genomics and proteomics, glycomics is a much more recent discipline and its achievements are far beyond those attained in genomics and proteomics. Increased understanding of the functions of glycans and of the importance of glycosylation has led to a growing interest in glycomics which has contributed to its own development. The limited and late development of glycomics is due to challenges unique to glycan analysis. The experimental and analytical methodologies inherent to glycomics have undergone several improvements which have simplified the procedures and decreased the time required for glycan analysis as well as enabled the analysis of a larger number of samples. Such improvements have allowed the first large-scale studies involving glycan structures and brought glycomics into line with the 'omics' approaches of genomics, proteomics and transcriptomics.

Nonetheless, the multiplicity of questions asked by all areas of glycobiology cannot be adequately addressed by analysing only the glycan moieties isolated from glycoproteins. For certain glycan-related subjects, such as changes in protein properties, bacterial binding and antigenicity specificities or therapeutic efficacy of glycoproteins, the analysis of intact glycoproteins or glycopeptides is required (Marino *et al.*, 2010). The wide range of glycoprotein analyses and the questions behind them developed into a field of its own – glycoproteomics – which is rapidly growing in parallel with glycomics. Despite having slightly different goals, glycomics and glycoproteomics complement each other and contribute to the main scope of elucidating the complex regulation of glycosylation. Although the overview of objectives and applications of glycoproteomics is out of scope of this thesis, it seemed worth to briefly mention

its importance and its coexistence with glycomics (for a review of glycopeptide analysis and glycoproteomics applications, see Wei & Li (2009) and Dallas *et al.* (2013)).

In glycobiology, as in most areas of science, the question to be answered dictates the type of glycosylation analysis strategy to be employed which should be selected to adequately suit the needs of the study.

1.2. Glycosylation

1.2.1. Principles overview

Glycosylation is an enzymatic process through which glycans are concurrently synthesized and typically attached to proteins and lipids producing glycoproteins and glycolipids, respectively.

Contrary to protein synthesis, where a single gene codes for a protein, there is no universal code for the structure of glycans. Glycan synthesis is not template driven but rather encoded by a complex network of glycotransferases, glycosidases, transcription factors, transporters and other proteins (Lauc *et al.*, 2010b). It is estimated that 1% of genes in mammalian genome participate in glycan formation and modification (Lowe & Marth, 2003).

The numerous enzymes and factors involved in glycan synthesis cooperate in an organized manner in stepwise reactions which lead to the final glycan structure. The resulting glycan moeities are assembled from only nine monosaccharides: glucose (Glc), galactose (Gal), fucose (Fuc), mannose (Man), xylose (Xyl), N-acetylglucosamine (GlcNAc), N-acetylgalactosamine (GalNAc), iduronic acid (IdoA) and sialic acid (SA) (Moremen *et al.*, 2012). Although the number of available monosaccharides may not appear sufficient to accomplish the claimed diversity of the glycome, a variety of combinations can be formed by establishing different glycosidic linkages between monosaccharides. In this way, a modest number of monosaccharides is able to generate a vast repertoire of glycan variants.

The coordination of glycosylation mechanisms is crucial for the accurate synthesis of glycans which were shown to be essential factors in the maintenance of an organism's homeostasis (Ohtsubo & Marth, 2006). Dysregulation of glycosylation pathways has been associated with several diseases, such as cancer and diabetes as well as cardiovascular, congenital and immunological disorders.

Investigating the behaviour of glycosylation-related factors and the interaction of glycan structures in either physiological or pathological conditions may help to gain a deeper knowledge of the intricate regulation of glycosylation.

1.2.2. Protein glycosylation

Glycosylation is the most complex and one of the most abundant post-translational protein modifications occurring in eukaryotes and prokaryotes. In fact, nearly all proteins in serum and in the plasma membrane are glycosylated (Narimatsu, 2006).

Protein glycosylation can be categorized into specific groups based on the nature of the glycan-peptide bond and the glycan attached. The most commonly detected types of glycosylation are N- and O-linked glycosylation whose glycans products are designated N- and O-glycans, respectively. In the case of N-linked glycosylation, glycans are covalently bound to the protein via the nitrogen atom of an asparagine (Asn) residue, while in O-linked glycosylation glycans are attached to the oxygen atom of serines (Ser) or threonines (Thr) residues.

The two types of glycosylation play distinct key roles in cell biology: N-linked glycosylation is important for processes such as protein folding and cell-cell recognition, whereas O-linked glycosylation is essential in the biosynthesis of the proteins that form mucus secretions – mucins. The principles of O-linked glycosylation and the description of the structures, biosynthesis and functions of O-glycans are behind the scope of this thesis (for further reading, see Hayes *et al.* (2012) and Van den Steen *et al.* (1998)). The N-linked glycosylation process is the focus of this thesis and its main aspects will be later described in greater detail.

1.2.3. Structural basis of glycan and glycoprotein diversity

The linkage between two monosaccharide units – a glycosidic bond – is at the basis of diversity existent among glycans. Contrary to peptide bonds, glycosidic bonds are extremely flexible, meaning they can be established in several different ways between two monosaccharides and allow the formation of isomers differing not only in their three-dimensional structures but also in their biological activities. The versatility of glycosidic bonds accounts for the fact that an ensemble of monosaccharides yields a greater number of possible final configurations than the same number of amino acids would yield. In particular, three different amino acids are able to form only six different chains of three residues each, whereas three different monosaccharides can produce more than thousand unique chains of three residues (Varki *et al.*, 2009). This

difference in complexity becomes even more visible as the number of monosaccharide units increases, leading to the theoretical presence of an infinite number of glycan structures in nature. However, glycan structures studied so far are composed of only some of the available monosaccharide units linked in a limited number of combinations, with many more structures expected to be discovered. The number and nature of the monosaccharide units and the conformational arrangements between them also contribute to the variety of existing sugar chains by influencing their length (short or long chains), composition (types of sugar in the chain) and structure (branched or unbranched chains).

When compared to other post-translational modifications of proteins, glycosylation is found to contribute to a higher degree to the diversity of proteome. Two main reasons are pointed out. First, due to the complexity of glycosylation and the non-template driven process of glycan synthesis, the molecular steps occurring during every glycosylation event are likely to vary, leading to slightly different final glycoconjugate products (Kung *et al.*, 2009). Second, in a glycoprotein, diversity arises due to not only the attachment of different glycan structures, but also because of the variable occupancy of glycosylation sites (Marino *et al.*, 2010).

1.3. N-glycosylation and N-glycans in Eukaryotes

The general term of glycosylation is often characterised as a post-translational modification. Although this is a fact for other types of glycosylation, it is not the case of N-linked glycosylation which mainly occurs co-translationally.

N-linked glycosylation (N-glycosylation in short) is the most common type of glycosylation with extreme importance for the normal metabolism of cells as evidenced by the multiple functions played by N-linked glycoproteins in the regulation of vital cellular processes. Moreover, N-glycosylation is essential to life as demonstrated by the fact that a lack of all N-glycans is lethal in species ranging from yeast to mammal (Freeze, 2006).

From this point forward and unless stated otherwise, the terms N-glycosylation and N-glycans will refer to N-linked protein glycosylation and its glycan products in eukaryotes, respectively.

1.3.1. Synthesis of N-glycans

N-glycosylation comprises a complex series of reactions catalyzed by two groups of enzymes having opposite activities: glycosyltransferases which synthesize glycan chains and glycosidases

which hydrolyze glycan linkages. Glycosyltransferases and glycosidases are responsible for the assembly and transformation of N-glycans and their attachment to proteins. The main processing steps occurring during glycan biosynthesis, from the N-glycan initiation in the endoplasmic reticulum (ER) to the complete maturation in the Golgi apparatus are depicted in a simplified manner in Figure 1 and are outlined and briefly described below.

N-glycan synthesis begins on the cytosolic side of the ER with the assembly of the N-glycan precursor by the addition of 14 monosaccharides to a lipid anchor molecule named dolichol phosphate (Figure 1, upper panel). As a result, a lipid-linked oligosaccharide carrying an N-glycan precursor composed of 14 sugars is formed. The lipid-linked oligosaccharide is then flipped across the ER membrane and re-oriented to the reticular lumen. Subsequently, a protein complex called oligosaccharyltransferase catalyzes the co-translational transfer *en bloc* of the N-glycan precursor from the lipid anchor to an asparagine residue of nascent proteins (newly synthesized proteins which are being translocated to the ER). The glycan precursor is directly linked to a specific asparagine residue through an N-glycosidic bond involving the nitrogen atom (N) of the asparagine, hence the term N-glycosylation (Snider, 2013). The asparagine residues candidates to receive N-glycans are usually part of the sequence motif Asn-X-Ser/Thr, where an asparagine (Asn) is followed by any amino acid (X) except proline and ends with a serine (Ser) or threonine (Thr). Following the co-translational attachment of the N-glycan precursor to a nascent protein, an initial trimming of the N-glycans occurs in the ER along with the protein folding.

Additional enzymatic processing and maturation of N-glycans are completed in the Golgi apparatus with the glycoprotein already folded. In the Golgi apparatus, N-glycans are further trimmed and extensively modified by the incorporation of new monosaccharides until a mature, complex N-glycan structure is produced (Figure 1, lower panel). Such modifications include the formation and elongation of branches, also called antennary structures, and the addition of terminal sugars such as *N*-acetylgalactosamine, galactose, sialic acid and fucose to the elongated branches.

In summary, the process of N-glycosylation can be divided into two spatially separated steps: the first step occurring in the ER and concerning the formation and transfer of the N-glycan precursor in association with protein folding; and the second step taking place in the Golgi apparatus and involving the modification and diversification of glycan structures (Helenius & Aebi, 2001). Regarding the glycan-binding site, it should be noted that not all asparagine

residues of a protein can accept an N-glycan. The Asn-X-Ser/Thr motif is considered the glycosylation sequon, i.e., a sequence of three consecutive amino acids in the glycan-acceptor polypeptide chains which is recognized by the oligosaccharyltransferase complex as the attachment site for glycans. Although the Asn-X-Ser/Thr sequon is the most frequently occurring site of glycosylation, N-glycans are also found to be linked in a smaller proportion to other non-standard sequences such as the Asn-X-Cys motif (where Cys is cysteine) (Moremen *et al.*, 2012).

In eukaryotes, the extensive and intricate biosynthetic pathways of N-glycosylation are able to transform a simple N-glycan precursor into a wide and diversified range of complex N-glycan structures. One of the main differences between N-glycosylation in eukaryotes and prokaryotes concerns precisely this source of variability of the N-glycans. While eukaryotes synthesize a conserved lipid-linked oligosaccharide structure and in later steps produce variable antennary structures, prokaryotes synthesize a diverse array of the initial lipid-linked oligosaccharides (Schwarz & Aebi, 2011). Despite the N-glycosylation properties characteristic of each domain of life, the principal events of N-glycosylation described above (the lipid-linked oligosaccharide assembly, flipping across a membrane and transfer to the protein) are shared among the three domains of life and occur in a similar manner (Dell *et al.*, 2010).

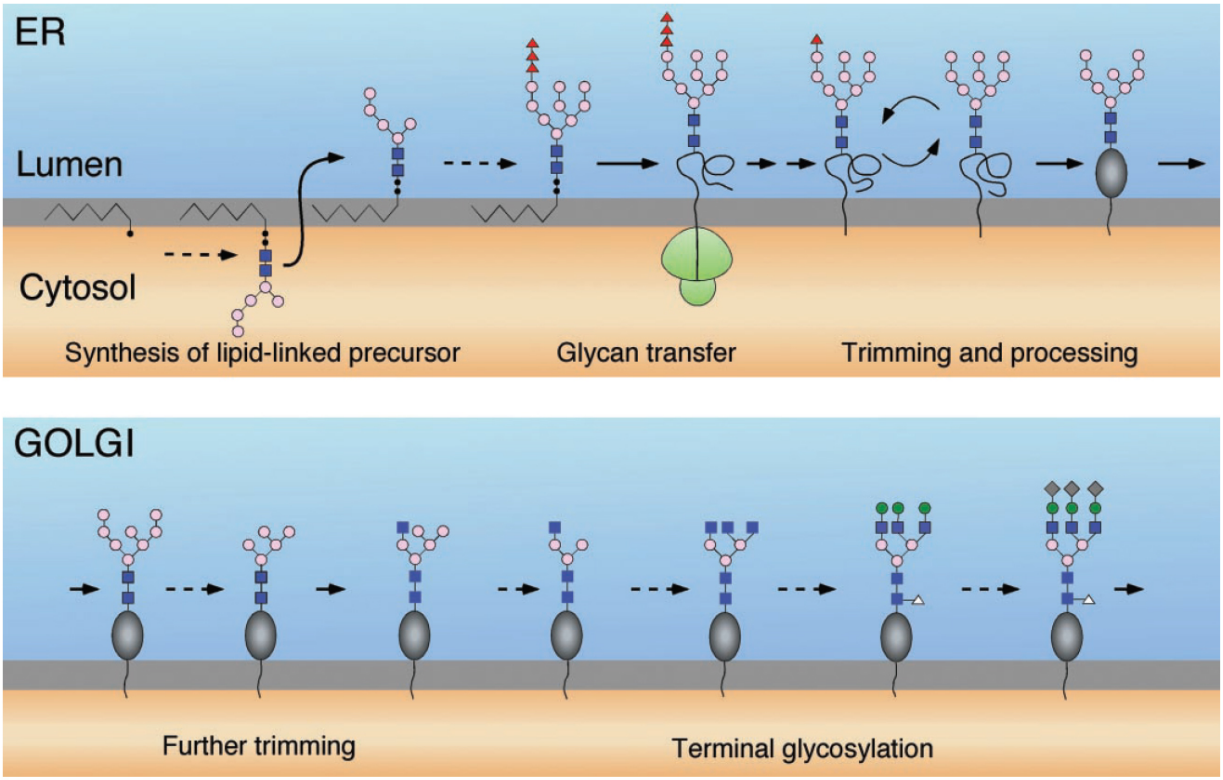


Figure 1. Biosynthesis of N-glycans. The N-glycosylation process can be divided into two spatially separated steps: the first step occurs in the ER (upper panel) and includes the assembly of the N-glycan precursor, the

transfer of the N-glycan precursor to the nascent protein and some minimal trimming; the second step takes place in the Golgi apparatus (lower panel) and involves trimming, elongation and maturation of the N-glycans. While the glycan precursors formed in the ER are conserved, the Golgi reactions generate highly diverse glycan structures that also differ widely between species. In the final mature N-glycan structure, the number and size of branches present is variable, as is the nature of the sugars added; only one of the many possible terminal glycosylation pathways is shown. Adapted from Helenius & Aebi (2001).

1.3.2. Structure of N-glycans

N-glycans are typically an ensemble of 10 to 15 monosaccharides. Unlike DNA and protein molecules which have a linear primary structure, N-glycans are often highly branched molecules of complex structure.

The different ways in which the initial and plain N-glycan precursor is trimmed and modified in the ER and, to a greater extent, in the Golgi apparatus generate three major structural classes of N-glycans: complex, hybrid and oligomannose or high-mannose (Figure 2). These structures vary in the number and size of the antennary structures as well as in the nature of their constituting sugars, while sharing a common core consisting of five monosaccharides kept from the original N-glycan precursor.

The core and branches of the major N-glycan structures are usually subjected to further modifications originating mature N-glycan structures. The main core modification in vertebrates is the addition of fucose to the core residues, called core fucosylation (Varki *et al.*, 2009). Another frequent modification of the N-glycan core is the transfer of an *N*-acetylgalactosamine residue (GlcNAc) to the mannose residue at the base of the N-glycan core, producing a bisecting GlcNAc structure. The elongated branches can be altered by the addition of terminal sugars through capping reactions such as galactosylation, sialylation and fucosylation which add galactose, sialic acid and fucose, respectively.

The human glycome is estimated to comprise more than 7000 N-glycan structures of which only circa 2000 structures have been described (Cummings, 2009). These glycan structures known so far are composed of only some of the available monosaccharide units linked in a limited number of combinations, with many more structures expected to be discovered.

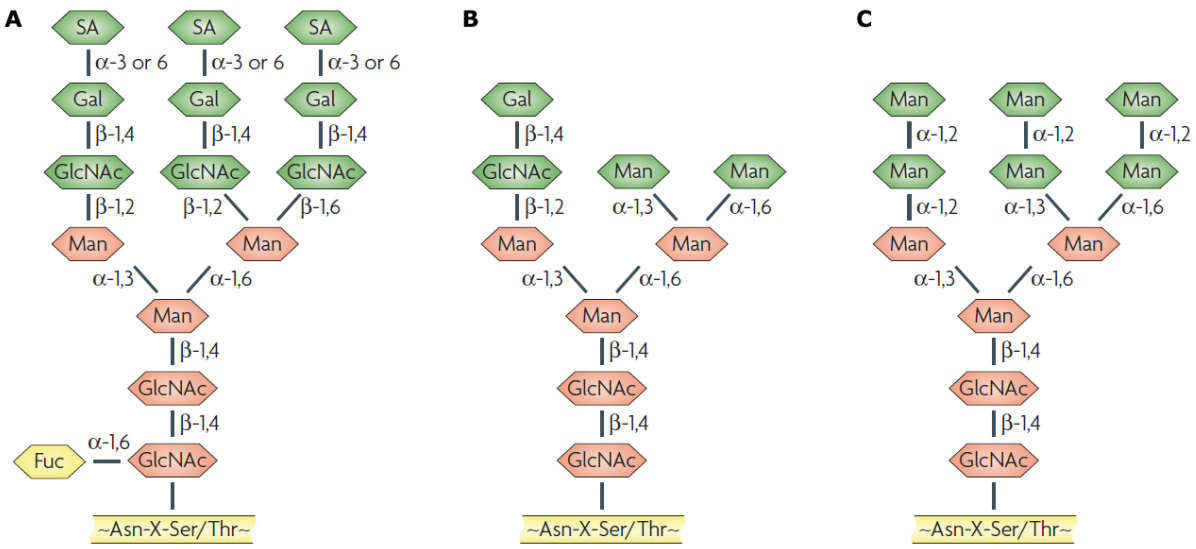


Figure 2. Major structural classes of mature N-glycans in eukaryotes. The structure of mature N-glycans can be divied into three major classes: complex (A), hybrid (B) and oligomannose or high-mannose (C). The monosacharides forming the common core of N-glycans are coloured in red while the monosacharides belonging to the antennary structures are coloured in green. A fucose attached to the core is also shown in the complex N-glycan structure (A). The Asn-X-Ser/Thr binding motif is represented at the bottom of each glycan (where Asn is asparagines, X is any amino acid except proline, Ser is serine and Thr is threonine). Monosaccharide abbreviations: Fuc - fucose; Gal - galactose; GlcNAc - N-acetylglucosamine; Man - mannose; SA - sialic acid. Adapted from Balzarini (2007).

1.3.3. Diversity of N-glycans and N-glycoproteins

The diversity found in N-glycans arises from the association between the nature and adopted configuration of the joined monosaccharide units and the complex network of different reactions involved in the N-glycosylation process.

In addition to the inherited basic features common to sugar chains (discussed in section 1.2.3), N-glycans exhibit their own diversity achieved through multiple elongation and modification reactions occurring during the N-glycosylation. Even though the reactions responsible for the assembly of the branches of the core of N-glycans are arranged in a stepwise manner, they follow a variable pathway which mainly depends on the localization of the glycotransferases and glycosidases through the ER and Golgi apparatus. The enzymatic activity flow associated with the regulatory activity of other factors influences the fate of the core of N-glycans by dictating the main composition and configuration of the antennary structures. Terminal sugars can be further added both to the core of N-glycans (such as fucose or GlcNAc) and to the antennary structures (such as fucose, sialic acid and galactose). Altogether, these modifications introduce

various degrees of structural variability to the common core of N-glycans and account for the enormous spectrum of mature N-glycan structures displayed at the cell surface.

Glycosylation is the most extensive source of protein heterogeneity and N-linked glycosylation in particular is a major contributor to that heterogeneity. N-glycosylation is characterised by a selective and diversified attachment of N-glycans to proteins which generates glycoproteins exhibiting macro- and microheterogeneity (Marino *et al.*, 2010). Macroheterogeneity concerns the glycosylation sites assigned for the attachment of N-glycans (not all available N-glycosylation sites on proteins are occupied) and the number of N-glycans simultaneously linked to the protein (typically between two and five glycans are attached to an average protein). Microheterogeneity refers to the diversity of glycan structures that can be found at a specific glycosylation site of a given protein, i.e. the same class of proteins might have distinct glycan structures attached to identical glycosylation sites. Macro- and microheterogeneity appear to be associated with the activity of glycosyltransferases and glycosidases. Since these enzymes have a remarkable degree of substrate specificity, their activity can be easily constrained due to protein sequence and conformation as well as environmental factors. Changes in enzymatic activity influence the fate of newly synthesized glycoproteins and lead to different glycosylation patterns characteristic of different cell types and stages of cell cycle, such as development, differentiation and maintenance.

The intrinsic variability of N-glycans and the heterogeneity of glycoproteins give rise to a vast glycome composed of thousands of glycan isomers and glycoprotein isoforms and increase the structural diversity of an already broad proteome.

1.3.4. Biological roles of N-glycans in health and disease

As opposed to the core synthesis of N-glycans which is mainly conserved, the assembly of antennary structures is often regulated in a tissue- or cell lineage-specific manner suggesting that the branches can be directly implicated in the different functions of N-glycans (Varki *et al.*, 2009).

N-glycans are complex extensions of the glycoproteins and their impact on the protein itself is not restricted to the structural level but extends to its biochemical and functional properties. Small structural modifications of N-glycans can be sufficient to cause loss or impairment of protein function showing that N-glycans are neither passive nor functional independent components of glycoproteins; on the contrary, N-glycans and proteins forming glycoproteins

work as one functional unit. Thus, the functional universe of proteins should not be explored without their glycan moieties since they are an integral part of the identity of glycoproteins.

Determining the functions of glycans and unravelling their contribution to the activity and properties of glycoproteins has been a challenging task. The strategies applied vary from the inactivation of enzyme-coding genes through genetic mutation methods to the use of inhibitors for specific N-glycosylation reactions and the study of features presented by mutant cells or organisms with a defect in N-glycosylation (Varki *et al.*, 2009). However, to fully understand the function of glycans it is necessary to have a detailed characterization of their structures which is technically difficult to achieve due to their tremendous structural diversity.

Regardless of the obstacles posed, advances in the area of functional glycobiology have been attributing fundamental roles to N-glycans in a multitude of key biological processes including protein folding, stability and targeting, molecular trafficking and clearance, cell adhesion, signal transduction and cell-cell interactions.

Like other co- and post-translational modifications occurring in the ER, N-glycosylation is important for correct protein folding. In the absence or failure of N-glycosylation, glycoproteins usually misfold and aggregate and, consequently, are subjected to degradation by quality control mechanisms in the ER. Therefore, the most basic known function of N-glycans is to facilitate protein folding in the ER which explains the fact that glycans are attached co-translationally to nascent polypeptide chains still in their unfolded state (i.e. not in their native structure). N-glycans ensure the proper folding of proteins by directly stabilizing their structure or by acting as recognition tags and promoting interactions between glycoproteins and enzymes involved in protein folding (Helenius & Aebi, 2004).

Concurrent with the aid in protein folding is the involvement of N-glycans in the protein quality control system in the ER. The main purpose of this quality control system is to monitor the integrity of protein synthesis and prevent aberrant proteins (misfolded or non-functional) from going further in the secretory pathway by assigning them for degradation. The mechanisms of quality control employ glycan moieties as tags to mediate the correct recognition of misfolded proteins which are then selectively retained and targeted for posterior degradation (Yoshida, 2003). It has been proposed that glycans attached to misfolded proteins might not be properly trimmed, thus functioning as indicators of the protein structure condition, i.e., whether the protein failed to fold or folded correctly (Gamblin *et al.*, 2009).

Furthermore, glycans might play a role in the secretion and intracellular transport of proteins and they appear to protect proteins from proteolysis as suggested by the fact that proteins lacking N-glycans are more susceptible to proteolytic degradation (Fiedler & Simons, 1995). Even though some glycoproteins have shown to be functional when lacking N-glycan moieties, they still require the presence of N-glycans for folding and transport out of the ER as N-glycans will affect protein conformation and stability (Trombetta, 2003).

The reported importance of N-glycans in cell differentiation, adhesion and migration as well as in cell-cell communication and signal transduction is somehow expected since most receptors and adhesion molecules on the cell surface are N-glycosylated (Gu & Taniguchi, 2008). For instance, it has been shown that the branching structures of N-glycans in growth factor receptors serve as important determinants for the signalling function of the receptors (Takahashi *et al.*, 2004). In addition, N-glycans have been found to have crucial roles in the nervous system development, regeneration and synaptic plasticity by mediating the formation of neural cell interactions (Kleene & Schachner, 2004).

Some of the hormones regulating major metabolic and reproductive functions of the body are also N-glycosylated. Evidence suggests that N-glycan moieties of these glycoprotein hormones have a biological role in hormonal control by being involved in their differential targeting and blood clearance (Thotakura & Blithe, 1995).

Due to the participation of N-glycans in an extensive list of vital processes, defects in N-glycan biosynthesis can compromise the course of these processes and, consequently, lead to disease. Not surprisingly, N-glycans have been associated with many pathological events, including host-pathogen interactions, tumour invasion and metastasis, diabetes, cardiovascular, immunological and genetic disorders, among others. Common to all these pathological conditions is the observation of an altered pattern of glycosylation.

N-glycans present in cell surface glycoproteins that mediate cell–cell and cell–matrix interactions have been described to be implicated and greatly contribute to the metastatic process (Zhao *et al.*, 2008). In epithelial tumours, including those of breast, colon and prostate, the adhesion and signalling properties of cells are affected due to modifications in N-glycan structures displayed at the cell surface (Rambaruth & Dwek, 2011). These structural alterations in N-glycans lessen the interactions between glycans and their binding partners and promote cell migration and invasion, thus providing favourable conditions for tumour progression and dissemination.

Among the large number of genetic disorders related to glycosylation which have been identified in recent years, the group of congenital disorders of glycosylation is one of the most explored. Congenital disorders of glycosylation are a group of rare but severe inherited metabolic disorders characterised by defects mainly in the N-glycosylation pathway (Marquardt & Denecke, 2003). In congenital disorders of glycosylation, genetically inherited mutations in glycosylation-related genes are the cause of deficiency in 34 different enzymes participating in the N-glycan synthetic pathway (Sparks & Krasnewich, 2005). Congenital disorders of glycosylation usually affect multiple organ systems (especially nervous, gastrointestinal, hepatic, visual and immune systems) and present a broad spectrum of clinical features ranging from psychomotor difficulties to mental retardation. The various different symptoms manifested by patients who suffer from congenital disorders of glycosylation pose an obstacle to a correct and early diagnosis of these diseases (Marquardt & Freeze, 2001).

As can be seen, N-glycans regulate many physiological and pathological processes and a correct N-glycosylation is a prerequisite for the normal function of the cells and, consequently, of the entire organism. One the one hand, understanding in more detail how N-glycans influence the behaviour of glycoproteins can help to clarify the precise function of N-glycans as well as to provide new insights on the biology of glycosylation-related disorders. On the other hand, understanding the mechanisms leading to disease and identifying specific alterations in glycosylation associated with it can aid the discovery of new biomarkers and therapeutic targets and promote the development of novel and more efficient diagnostic and treatment solutions.

1.4. Human Plasma N-glycome

1.4.1. Challenges of structural analyses of N-glycans

Understanding the biological roles of glycans and their involvement in diseases or establishing the cause of different glycosylation patterns are relevant topics in the field of glycobiology. Although developing at a slow pace over the years, glycobiology has given crucial insights into the importance of glycans and glycosylation which have contributed to the recent growing interest in glycan analysis.

Glycan analysis provides a structural description of glycans that can be valuable for a more comprehensive view of the functional significance of glycans (and glycosylation). However, the challenges faced by glycan analysis account for the fact that the knowledge of glycan structures

and their synthesis lags behind the knowledge acquired in protein or nucleic acids research which are not affected by such problems.

The heterogeneity and structural complexity exhibited by glycans and the inexistence of a universal glycan structure code capable of explaining such diversity have been a bottleneck in the determination of glycan structures and a restraint to glycan analysis. Such extremely challenging nature of glycan structures demands adequate, robust and efficient methods that can provide a correct and detailed analysis of glycans. Chromatographic and mass spectrometry-based methodologies are the principal strategies used for structural analysis of glycans (Stumpo & Reinhold, 2010). While mass spectrometry techniques have higher resolution and are able to identify a greater number of glycan structures, chromatography methods are better at separating isomers despite their limited resolution. The methodology-related variability yields slightly different results creating comparability issues and making the validation of results obtained from different sources a more difficult task (Thobhani *et al.*, 2009). Since no reference standards are available, results from glycan analysis should be interpreted in the light of the methodology of choice.

Additionally, structural analyses of glycans have been restricted to a reduced number of samples due to technological limitations and, thus, a complete and detailed characterization of the glycome composition has remained scarce. High-performance liquid chromatography, the simplest technique used for a broad profiling analysis of glycans, has been recently adapted for high-throughput glycan quantification (Royle *et al.*, 2008). The possibility to quantify glycans in a relatively large number of samples opens new venues for the study of glycans and the investigation of factors associated with glycosylation in a large scale.

1.4.2. Variability, heritability and stability of N-glycans

Recent developments in methodological procedures, namely the adaptation of high-performance liquid chromatography for high-throughput analysis of glycans, has allowed the first large scale study evaluating the variability and heritability of the human plasma N-glycome (Knezevic *et al.*, 2009). The plasma N-glycan profiles of 1008 individuals were analysed based on a chromatographic division of 33 peaks containing similar glycan structures. The observed variability of glycans at the population level was larger than expected emphasizing the need for a careful approach when using glycan levels for diagnostic purposes. A broad range of variation in heritability of glycans was also found suggesting that the influence of genetic and environmental factors varies according to different structural glycan groups.

Since the variability found at the population level was larger than changes reported to be associated with disease, a follow up study was conducted to test the stability of the human plasma N-glycome over a period of time and evaluate the validity of the use of glycan changes for diagnostic purposes (Gornik *et al.*, 2009). Several plasma N-glycan profiles were obtained for 12 healthy individuals during 5 days and the profiles within an individual were compared. The plasma N-glycome showed a good temporal stability suggesting a significant genetic background control. Thus, glycan changes arising from environmental factors and/or altered physiological processes present themselves as potential diagnostic markers for diseases.

A comprehensive analysis of association between N-glycans of human plasma and several environmental factors and biochemical traits reported smoking, diet, lipid status, gender and age to affect different glycosylation features (Knezevic *et al.*, 2010). However, the parameters analysed explained only a small fraction of the variability observed in glycan levels supporting the previous evidence that glycans are under great genetic control.

The relation between glycosylation and ageing is attracting a lot of attention and few studies have investigated how plasma N-glycans change during ageing (Ding *et al.*, 2011; Knezevic *et al.*, 2010; Vanhooren *et al.*, 2010; Vanhooren *et al.*, 2008). Similar age-related structural changes in N-glycan profiles were consistently reported in all studies regardless of the ethnic origin of the populations considered (Belgian, Chinese, Croatian and Italian). Some of these studies also analysed the relation between glycan levels and gender. However, in this case, glycan differences between males and females at different age stages were not reproducible in all studies. This age and gender dependence of glycans should be taken into account in the development of glycan-based diagnostic tools as well as in the data analysis of studies comparing different groups.

Glycosylation changes associated with the intake of different medications were analysed and few sporadic associations were identified but not demonstrated in all tested groups (Saldova *et al.*, 2012). Additional studies analysing a larger number of samples are necessary to validate the associations observed.

The recent availability of considerable amounts of glycomic and lipidomic data and the biological importance of these two major classes of molecules motivated the first glycome and lipidome-wide association study intended to reveal possible interactions between 46 plasma N-glycan structural features and 183 lipid traits in individuals from three geographically distinct population cohorts (Igl *et al.*, 2011). Although strong associations between N-glycans and lipids

were found in each individual population, these patterns of association were different and not completely replicable across populations. The observations suggest potential interactive metabolic pathways between glycans and lipids and show the presence of population-specific correlations which are thought to derive from exposure to different environments and genetic background.

All these studies on plasma N-glycans are an attempt to characterise in detail the human N-glycome and provide a complete description of its behaviour and peculiarities at a large scale and at the population level. The overall findings of the overviewed analysis suggest a strong influence of the genetic component on the glycan levels and possible associations of glycans with several (patho)physiological phenotypes. The integration of glycan traits, phenotypes and genotype data is required in further studies in order to determine the extent of validity of these preliminary findings.

1.4.3. Potential diagnostic value of the N-glycome

Blood is the central transport medium across the human body and is composed of blood cells suspended in blood plasma. Blood plasma is the liquid component of blood containing dissolved proteins, among other substances such as glucose and clotting factors.

The majority of proteins in blood plasma are glycosylated. As said before, glycosylation is the most common post-translational modification of proteins and a biological important process for the human physiological metabolism. Dysregulation of glycosylation has been implicated in several diseases, making it plausible to assume that glycosylation alterations can be a sensitive indicator of changes in the external and internal environment of the cell. Modifications in the mechanisms controlling and changing glycosylation will ultimately lead to glycan structural variation which can be examined and determined through glycan analysis. Therefore, the composition of plasma N-glycome is expected to reflect diverse physiological status of the organism and to be able to act as a biomarker.

The potential value of glycan profiles as a diagnostic biomarker for a type of maturity-onset diabetes of the young (MODY) has been assessed (Thanabalasingham *et al.*, 2013). Glycan profiles were shown to be altered in individuals presenting the condition and the particular changes identified were suggested to be used together with existing biomarkers to improve the diagnosis of the disease. Probable genotype-phenotype relationships were also indicated but the validation of such evidences requires more extensive studies. Similar studies involving other

diseases can help to clarify the role of glycans in pathological conditions and to explore their use as biomarkers.

In the context of biomarkers application, it should be noted that to date the majority of studies involving the analysis of N-glycans report findings exclusively in adult populations. However, the glycosylation profiles in childhood are of equal importance and should not be underestimated, especially when children are largely affected by congenital disorders of glycosylation. The composition of plasma and IgG N-glycome was analysed in childhood and was reported to vary from the glycosylation profiles observed in adulthood (Pucic *et al.*, 2012). Thus, the use of glycans as diagnostic biomarkers and the development of glycan-based therapeutics should take into account the different glycosylation patterns found in children and adults.

Analysing the plasma N-glycome and monitoring glycosylation changes can bring insights into the mechanisms of glycosylation in health and disease as well as open new possibilities to the clinical application of glycans in medical prognosis, diagnosis and therapy procedures. The great and promising potential of the human N-glycome as a disease biomarker is further strengthened by the ready availability of plasma and by the simplicity and non-invasiveness of blood sampling procedures.

1.4.4. Genome and Glycome-wide association studies

Genome-wide association studies (GWAS) aim to find associations between genotype and phenotype. The genotype is represented by single-nucleotide polymorphisms (SNPs) while the phenotype is represented by a disease trait or a biochemical/physiological feature. Although it is the current strategy of choice for screening relevant genetic variants underlying human diseases and traits, GWAS shows major drawbacks.

The first limitation is related with the application of GWAS to the case of polygenic diseases or traits. GWAS can accurately detect mutations responsible for single gene disorders (also known as Mendelian disorders) due to the fact that a certain disease is caused by SNPs in a single gene. On the contrary, complex diseases and traits involve complicated interconnections between multiple genes as well as environmental factors and gene-environment interactions. As such, the cause-effect relation between genotype and phenotype is not as direct as in the single-gene disorders and establishing associations between variants and complex diseases and traits becomes a less simple and straightforward task (Hirschhorn & Daly, 2005). Nonetheless, GWAS

has been widely applied to the study of polygenic diseases and traits, such as diabetes, asthma, cancer, cardiovascular and neurological disorders, obesity and elevated blood cholesterol levels. Although several common genetic variants influencing complex human diseases and traits have been indentified, GWAS is unable to recover all loci involved (Cooper & Shendure, 2011). Moreover, it should be emphasized that the majority of the variants discovered by GWAS only explain a small proportion of the genetic contribution to the phenotype variance and thus, cannot be taken with full reliability as risk factors for disease (Queitsch *et al.*, 2012).

The second limitation concerns the rationale behind the GWAS approach. GWAS are mainly based on a series of single-locus analysis where each SNP is examined independently for association with the phenotype through a statistical test that depends on the nature of the phenotype (quantitative or categorical). Such gene selection approaches using univariate (gene-by-gene) analysis are easy to implement and to interpret and are computationally inexpensive (Moore *et al.*, 2010). However, these traditional univariate models are often unable to deal with nonlinear relationships and high-dimensional data which are characteristic to large studies. Additionally, univariate methods assume the existence of a simple genetic architecture excluding possible gene-gene interactions, which are known to occur in complex diseases and traits. While being part of the genetic architecture, gene-gene interactions are likely to play an important role in the genotype to phenotype mapping relationship and should be considered in GWAS.

Modified versions of GWAS and complementary approaches have been proposed through the years to overcome the drawbacks and improve the potential of GWAS. In particular, multi-locus analyses that explore the interactions between SNPs have been investigated. However, genome-wide studies currently generate between 500,000 and 1,000,000 markers and combinatorially examining all possible pairwise or higher-order SNP interactions is a computationally infeasible approach (Bush & Moore, 2012).

Strategies aimed to reduce the number of tested SNPs propose that the analysis should be performed in two-stages: in the first stage, a subset of likely associated genetic variants is selected based on a chosen method (usually single-locus methods); and in the second stage, a desired multi-locus analysis is performed on this filtered and reduced subset (Cordell, 2009). Although often employed in genetic analyses, filtering strategies might miss potential interacting markers with small marginal effects as these will be missed and eliminated in the first stage of SNP selection.

Machine learning approaches such as tree-based methods or support vector machines have been used as a promising alternative to filtering algorithms in several association studies (Jiang *et al.*, 2009; Li *et al.*, 2011; Mittag *et al.*, 2012; Yao *et al.*, 2009). Popular tree-based methods are the Random Forests and the Random Jungle which is an improved version of Random Forests implemented to allow the analysis of high-dimensional data (Schwarz *et al.*, 2010; Winham *et al.*, 2012). These ensemble learning methods do not include the interaction between SNPs per se but allow for their interaction during the process of tree contruction. In other words, the paths in the tree-like structures correspond to particular combinations of SNPs which might mirror potential interactions between them.

Analysing all SNPs available for a genome-wide study as well as existent SNP-SNP interactions would be the ideal scenario for GWAS. Despite the great effort put into trying to solve this problem, there are still computational, statistical and logistical challenges which need to be overcome. Due to the computational and memory requirements inherent to the analysis of high-dimensional data, algorithms with high statistical efficiency and computational performance are necessary to provide faster analyses and to improve the findings discovered by current genome-wide approaches. Recently, multivariate methods have been developed to address the problem of SNP selection and their use in GWAS has shown satisfactory results (Rotival *et al.*, 2011; Zhou *et al.*, 2013; Zuber *et al.*, 2012). Such multivariate methods implement polygenic modelling algorithms which are able to simultaneously analyse multiple SNPs and account for their dependencies while executing the task in an acceptable amount of time. Polygenic modelling is viewed as a promissing and valuable approach in genome-wide studies and has been gaining more attention over the commonly used standard univariate methods.

Glycosylation is a polygenic trait characterised by the production of different glycan structures. The levels of these glycan products can be measured and considered as individual phenotypes in analyses having the same rationale as GWAS. Such glycome wide studies aim to get more insights into the genetic regulation of the glycosylation process as well as into the association of glycans with diseases.

This was the fundament for the first comprehensive analysis of common genetic polymorphisms affecting protein glycosylation which was recently performed by combining high-throughput glycan analysis with the GWAS approach (Lauc *et al.*, 2010a). This pilot study conducted a meta-analysis of GWAS data for 13 N-glycan features in individuals from three European populations (Vis, Korčula and Orkney). N-glycan levels in human plasma were found to be

influenced by a set of polymorphisms located at three loci comprising the fucosyltransferase 8 (FUT8), the fucosyltransferase 6 (FUT6) and the hepatic nuclear factor 1 alpha (HNF1A) genes. A complementary study was performed to include an additional population sample (Sweden) and to extend the analyses to 46 glycosylation traits (Huffman *et al.*, 2011). The results of the pilot study were reinforced and three novel associations with glycan features were identified for b-1,3-glucuronyltransferase 1 (B3GAT1), solute carrier family 9, member 9 (SLC9A9) and mannosyl(a-1,6-)-glycoprotein b-1,6-N-acetyl-glucosaminyltransferase V (MGAT5) genes. For the first time, high-throughput data from genomics and glycomics is brought together in an effort to map the complex network of genes involved in the regulation of protein N-glycosylation and to unravel the mechanisms behind the genetic associations observed.

Despite the optimistic results achieved in these preliminary studies, the polygenic nature of glycosylation reflects itself in the still scarce understanding of the genetic regulation of glycosylation. In this context, the polygenic modelling methods could show their usefulness in fetching new polymorphisms related to glycosylation or to be used as a faster alternative to GWAS analysis.

1.5. Immunoglobulin G: an N-linked glycoprotein

1.5.1. Structure and function of IgG

Antibodies, or immunoglobulins, are glycoproteins produced by the immune system in response to bacteria, virus, toxins or other pathogens. Antibodies are released throughout the body to mediate a variety of effector functions, aimed at identification, neutralization and removal of infectious agents and their products. Usually, the antibody is required to bind its antigen (a specific part of pathogens which is unique to each of them) in order to trigger the effector functions.

In mammals, antibodies can be grouped according to their structure into five major classes, also called antibody isotypes: IgA, IgD, IgE, IgG and IgM; where the prefix Ig stands for immunoglobulin. These antibody isotypes differ in their biological properties, functional locations and each of them helps to coordinate an appropriate immune response for a given pathogen.

Immunoglobulin G, or IgG, is the most abundant antibody isotype found in human blood accounting for approximately 75% of the total immunoglobulins in plasma of healthy

individuals. IgG is a major effector molecule of the humoral immune response by activating the complement system and inducing phagocytosis in order to protect against bacterial and viral infections.

IgG is a glycoprotein composed of two identical light chains and two heavy chains connected by disulfide bonds and forming a tetramer with a Y-shaped structure (Figure 3). The structure of IgG can be divided into two regions: the antigen-binding fragment (Fab) comprising the arms of the Y structure and the crystallizable fragment (Fc) forming the tail region of the Y structure. These two regions account for the main biological activities of IgG: the Fab portion is responsible for the recognition of pathogens by bearing the site to bind antigens; and the Fc domain initiates the effector functions by interacting with cell surface receptors.

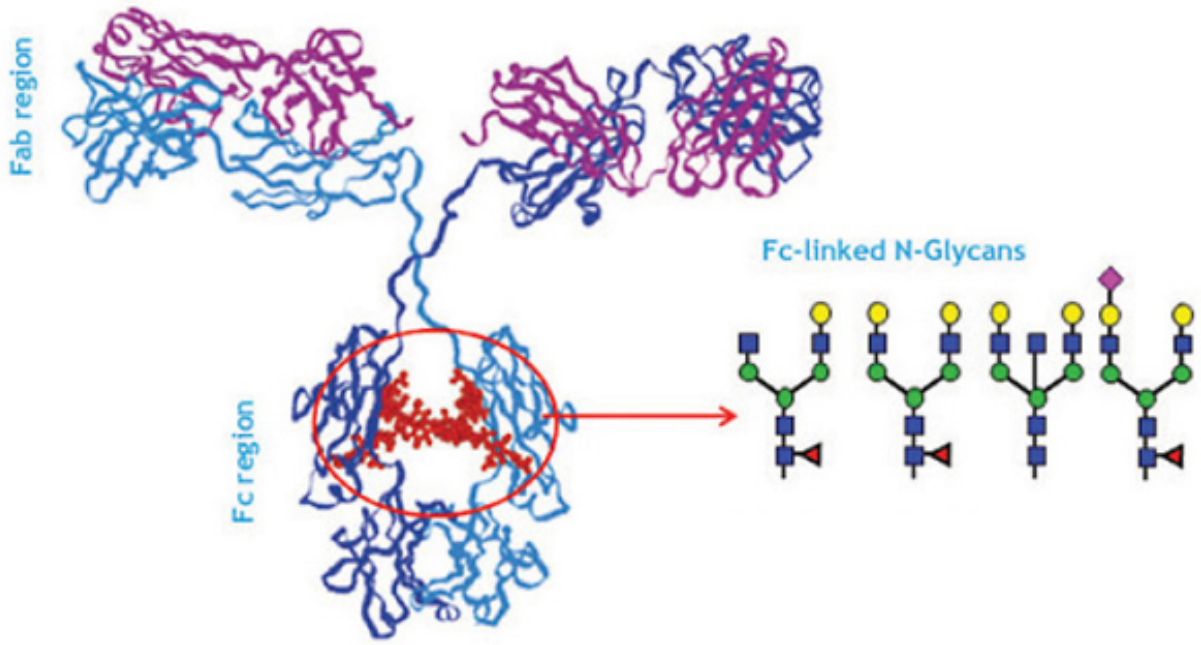


Figure 3. Immunoglobulin G structure. The Y-shaped structure of the human IgG is composed of the Fab region (responsible for the antigen-binding activity) and the Fc region (responsible for the effector functions). A single N-glycan is attached to each asparagine-297 residue in the Fc portion and is shown in red in the tri-dimensional molecule. Four possible structures of Fc-linked N-glycans are shown; the monosaccharides represented are N-acetylglucosamine (blue squares), mannose (green circles), fucose (red triangles), galactose (yellow circles) and sialic acid (pink diamonds). Adapted from New England BioLabs (2013).

1.5.2. Glycosylation of IgG

Glycosylation of Fc fragments is fundamental for this domain to be able to mediate the effector functions of the IgG. In the glycosylated state, each heavy chain of IgG has a single biantennary N-glycan attached to the highly conserved N-glycosylation site asparagine 297 of the Fc region (Figure 3). The two N-glycan moieties, also known as Fc glycans, are crucial for the interaction between IgG and the receptors, as demonstrated by the fact that interaction between the two parts is lost in the absence of glycosylation (Kaneko *et al.*, 2006). Fc glycans are believed to maintain an open conformation of the heavy chains which is favorable to the binding of IgG to receptors (Anthony & Ravetch, 2010).

Glycosylation of IgG varies considerably due to modifications of the biantennary core or elongation of the arms of the Fc glycans through sugar additions. These structural alterations are frequent and over 30 different glycans have been detected on IgG in healthy individuals (Anthony *et al.*, 2012). However, certain glycan structures might alter the conformation of the Fc region in a way that affects its affinity to receptors and, as a result, have a profound impact on the effector functions of the IgG.

One of the most striking examples is the presence of terminal sialic acids which totally reverts the innate function of IgG. Sialylation alters the binding ability of IgG and converts IgG from having pro-inflammatory into having anti-inflammatory activity (Kaneko *et al.*, 2006). The potential anti-inflammatory behaviour manifested by IgG has been used successfully in the intravenous IgG therapy, a common treatment for a number of autoimmune diaseases (Lux *et al.*, 2010).

A fucose residue attached to the glycan core is present in the majority of IgGs but rarely found in other plasma proteins. This core-fucosylation seems to negatively influence the action of IgG on the antibody-dependent cell-mediated cytotoxicity mechanism, as opposed to a lack of core-fucose which enhances the affinity of IgG to bind receptors of cells involved in the antibody-dependent cell-mediated cytotoxicity process (Gornik *et al.*, 2012).

The first report implicating IgG glycosylation in disease dates back to 1985 and describes decreased IgG galactosylation to be associated with rheumatoid arthritis (Wuhrer *et al.*, 2007). Since then, the interest in the potential role of IgG glycosylation in disease increased and several studies followed reporting the presence of characteristic IgG glycosylation patterns in other autoimmune diseases, infectious diseases and cancer.

A correct glycosylation is of physiological importance for glycoproteins and IgG is a clear example of that. Alternative glycosylation of IgG induces structural alterations in its Fc domain which enable IgG to perform completely different functions. In this way, IgG glycoforms play an important role in the modulation of inflammatory responses (Hounsell & Davies, 1993).

Although the functions of alternative glycosylation of IgG have been analysed in health and disease, the molecular significance of these changes and the specific regulation of the glycosylation process are still mostly unknown.

1.5.3. Structural analyses of IgG

IgG is one of the most studied glycoproteins in terms of structural and functional aspects of glycosylation. The interest in IgG lies not only on its important biological activity in humoral immune responses but also on its glycosylation patterns which have been shown to be altered under various physiological and pathological conditions. Understanding the alternative glycosylation of IgG requires a detailed analysis of the composition of the IgG N-glycome.

Recently, a high throughput method for the isolation of IgG was developed and applied to the first large-scale study of the IgG N-glycome which showed a higher variability between individuals than that reported for the total plasma N-glycome (Pucic *et al.*, 2011). Associations between certain IgG glycosylation features and age, such as an increase of structures with bisecting GlcNAc and a decrease in galactosylation and sialylation, were observed in accordance with previous studies which also reported the dependence of glycosylation features sex and pregnancy (Huhn *et al.*, 2009).

While the present thesis was in progress and the analyses in completion, a genome-wide association study of the human IgG N-glycome was published (Lauc *et al.*, 2013). Significant associations with IgG glycans were found for nine genetic loci. While four loci comprise genes encoding known glycosyltransferases, the remaining five loci have not been previously implicated in protein glycosylation but comprise genes reported to be related with autoimmune and inflammatory conditions.

In the studies concerning the analysis of human plasma N-glycome, the broad spectrum of N-glycan structures considered are carried by many diverse glycoproteins which might be under distinct glycosylation regulation. However, these glycan analyses do not include any information about the glycoproteins themselves in the sense that the glycan moieties examined are not differentiated according to the glycoproteins from which they were released. Thus, the analysis

of the entire set of N-glycan moities existent in the plasma provides an overall descriptive and quantitative view of the N-glycome but overlooks subtle glycan structural fluctuations occuring at the individual level of glycoproteins. Separating the glycans by glycoproteins and analysing them independently has the potential to detect glycoprotein-specific glycan changes dissimulated at the plasma level, thus adding another dimension to the knowledge about glycosylation regulation.

1.6. Aim and objectives of the thesis

Major breakthroughs in methodological procedures created the possibility to reliably quantify glycans in a high-throughput manner and allowed the first large scale studies reporting a comprehensive description of the behaviour of human N-glycans and of possible causes behind that behaviour. These studies and their encouraging and promising results constitute the basis and the motivation for the present research.

The aim of this thesis is to gain more insights into the genomic and environmental regulation of glycosylation by using advanced bioinformatics tools. At the present stage of glycome research, the field of bioinformatics is required to develop, adapt and improve computational algorithms for a more thorough exploration and accurate characterization of the available spectrum of glyco-related data (Aoki-Kinoshita, 2008). Three isolated population cohorts characterised on the level of the glycome, genome and physiological/biochemical parameters were analysed as case studies.

Within the aim of the thesis and regarding the available data, the following objectives were defined:

- to develop a general data processing pipeline to treat and prepare the data for further analysis;
- to investigate the existence of glyco-phenotypes, in particular glycan changes associated with medical conditions such as diabetes;
- to examine the presence of population-based glycosylation patterns that could characterise geographically distinct cohorts;
- to explore associations between glycans and genotypes that could reveal new variants influencing the glycosylation process.

Computational methods/algorithms present in literature were researched and evaluated for their suitability to fulfil the particular needs of each of the mentioned objectives. Several methods considered to be appropriate for the analyses were chosen. Since methods intended for the same type of analysis usually differ in their principles and might yield different results, for some of the analyses more than one method was investigated. In such cases, the performances of the methods applied were compared and the agreement of their results was assessed. Exploratory graphical methods were employed to enable the visualisation of the results obtained with the computational methods and facilitate their interpretation

2. MATERIALS & METHODS

2.1. Study Populations

Data consists of human samples from three different isolated population cohorts: the islands of Vis and Korčula in Croatia and the Orkney archipelago in Scotland. Individuals were recruited as part of larger genetic epidemiology studies intended to investigate genetic variability and map genes associated with common complex diseases and disease traits in genetically isolated populations.

The "10 001 Dalmatians" study of Croatian island isolates includes 1008 individuals from Vis island (Vis cohort) and 969 individuals from Korčula island (Korčula cohort) (Rudan *et al.*, 2006; Rudan *et al.*, 1999; Rudan *et al.*, 2009).

The Orkney Complex Disease Study (Orkney cohort) includes 2095 individuals from Orkney Islands (Igl *et al.*, 2010).

All individuals are adults over 18 years of age and the mentioned studies were approved by the appropriate ethical committees. In all three population studies, blood samples were drawn, biochemical and physiological traits were measured, lifestyle and medical-related information was acquired and DNA samples from individuals were genotyped following similar protocols.

2.2. N-glycan Quantification Analysis

2.2.1. Plasma N-glycans

A high-throughput method was used to isolate and quantify the glycan structures present in the plasma samples of the individuals. The developed methodology allows a rapid and detailed analysis of a large number of samples by combining a 96-well plate platform with quantitative high-performance liquid chromatography (HPLC) profiling in an automated manner (Royle *et al.*, 2008).

Prior to the chromatographic analysis, plasma samples are required to be preprocessed for the release and labeling of N-glycans. First, N-glycans are enzymatically released from glycoproteins using peptide N-glycosidase F (PNGase F) which cleaves the linkage between the core of the glycans and the asparagine residue of the protein. Second, since isolated N-glycans

do not present any chromophores, they are labeled with 2-aminobenzamide (2-AB) for fluorescent detection (Adamczyk *et al.*, 2012). The labeling of glycans is nonselective allowing charged and neutral glycans to be analysed simultaneously.

The released and labeled N-glycans are then analysed by hydrophilic interaction high performance liquid chromatography (HILIC) with fluorescent detection to identify and quantify individual glycans present in the samples.

Glycan profiling aims to identify and assign glycans structures to the various peaks in the chromatogram obtained on the basis of the elution positions of different glycans. For this specific purpose, the measured elution positions of glycans are converted to glucose units which are then matched against reference values in the GlycoBase database for structure assignment.

A 2AB-labeled glucose ladder is used as an external calibration standard for the assignment of glucose units. The glucose ladder chromatogram contains the elution positions (or retention times) of glucose homopolymer species with different degrees of polymerization allowing each chromatographic peak to be expressed as a glucose unit. Thus, a chromatogram of a certain glycan pool can be compared to the reference glucose ladder and the elution positions of individual glycans can be assigned with glucose units.

The prediction of glycan structures based on glucose unit values is possible due to the fact that each monosaccharide present in the structure has its own additional value. In this way, the glucose unit value of each glycan structure is directly related to the number and type of linkage of its constituent monosaccharides, i.e., higher glucose unit values correspond to larger glycans. GlycoBase contains the HPLC elution positions expressed as glucose unit values for more than 700 2AB-labelled glycan structures both N-linked and O-linked (Campbell *et al.*, 2008).

The identification of different glycan structures is followed by the division of the chromatogram into certain chromatographic areas and the later quantification of glycans on those areas. The chromatograms are divided into several peaks based on peak resolutions and similarity of individual glycan structures; each peak containing more than one glycan structure (Figure 4). The amount of glycans present in each peak is expressed as a percentage of the total integrated area of the chromatogram and calculated as the amount of total glycan structures in the peak divided by the total serum N-glycome; the percentages of all peaks add up to 100% for a single chromatogram (Knezevic *et al.*, 2009).

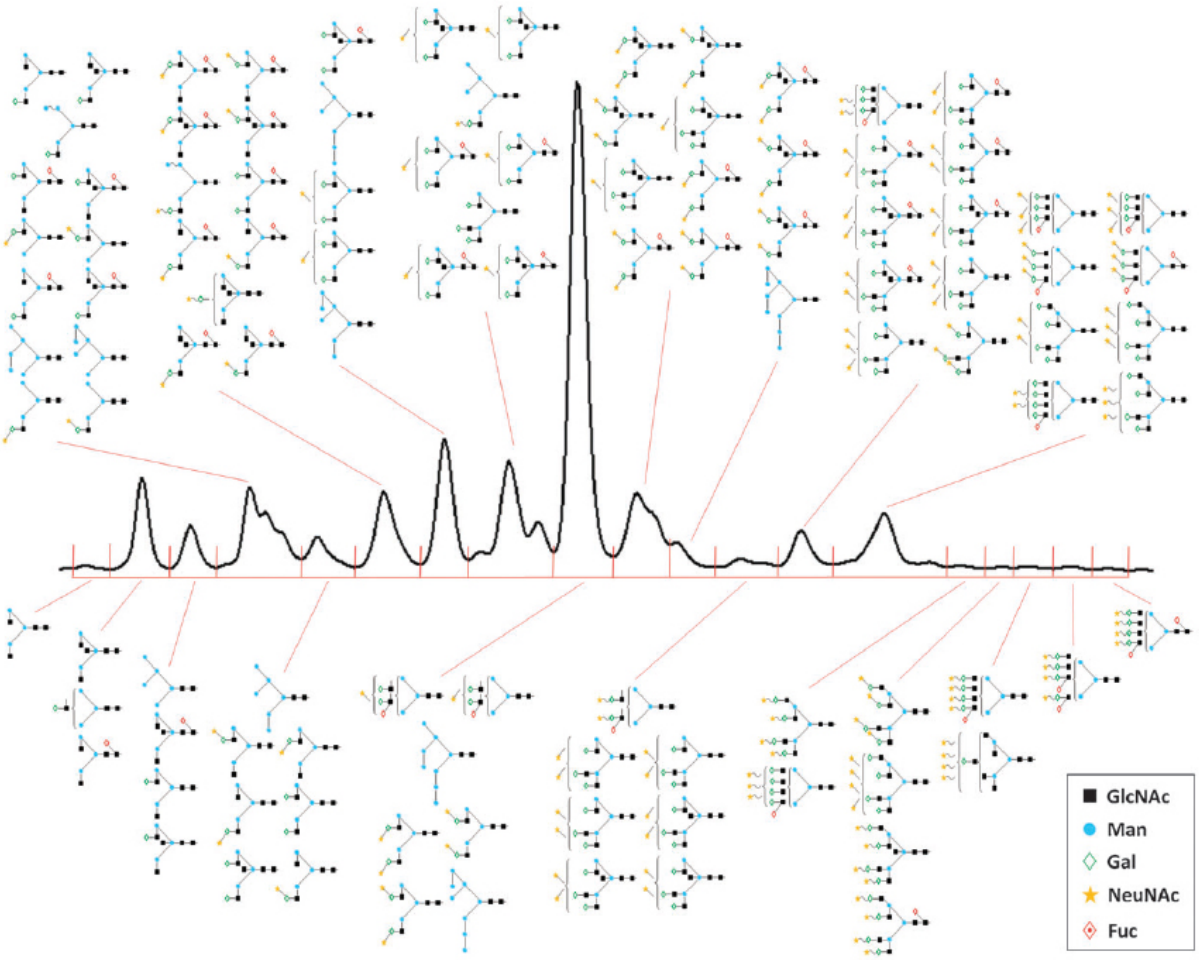


Figure 4. Example of a chromatographic peak division for glycan quantification. Chromatogram was obtained by HILIC and divided into 16 chromatographic peaks. Each peak contains more than one similar individual glycan structures as shown. The amount of glycans present in each peak is expressed as a percentage of the total integrated area of the chromatogram and calculated as the amount of total glycan structures in the peak/the total serum N-glycome; the percentages of all peaks add up to 100%. Monosaccharide abbreviations: Fuc - fucose; Gal - galactose; GlcNAc - N-acetylglucosamine; Man – mannose; NeuNAc - N-Acetylneuraminic acid. Adapted from Lauc & Zoldos (2010).

Three separate chromatographic methods were used to analyse the glycans: HILIC, HILIC of desialylated glycans and weak anion exchange high-pressure liquid chromatography (WAX-HPLC). HILIC analysis chromatograms were divided into 16 groups named GP1-GP16 (Figure 5A). HILIC of desialylated glycans was performed on released N-glycans after the removal of sialic acids by sialidase digestion treatment and the chromatographic division resulted in 13 groups of desialylated glycans named DG1-DG13 (Figure 5B). The individual glycan structures present in each of the chromatographic peaks of HILIC analysis and HILIC analysis after sialidase treatment are specified in Supplementary table 1. WAX-HPLC separated glycans

according to the level of sialylation, i.e. the number of attached sialic acids into monosialylated, disialylated, trisialylated and tetrasialylated (Figure 5C). In WAX-HPLC, compounds are separated and quantified depending on their charge density with higher charged compounds having longer retention times.

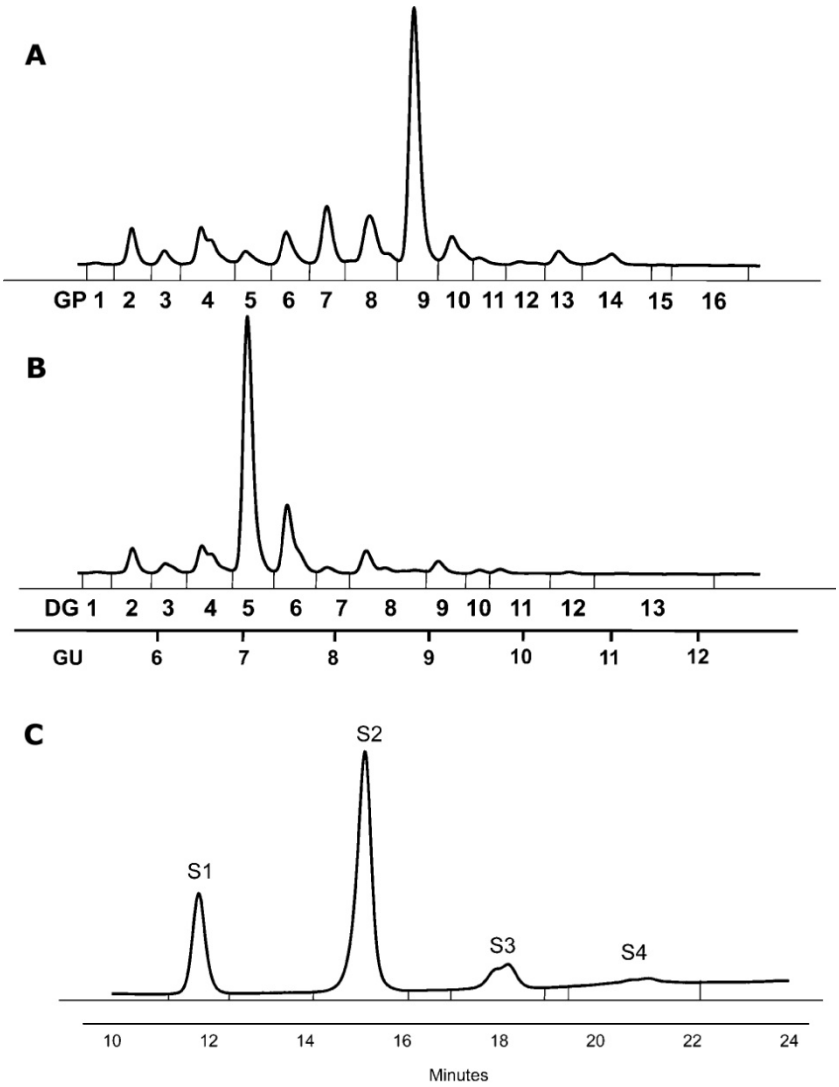


Figure 5. Plasma N-glycome chromatographic and quantification analysis. Typical chromatograms from HILIC (A), HILIC after sialidase digestion treatment (B) and WAX-HPLC (C) of N-glycans released from human blood plasma. These three chromatographic methods allow the division of a plasma N-glycome profile into 33 chromatographic peaks. N-glycans are separated into: 16 peaks with HILIC analysis (named GP1-GP16), 13 peaks with HILIC analysis after sialidase treatment (named DG1-DG13) and 4 peaks with WAX-HPLC analysis (named S1 - monosialylated, S2 - disialylated, S3 - trisialylated, S4 - tetrasialylated). Adapted from Saldova et al. (2012).

Additional glycan structural features, such as fucosylation, level of galactosylation, level of sialylation of biantennary structures and degree of branching, were approximated by adding the glycans sharing the same structural characteristic from either HILIC or HILIC after sialidase treatment integrated glycan profiles. A total of 13 glycan structural features were derived and are presented in Supplementary table 2.

HPLC analyses of plasma N-glycans were performed by collaborators in the Glycobiology Laboratory of Genos Ltd in Zagreb, Croatia, and in the National Institute for Biotechnology and Training (NIBRT) in Dublin, Ireland.

2.2.2. IgG N-glycans

IgG proteins were isolated and purified from plasma using a novel 96-well protein G monolithic plate followed by the release and labelling of N-glycans (Pucic *et al.*, 2011). Fluorescently labelled N-glycans were separated by ultra performance liquid chromatography (HILIC-UPLC).

There are two versions of this technique: the "in-gel" approach used for the quantification of IgG glycans in the Vis and Korčula populations; and the "in-solution" approach used to quantify the IgG glycans from the Orkney samples. The two approaches mainly differ in the methodology of the steps involved: the filtration of plasma before isolation and purification of IgG proteins was introduced in the in-solution version, the deglycosylation is done in solution conditions in the in-solution method as opposed to gel blocks used in the in-gel method, microcrystalline cellulose is used for solid-phase extraction to remove excess of 2-AB dye in the in-solution method while chromatography paper is used in the in-gel method. Overall, the in-solution method has shown to be less laborious, much faster and cheaper than the initial in-gel approach. It should be noted, however, that these differences in methodological procedures will lead to slightly different quantification results.

Individual glycan structures in the chromatographic peaks were identified by mass spectrometry. The amount of glycans present in each peak is expressed as a percentage of the total integrated area of the chromatogram and the percentages of all peaks add up to 100% for a single chromatogram.

IgG N-glycan chromatograms obtained with HILIC-UPLC were divided into 24 peaks named GP1-GP24 (Figure 6) and the composition of individual glycan structures contained in each peak is presented in Supplementary table 3. The minor peak GP3 was excluded from all the calculations because its value was significantly contaminated as explained in Pucic *et al.* (2011).

Additional glycan structural features were derived and approximated from the ratios of the 23 IgG original N-glycan peaks sharing similar structural features. A total of 54 glycosylation traits were derived and both their description and calculation formula are available in Supplementary table 4.

The analyses and quantification of IgG N-glycans were performed by collaborators in the Glycobiology Laboratory of Genos Ltd in Zagreb, Croatia.

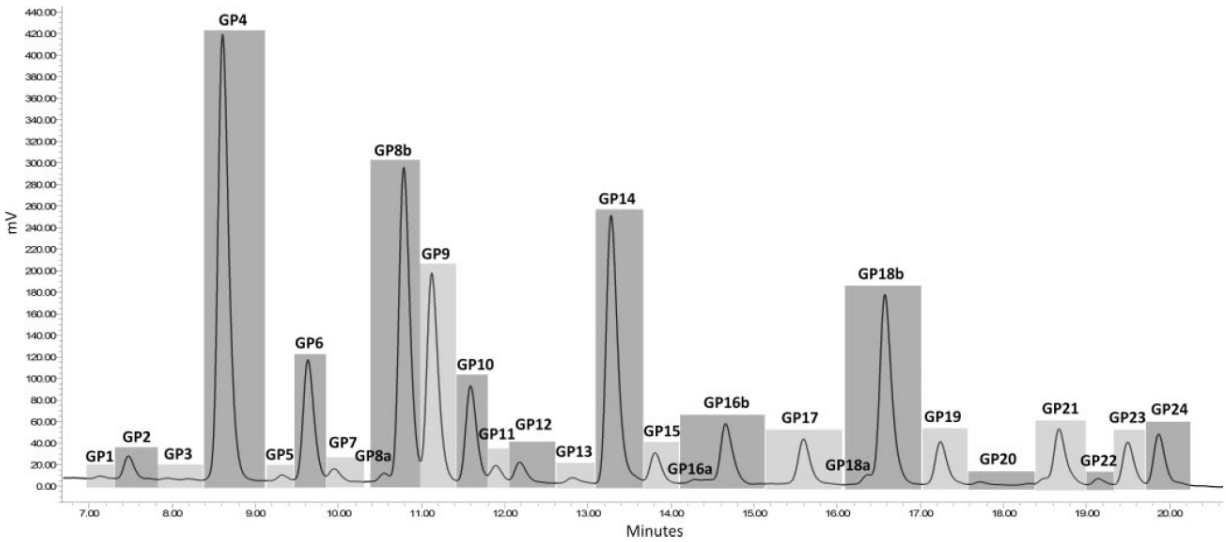


Figure 6. IgG N-glycome chromatographic and quantification analysis. The IgG N-glycome was separated into 24 chromatographic peaks (named GP1-GP24) by HILIC-UPLC. The amount of glycans present in each peak is expressed as a percentage of the total integrated area of the chromatogram; the percentages of all peaks add up to 100%. Adapted from Pucic et al. (2011).

2.3. Feature Data Sets

Each individual sample is represented by a plasma N-glycan profile, an IgG N-glycan profile, a set of phenotype traits and a list of genotypes for several SNPs.

2.3.1. Plasma N-glycan profile data

The human plasma N-glycome was separated by three chromatographic analyses into 33 different chromatographic peaks: 16 groups of glycans before desialylation from HILIC, 13 groups of desialylated glycans from HILIC after sialidase treatment and 4 groups of differently charged glycans from WAX-HPLC. An additional group of 13 glycan structural features were

derived based on the above measured chromatographic peaks. The plasma profile of an individual consists of a total of 46 traits divided into the 4 groups previously mentioned, further referred to as GP (n=16), DG (n=13), Sialos (n=4) and Structural (n=13) glycans.

2.3.2. IgG N-glycan profile data

The human IgG N-glycome was divided into 23 chromatographic peaks which were used to derive 54 additional glycan structural features. The IgG profile of an individual comprises 77 traits divided into 4 groups, further referred to as Initial (n=23), Charged (n=17), Neutral (n=14) and Neutral derived (n=23) glycans.

2.3.3. Phenotype data

Phenotype data is composed of personal and health-related data as well as physiological and biochemical traits. Personal data (such as name, age, gender and education level), lifestyle variables (such as smoking status and diet) and medical conditions (such as presence of certain diseases and drug intake) were collected based on extensive questionnaires. Several physiological traits were measured including height, weight, blood pressure, waist circumference and skinfold. Biochemical traits measured through biochemical analyses include levels of creatinine, uric acid, HDL, LDL, total cholesterol, triglycerides, insulin, fibrinogen and blood glucose.

Due to the fact that different phenotypes were available for each population, a set of phenotype traits collected for all populations was selected. The following 21 phenotype traits were considered for the present analyses: age, sex, systolic and diastolic blood pressure (Sys and Disys), total cholesterol (Cholest), HDL, LDL, triglycerides (Trigy), blood glucose, insulin, glycosylated haemoglobin (HbA1c), fibrinogen (Fibrin), creatinine (Creat), calcium, uric acid, albumin, body mass index (BMI), waist-to-hip ratio (WaistHip), FAT, waist and hip circumference (WaistCir and HipCir). In parenthesis are indicated the short names of the phenotypes which will appear in the figures shown in the present thesis.

The diabetes status of an individual was one of the medical conditions collected. The level of glycosylated hemoglobin (a marker for average blood glucose levels over prolonged periods of time; HbA1c in short) was used to classify the individuals into one of the three groups: non-diabetics (HbA1c < 6%), pre-diabetics (HbA1c 6-6.49% and no record of diabetes in medical history) and diabetics (HbA1c >= 6.5% or physician reported diabetes in medical history or

treatment with anti-diabetic medication). The diabetes status was available for a total of 3248 individuals (692 from Vis, 495 from Korčula and 2061 from Orkney) of which 2736 were assigned as non-diabetics, 233 as pre-diabetics and 279 as diabetics.

2.3.4. Genotype data

DNA samples were genotyped according to the manufacturer's instructions on Illumina Infinium SNP bead microarrays (HumanHap300v1for the Vis cohort, HumanCNV370v1 for the Korčula cohort and HumanHap300v2 for the Orkney cohort). Genotypes were determined using Illumina BeadStudio software (Lauc *et al.*, 2010a).

Approximately 300.000 SNPs were genotyped (referred further on as the all SNPs set), including about 900 SNPs known to be related with glycosylation (referred further on as the glycan-related SNPs set). Genotyping was successfully completed on 986 individuals from Vis, 944 from Korčula and 890 from Orkney.

2.4. Data Preprocessing

Real-world data tends to be incomplete (lacking attribute values of interest), noisy (containing errors or outliers) and inconsistent (containing discrepancies in codes or names) (Chakrabarti *et al.*, 2009).

Data preprocessing consists of an ensemble of techniques, including data cleaning, data integration, data transformation and data reduction, aimed to improve the quality of the data to be analysed. Data preprocessing methods are applied to rectify the data by filling in missing values, removing errors, correcting inconsistencies or transforming data into appropriate format for analysis. The use of these techniques is not mutually exclusive and usually a few of them are applied sequentially to the same data set.

The majority of studies concerning large size populations and the collection of various feature data sets for those populations frequently face the problem of having several records which are not complete. The reasons accounting for incomplete and inaccurate records vary from the impossibility of trait measurements and incorrect data entry to errors occurring during the methodological procedures.

The quality of data is of extreme importance and has a great impact on the accuracy and interpretation of the results for knowledge discovery. Therefore, data preprocessing is an important routine to consider/bear in mind prior to data analysis.

The four available feature data sets (plasma N-glycan profiles, IgG N-glycan profiles, phenotypes and genotypes) follow the real-world data behaviour and as such were subjected to a data preprocessing pipeline as described below.

2.4.1. Data quality control

Data quality control was performed in order to eliminate the most incomplete samples and decrease the amount of missing data. The quality control procedure was applied in the same manner for all populations.

Plasma and IgG N-glycan and phenotype data sets were filtered by removing samples missing 50% or more of the features in question. The removal of all samples lacking at least one trait was not applied because it would greatly reduce the size of the study populations and lead to loss of information.

Quality control of genotype data aims to filter out not only individuals with a small amount of genotype data but also SNPs which were not resolved for a large number of individuals. Genotyping quality control was performed on the basis of the following inclusion thresholds criteria: individuals were excluded when having a genotype rate less than 97%, i.e. with more than 3% of genotypes missing; SNPs were removed when having a call rate less than 95%, minor allele frequency less than 2% or Hardy-Weinberg equilibrium p-value less than $1x10^{-7}$.

While the IgG and phenotype profiles were rather complete in all populations, the plasma profiles of a large part of the Orkney samples was lacking more than 50% of data and the genotype data was mainly incomplete for Vis and Korčula samples. The influence of data quality control on population sample size is summarized in Table 1.

Table 1. Data summary for the study population cohorts. For each population the following values are indicated: the initial number of individuals, the remaining individuals after data quality control by feature data set and the final sample size after data integration. Pop.cohorts: Population cohorts; Plasma and IgG profiles: refer to plasma and IgG N-glycan profiles.

Pop. cohorts	**Plasma profiles**		**IgG profiles**		**Phenotypes**		**Genotypes**		Common Individuals[c]
	initial	after filtering[a]	initial	after filtering[a]	initial	after filtering[a]	initial	after filtering[b]	
Vis	1008	995	890	890	1008	1006	986	858	735
Korčula	969	949	914	914	969	959	944	887	823
Orkney	2095	1475	1770	1770	2095	2077	890	890	770

[a] Filtering was done by excluding samples with 50% or more of the features missing.

[b] Filtering was done by applying a missing rate per person of 90%.

[c] Individuals present in all four feature data sets within a population cohort.

2.4.2. Data integration

Gathering different kinds of data sets for a single study population of a large sample size clearly provides a significant amount of data (and information) for the analysis. However, dealing with these various data sets afterwards can be challenging and troublesome. Particularly, when preparing data for analysis a certain level of data inconsistency is often encountered, for instance it might happen that some samples are present in some of the data sets but nonexistent in the rest. In such cases, data integration is performed to properly combine data extracted from multiple sources into a coherent whole.

In order to achieve data consistency within a population, the individuals presenting all feature profiles were identified and their corresponding data selected to be used in the analyses. The final number of individuals after data integration is shown in Table 1. This step greatly reduced the number of available samples for each population, especially for Orkney where approximately half of the samples was not genotyped. Although global data integration was performed, pairwise integration, i.e., integrating data from only two feature data sets (for example, plasma profiles and phenotypes) according to the analysis to be performed might be a better approach since a smaller number of samples would be eliminated and, consequently, more samples would be considered for analysis.

Data integration and data quality control also affected the number of individuals with available diabetes status which was drastically reduced to less than half of the original number (Table 2).

Table 2. Data summary for the diabetes data set. Number of samples per diabetes group (non-diabetics, pre-diabetics and diabetics) and the composition of each group in terms of populations (Vis, Korčula and Orkney). In parenthesis is shown the number of samples for each group and population before the data quality control and data integration steps.

Diabetes status	Populations			Total for groups
	Vis	**Korčula**	**Orkney**	
Non-diabetics	449	270	585	(2736) 1304
Pre-diabetics	33	59	58	(233) 150
Diabetics	47	45	35	(279) 123
Total for populations	(692) 529	(495) 374	(2061) 674	(3248) 1577

The use of different SNP arrays for genotyping and the thresholds imposed in the quality phase control resulted in different set of SNPs available for each population. In order to ensure that the same core of SNPs was used throughout the analyses and be able to compare results between populations, the set of SNPs shared by all three populations was obtained and the corresponding genotype data extracted for each population. The final common set of SNPs comprises a total of 275895 SNPs, including 971 glycosylation-related SNPs.

2.4.3. Data normalization

Normalization is commonly used to obtain a data set where all the variables are within the same value range and can be fairly compared. Several data normalization procedures exist; however, it is possible that they might affect the outcome of the analysis.

Data normalization was used to adjust the values of the features to a common scale and allow a better comparison of features across populations. The approach selected to normalize the glycan and phenotype data was the median normalization which was intended to center the data to have median zero while keeping the data distribution specific of each population. The transformation was accomplished by subtracting the median value of each feature to the corresponding initial individual values.

2.4.4. Data correction

Data correction concerned the removal of batch effects present in IgG glycans and the age and sex correction of both plasma and IgG data sets.

The 23 Initial IgG glycans were corrected for batch effects resulting from the use of different plates in the IgG quantification analysis. First, a log transformation was applied to each glycan group to obtain normally distributed variables. Second, batch correction was performed using a linear mixed model where the methodological sources of variation (plates and columns in the plates) were described as random effects. The estimated batch effect (random component) was disregarded in the calculation of the corrected values which express only the normal biological variation of glycans. Third, the exponentials of the corrected values are taken to return the values to their original scale. The *lmer* function as implemented in the *lme4* package for R was used for the purpose of the batch effect correction (Bates *et al.*, 2013).

N-glycans have been reported to be associated with age and gender and it has been suggested that the influence of these variables should be taken into account when investigating the relationship between glycans and phenotype traits and in genome wide association studies (Ding *et al.*, 2011; Huhn *et al.*, 2009; Knezevic *et al.*, 2009). In this way, it is excluded the possibility that observed associations between certain features and glycans are a reflection of a background influence of aging or gender upon these features. Plasma and IgG N-glycan data was corrected for age and sex to eliminate any dependencies of these two variables with N-glycans. The correction was performed with a generalized additive model and the resulting residuals were considered for analysis. Generalized additive models use a local scoring algorithm which iteratively applies a smoothing function to the data, similar to a locally weighted regression. For each predictor variable in the model, the smoothing function fits the data by taking into account the neighbourhood of each point being fitted. The *gam* function as implemented in the *mgcv* package for R was used for the age and sex correction of glycans (Wood, 2011).

The age and sex correction was done in two different ways based on whether the purpose of the analysis was to compare the glyco-phenotype characteristics of populations or to find general associations between glycans and phenotypes/SNPs. To compare populations and search for particular glyco-phenotype features capable of distinguish them, glycans were first normalized independently for each population and then the age and sex correction was applied to the pool of the three populations, resulting in a data set having a total of 1990 individuals. Performing the age and sex correction while considering the three populations as a whole assumes that the age covariate has the same distribution across populations. In this way, the age effect is kept constant across populations and noticeable differences between populations can be regarded as a consequence of the population structure itself. To investigate potential association patterns existing between glycans and phenotypes and between glycans and SNPs that can be

generalized, the age and sex correction was done separately for each population, resulting in a data set having a total of 2063 individuals. This approach ensures that the population effect on data is removed and, consequently, analysis intended to search general association trends between variables can be compared across populations and can be done on the three populations as a whole.

2.4.5. Data removal of outliers

Outlier samples were removed after age and sex correction for each glycan measure to account for errors in quantification and to eliminate individuals not representative of normal variation within populations. An individual was classified to be an outlier if its residual measure for the trait was more than 4 standard deviations away from the mean. Although in common practice a data point is considered an outlier if the corresponding residual is 3 or more standard deviations from the mean, a less conservative threshold of 4 was applied in order remove only extreme outliers. The age and sex correction model was again fit to the data sets without the outliers.

2.4.6. Data imputation

Several statistical tools and algorithms either discard by default any record that has a missing value or require complete records. To overcome these problems imputation techniques are often carried out. Imputation is the process of replacing missing values with a probable value based on the rest of the available data while preserving all the records in the data set.

Although the major part of missing data in the four feature data sets was removed when data quality control was performed, incomplete records with minimal amount of missing data remained. Missing data in N-glycan profiles and phenotypes was handled by imputing the missing values with the median of the corresponding trait whereas missing genotypes were replaced with the most common genotype found for each SNP. Similarly to data normalization, various techniques can be applied to impute missing data and the approach chosen here was just one technique among the various existing possibilities.

2.4.7. Data comparison

The in-solution and in-gel methods used for high-throughput quantification of IgG N-glycans were compared by analysing 473 samples from the Orkney cohort which had the IgG N-glycan profiles measured with both methods. The agreement between the two methods was assessed by

computing their linear correlation after age and sex correction of the data. The *lm* function of the *stats* package for R environment was used to compute the correlation coefficients.

2.5. Computational Tools

2.5.1. R statistical package

R is an open source programming language and environment for statistical computing and graphics (R Core Team, 2013). R integrates a wide range of methods designed to explore data in a variety of ways, such as modelling, classification and statistical analysis, and to graphically display data in a comprehensive manner to facilitate and improve data evaluation. R core functionality is extended by allowing users to define new functions and via additional packages which are freely available and provide groups of functions developed for specific analysis.

The exploratory analysis of the data was mainly performed in the R programming environment (version 3.0.1) using several specialized packages according to the needs of the analysis to be carried out.

2.5.2. PLINK

PLINK is a free, open-source whole genome association analysis toolset developed to improve and facilitate computational analyses of large-scale genotype data (Purcell *et al.*, 2007).

PLINK was used to perform the genotype quality control through the commands reserved to specify the inclusion thresholds: --*mind* for the missing rate for person (value of 0.03), --*geno* for the missing rate per SNP (value of 0.05), --*maf* for the minor allele frequency (value of 0.02) and --*hwe* for the Hardy-Weinberg test (value of 0.0000001).

2.5.3. Perl programming language

Perl is a high-level and general-purpose programming language initially developed for text processing but rapidly extended to areas like system administration, web development and graphical programming (Perl, 2013).

Perl was used to write auxiliary scripts mainly intended to perform tasks of data manipulation with the purpose of transforming and modifying the data format and/or structure into a suitable format for following computations.

2.6. Computational Methods/Algorithms

Several machine learning, data mining and statistical methods as well as different graphical representation approaches were applied to analyse, visualise and model the data and to derive relevant biological information.

2.6.1. Nearest neighbours computation

Although the human plasma glycome is very stable and is similar among most individuals, some individuals with a glycan profile showing deviations from this normal glycan profile were observed and referred to as outliers. To form limited size groups of individuals sharing the same profile characteristics as the outliers, computational identification of groups of nearest neighbour individuals was carried out as described below (Papadias *et al.*, 2004).

Glycan profiles were normalized for age and gender differences and scaled to the mean residuals of linear regression. Individuals presenting the most similar glycan profiles to single identified outliers were determined using a consensus scoring of pairwise distances between vectors containing measured glycan values. Basically, the five nearest neighbors (i.e. the individuals with the smallest respective profile distances) were calculated for each outlier using five distance calculation methods: maximum value (maximum difference in any coordinate dimension); Manhattan (city block); Euclidean (square root of the sum of squared vector coordinates); Canberra (sum of differences between the vector coordinates); and Minkowski generalized distance of order 4 (fourth root of the sum of vector coordinates raised to the fourth power). Neighbors occurring in a group of five nearest neighbors using at least two different methods were selected as true neighbors and treated as a group.

2.6.2. Clustering

One of the most important goals of unsupervised learning is to discover meaningful clusters in data. Cluster analysis, an approach to unsupervised learning, aims to discover groups, or clusters, of data points which belong together because they are in some way similar to each other. Although there are hundreds of published clustering algorithms, there is no correct algorithm that can be applied to all cluster-related problems. Instead, the most appropriate algorithm should be chosen based on the capacity of its underlying cluster model to fit the data set properties in question (Andreopoulos *et al.*, 2009).

A popular clustering algorithms employed in a broad range of areas is the K-means algorithm which attempts to partition the data points into *k* clusters so as to minimize the sum of squared distances between the data points and their nearest cluster center. A major drawback of K-means and similar algorithms is that the number of clusters (*k*) is often required to be specified prior to the analysis which is not always desirable. Another issue concerns the assignment of the initial cluster centers which is randomly performed; an inadequate initial choice of centers might lead to poor results. This is the reason why K-means is often rerun several times with different initialization centers in order to be able to find an acceptable solution.

Affinity propagation is a clustering approach that simultaneously considers all data points as potential centers, or exemplars, and recursively exchanges value-encoded messages along the network formed by these data points until a high-quality clustering solution is achieved, as illustrated in Figure 7A (Frey & Dueck, 2007). Affinity propagation takes measures of similarity between data points as input and transmits two types of messages between data points which are updated during the message-passing procedure. The "responsability" message, defined as $r(i,k)$, is sent from data point *i* to candidate exemplar *k* and reflects the accumulated evidence for the affinity that point *i* has for choosing *k* as its exemplar (Figure 7B). The "availability" message, defined as $a(i,k)$, is sent from candidate exemplar *k* to point *i* and reflects the accumulated evidence for how appropriate it would be for point *i* to choose point *k* as its exemplar (Figure 7C). Combining these two messages allows the identification of exemplars. Through this dynamic process of exchanging messages, the appropriate number of centers (and thus clusters) emerges iteratively without having to specify it beforehand. Besides the similarity matrix, the individual tendencies of data points to become exemplars, called input preferences, can be specified in the affinity propagation clustering. Input preferences can be chosen individually for each data point or can be a shared value among all data points (meaning that all data points are equally suitable as exemplars). The value of the input preferences influences the number of clusters produced and is usually set to the median of the input similarities (resulting in a moderate number of clusters) or to their minimum (resulting in a small number of clusters).

The affinity propagation clustering was applied to analyse the internal structure of the population cohorts and to explore the glyco-phenotype signatures of the observed clusters.

The *apcluster* package for R environment implements the affinity propagation clustering (Bodenhofer *et al.*, 2011). The *apcluster* function was employed when the analyses were performed without a pre-defined number of clusters. The input preferences were choosen as a

common value for all data points and its optimal value for clustering was searched by setting the *q* parameter to 0 and 0.5, corresponding to the minimum and median values of the input similarities, respectively. The *apclusterK* function was used to analyse a specific number of clusters. The desired number of clusters was set with the *K* parameter and the parameter *prc* which controls the percentage that the number of clusters is allowed to deviate was set to 0 to have exactly *K* clusters. In both cases, the measure of similarity between samples was taken as the negative Euclidean distance computed based on the glycan or phenotype profiles.

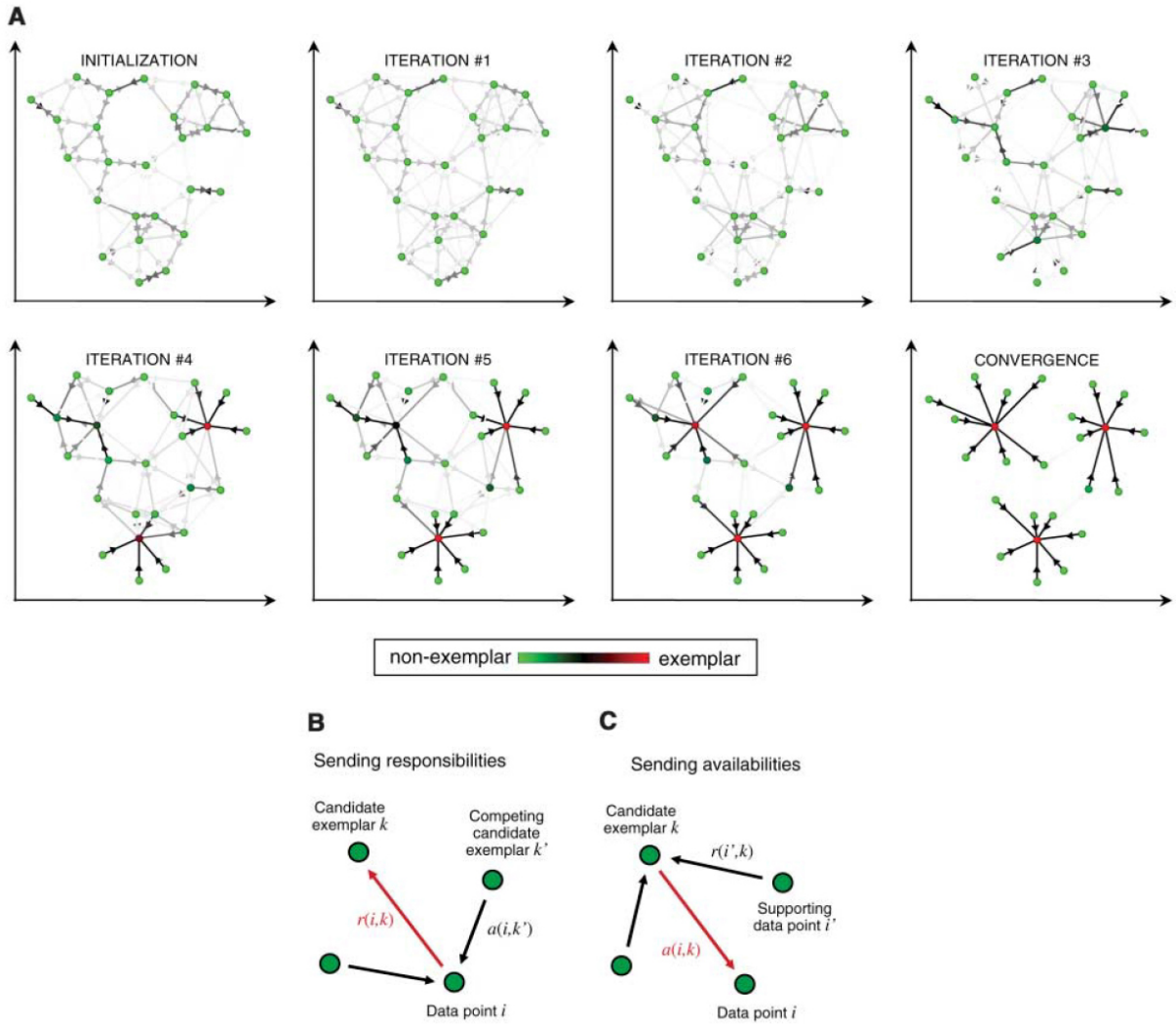


Figure 7. Principles of the affinity propagation algorithm. (A) Gradually emerging clusters during the message-passing procedure. (B) The "responsability" message, r(i,k), sent from data point i to candidate exemplar point k. (C) The "availability" message, a(i,k), sent from candidate exemplar point k to point i. Adapted from Frey & Dueck (2007).

2.6.3. Principal component analysis and partial least squares regression

Multivariate statistical analysis involve modelling data sets often with a large number of explanatory variables which do not have equal relevance to the model. Additionally, the organization of such high-dimensional data cannot be spatially visualized. Thus, variable selection and dimension reduction are important tasks in multivariate analysis.

Principal Component Analysis (PCA) and Partial Least Squares (PLS) are multivariate techniques for dimension reduction and are particularly useful when the explanatory variables display a high degree of correlation (Maitra & Yan, 2008). The principle of both methods is to convert a set of correlated explanatory variables to a set of independent synthetic variables (defined as linear combinations of the initial variables) by transforming the data into a new coordinate system. Despite the fact that the basic idea is similar, it should be noted that PCA is a type of unsupervised analysis used to explore and visualise a single set of variables (explanatory), while PLS is a supervised analysis for correlating two sets of variables (explanatory and response). The main characeristics of each method are outlined below.

PCA determines linear combinations of the explanatory variables, called principal components, that explain most of the data variability with the first principal component accounting for as much of the variability in the data as possible followed by the other components ordered by the amount of variance explained (Mörtsell & Gulliksson, 2001). In this way, PCA projects the data into a lower and more tractable dimension without losing too much information.

PLS decomposes simultaneously explanatory and response variables into linear combinations, called latent variables, such that the covariance between them is maximized . In an iterative process, PLS seeks for the latent structure in the explanatory variables that best explains the latent structure accounting for the maximum variance in the response variables (Tobias, 1995). Partial Least Squares Discriminant Analysis (PLS-DA) is a variant of PLS used when there is a single response variable.

The PCA and PLS-DA methods were used in an attempt to differentiate the samples according to populations based on plasma profiles, IgG profiles and phenotypes and to summarize the differences that most influence the achieved separation.

The *mixOmics* package for the R environment, dedicated to the integrative analysis of ‘omics’ data, contains an implementation of these two techniques (Le Cao *et al.*, 2009). The functions *pca* and *plsda* were used to perform the PCA and PLS-DA analyses, respectively. The package

also provides several integrative techniques to analyse highly dimensional data sets as well as numerous possibilities of graphical representations to help interpret the results.

2.6.4. Discriminant analysis of principal components

Multivariate statistical approaches have been applied to investigate the genetic structures of biological populations. In such studies, the aim of multivariate methods is to detect a set of alleles that best reflects the genetic variation present among the analysed individuals. This genetic variability can be decomposed into two components: the between-group variability concerning the genetic structure of populations and the within-group variability related to the general random genetic diversity existent (Figure 8A).

Approaches like the previously mentioned PCA seek to describe the overall variability of data (including both between and within-group variability) and tend to overlook the divergence between groups (Figure 8B). Discriminant Analysis (DA) is an alternative method that has almost the opposite rationale in the sense that it tries to model genetic differences by maximizing the between-groups variability while minimizing the within-group variability (Figure 8C). The linear combinations of explanatory variables resulting from DA are called discriminant components or functions. Unlike PCA which does not provide a group assessment measure essential to study population structures, DA is used when groups are known a priori and is able to predict category membership. However, the performance of DA when applied to genetic data is compromised by the inherent characteristics of the data sets such as the larger number of SNPs when compared to the number of samples and the high level of correlation present between SNPs.

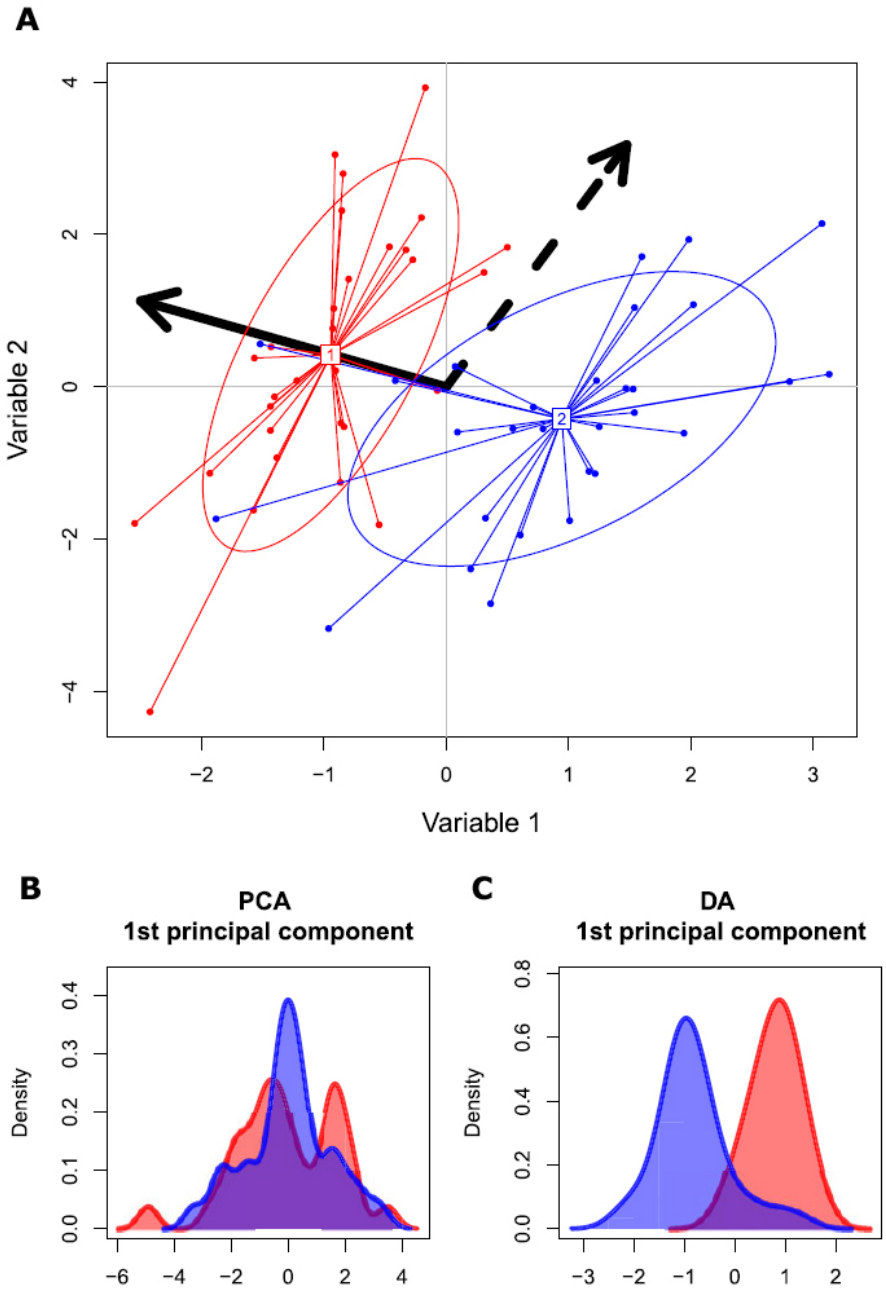


Figure 8. Fundamental difference between PCA and DA. (A) The diagram shows the essential difference between Principal Component Analysis (PCA) and Discriminant Analysis (DA). Individuals (dots) and groups (colours and ellipses) are positioned on the plane using their values for two variables. In this space, PCA searches for the direction showing the largest total variance (doted arrow), whereas DA maximizes the separation between groups (plain arrow) while minimizing variation within group. As a result, PCA fails to discriminate the groups (B), while DA adequately displays group differences (C). Adapted from Jombart et al. (2010).

Discriminant Analysis of Principal Components (DAPC) is a new multivariate method for the analysis of genetically structured populations developed to allow the DA principle to be applied in the analysis of large genetic data (Jombart *et al.*, 2010; Rasmussen *et al.*, 2011). DAPC combines the capabilities of both PCA and DA for a better discrimination of genetically related individuals into pre-defined groups. PCA is initially employed for dimensionality reduction and for elimination of correlations between variables, in order to obtain a small number of uncorrelated variables which can then be subjected to DA. The DAPC method allows for a graphical assessment of between-population differentiation through scatterplots of discriminant functions and derives group membership probabilities which can be considered as indicators of how clear-cut the population clusters are. Moreover, DAPC provides a measure of allele contributions to the structures identified which can be used to fetch the alleles that most differ across populations.

The DAPC technique was applied in an attempt to classify the three population cohorts and to investigate the genetic background behind it.

The *dapc* function of the *adegenet* package for the R software implements the DAPC method and was used to perform the analyses (Jombart, 2008; Jombart & Ahmed, 2011). The optimal number of axes to retain in the PCA step of the DAPC algorithm (defined by the n.pca parameter) was estimated using the *optimalPC* function and the obtained result used in the DAPC. In some cases, the analysis was also performed with a different number of principal components for comparison purposes.

2.6.5. Random Forests and Random Jungle

Random Forests (RF) are an effective machine learning algorithm used for both problems of supervised (classification and regression) and unsupervised learning (Shi & Horvath, 2006; Svetnik *et al.*, 2003). RF grows an ensembl of classification trees which individual results are aggregated to obtain the final predictions. The layer of randomness in the forests is introduced by two main aspects in which the RF trees differ from the standard decision trees: random inputs, each tree is independently constructed using a bootstrap sample from the data set (know as the bagging method); and random features, each node of the tree is split using the best among a subset of variables (predictors) randomly chosen at that node. The construction of the individual trees in RF is depicted in Figure 9 and summarized below. Considering a data set having N samples and M predictor variables:

1. Draw with replacement a bootstrap sample consisting of N samples from the original data.
2. At each node in the tree, randomly select m variables from the entire set of M posssible variables.
3. Find the best split at that node among the m randomly selected variables.
4. Iterate the second and third steps until the tree is fully grown.

A specific number of trees can be achieved by repeating steps 1 to 4 a desired number of times (for the theoretical background behind RF see Breiman (2001)).The number of variables randomly sampled as candidates at each node and the number of trees in the forest are the only parameters that need to be specified when running the algorithm. The prediction of a new sample is done by running down its corresponding vector of variables through each of the grown trees in the forest. Each tree will give its own classification for the new sample and the forest will choose the classification having more votes. Besides yielding a classsification result, RF additionally provides a measure of the importance of each predictor variable; an useful feature to estimate the contribution of the variables to the classification.

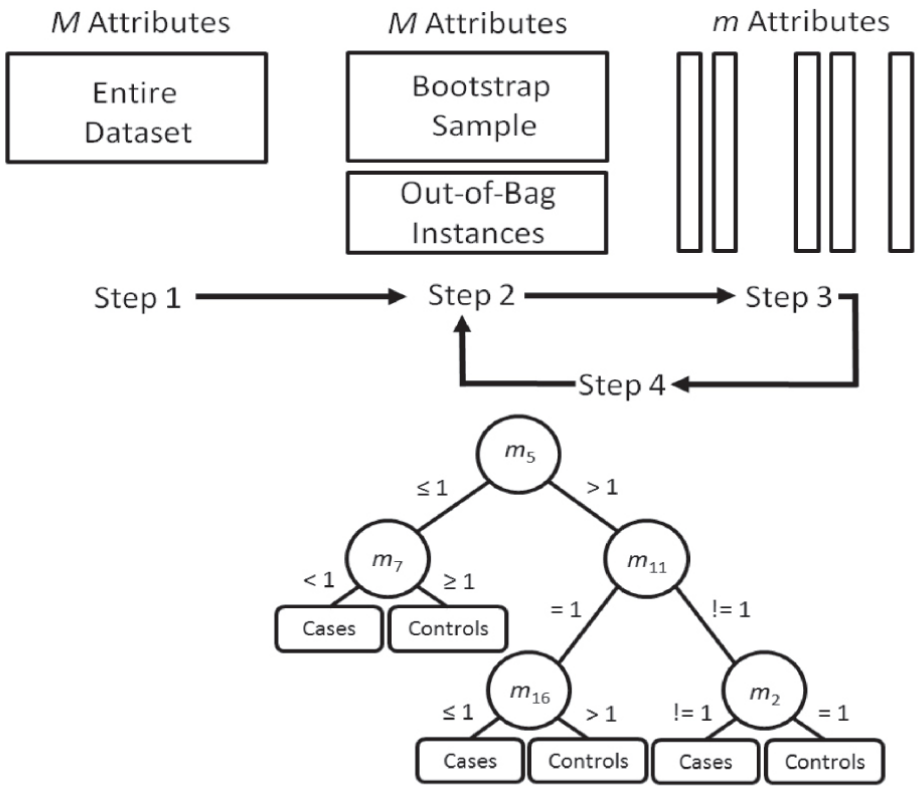


Figure 9. Diagram of the Random Forest algorithm. The description of each step is described in the main text. Adapted from Moore et al. (2010).

RF was employed to undertake the classification of distinct populations and the classification of diabetes status based on glycan profiles and phenotypic data. Plasma glycans, IgG glycans and phenotypes feature data sets were separately used as predictor variables while population and diabetes classes were used as response variables. For the classification problem, the samples were divided into a training set contaning 70% of the samples and a testing set containing 30% of the samples.

The *randomForest* package for R environment implements the algorithm described and was used to address the two classification problems (Liaw & Wiener, 2002). The number of trees to grow was set to 5000 (defined by the *ntree* option) and the number of variables randomly selected at each split was left to default that is sqrt(M) where M is the total number of variables considered (defined by the mtry option).

Random Jungle (RJ) is a recently developed alternative which allows the rapid analysis of large scale data present in genome-wide association studies (Schwarz *et al.*, 2010). RJ implements all the features available in the original RFs but it is structured and designed to analysed large data sets. Additionally, RJ is able to perform on multiple CPUs when available. When applied to genome-wide data, the computational performance of RJ in terms of computing time and memory usage were shown to be superiror to those of the original RF implementations while still yielding valid results.

RJ was employed in three different scenarios: the classification of distinct populations, the classification of diabetes groups and the investigation of possible associations between glycan profiles/phenotypes and SNPs. In all cases, genotypes were taken as predictor variables.

The *RandomJungle* software implements this improved version of RF and is freely available for download (Random Jungle, 2013). In the classification problems the *rjunglesparse* function was used while in the regression problem the *rjungle* function was applied. The *rjunglesparse* is the same program like *rjungle* but uses less memory and data values can only be 0, 1, 2 (and 3 as missing coding). The RJ parameters defined for each problem are shown in Table 3.

Table 3. Random Jungle algorithm parameters. Random Jungle was applied to three case problems: the classification of distinct populations, the classification of diabetes groups and the prediction of glycan levels based on the genotype. The tree type (or classifier type) and the number of trees in the jungle were set according to each problem while the number of predictor variables randomly sampled at each node of the trees was left to default in all cases. Below the name of the parameter in parenthesis are the corresponding options used when invoking the *rjungle* or *rjunglesparse* commands.

Case problems	**Random Jungle parameters**		
	Tree type[a] (-y)	Number of trees [b] (-t)	Number of sampled variables[c] (-m)
Classification of Populations	1	800/1000	default
Classification of Diabetes	1	800	default
Regression on glycan levels	3	100	default

[a] The tree type was set to y=1 for classification with numeric predictor variables and categorical response variables and to y=3 for regression trees with both numeric predictor and response variables.

[b] In the classification of populations case, the number of trees was chosen to be t=800 when using all SNPs and to be t=1000 was set when using the set of GlycansRelated SNPs.

[c] The default is the square root of the number of predictor variables.

2.6.6. Correlation adjusted scores

The correlation-adjusted t-score (CAT score for binary responses) and the correlation-adjusted marginal correlation (CAR score for quantitative responses) are two multivariate statistics recently introduced (Zuber & Strimmer, 2009). These two measures are multivariate generalizations of the standard univariate test statistics that explicitly included in their formulation the correlation existent among SNPs. Although initially suitable to analyse only relatively large data sets, new improvements in the algorithms currently allow their computation on large scale data. The variable selection based on these measures is shown to be highly efficient and to even outperform both uni- and multivariate competing approaches (Zuber *et al.*, 2012). Additionally, the squared scores of these statistics can be regarded as natural measures for SNP importance and the cumulative sum of SNP importance can be regarded as the coeficient of determination (proportion of phenotypic variance explained by SNPs).

These adjusted scores were applied to the classification of distinct populations, the classification of diabetes groups and the investigation of possible associations between glycan profiles/phenotypes and SNPs.

The *care* package for R software implements the original and improved versions of the measures and was used for the above mentioned analyses (Zuber & Strimmer, 2011).

2.6.7. Genome-wide efficient mixed model association

Despite being widely used in genetic analyses with rather overlapping purposes, linear mixed models (LMM) and sparse regression models have a quite different rationale. LMM applied to polygenic modelling assume that every genetic variant affects the phenotype, whereas sparse regression models assume that a relatively small proportion of all variants affect the phenotype. These two different assumptions will yield different results depending on the real genetic background of the phenotype.

Bayesian sparse linear mixed model (BSLMM) is a hybrid type of modelling that combines the advantages of both LMM and sparse regression models (Zhou *et al.*, 2013). The model behind can be interpreted as assuming that all variants have at least a small effect and that a part of the variants have an additional effect. BSLMM yields two important estimation measures: the total proportion of variance in phenotype explained by both random and sparse effects together, denoted as PVE, and the proportion of genetic variance explained by the sparse effects terms (i.e. by the additional effects of certain variants), denoted as PGE. Although the PVE estimation can also be obtained with LMM and sparse regression models, PGE is a feature specific of BSLMM. These estimates can help in the persistent problem of "missing heritability" by unveiling new potential effects of variants and, thus, contributing to a better understanding of the underlying genetic architecture of complex diseases.

The univariate linear mixed model and the bayesian sparse linear mixed model were applied to explore associations between glycan profiles and SNPs.

The *GEMMA* software (GEMMA stands for Genome-wide Efficient Mixed Model Association algorithm) implements both algorithms and is freely available for download (GEMMA, 2013; Zhou & Stephens, 2012). The main *gemma* command was run with the *-bslmm* option to fit a BSLMM (with sampling-related parameters set to w=1000 and s=1000) and with the *-lmm* option to perform association tests with a linear mixed model.

2.7. Statistical Methods

A quantile-quantile plot (known as Q-Q plot) was used for assessing whether the glycan variables were approximately normally distributed (data not shown). Since the majority of glycans showed a non-normal distribution, the nonparametric Wilcoxon rank-sum test (also called Mann-Whitney U test) was used to assess the statistical significance of pairwise

differences between glycan and phenotype levels of the particular groups analysed. The built-in functions corresponding to the mentioned tests were used as available in the *stats* package for R environment.

Bonferroni correction was applied to adjust p-values derived from multiple statistical tests. The corrected significance level varied according to the analyses performed and is indicated through the text whenever the analyses are described.

3. RESULTS

3.1. Data Preprocessing/Analysis Pipeline

In order to improve data quality, it is not uncommon to apply certain preprocessing methods to the data of interest prior to data analysis. The choice of the preprocessing methods to be used depends on the nature of the data and varies with the type of analyses to be performed. The feature data sets analysed in the present study – plasma profiles, IgG profiles, phenotypes and genotypes – were subjected to a data preprocessing pipeline which included data quality control, data integration, data normalization, data correction and data imputation, as described in detail in section 2.4.

Additionally to the above described preprocessing methods, a comparison of the gel and solution methods used for IgG glycan quantification was performed. Since the IgG glycan profiles were measured with the gel method for Vis and Korčula cohorts and with the solution method for Orkney, an evaluation of the agreement between the two methods was necessary to allow a proper comparison and interpretation of results from analyses involving the IgG profiles of the three populations.

For the purpose, the IgG glycan levels measured with both methods for a small set of Orkney samples were quantitatively compared. Several glycan groups presented a smaller range of values with the solution method than with the gel method, such as IGG1, IGG2, IGG16, IGG19, IGG20 and IGG21 (Supplementary figure 1). This behaviour was also observed when comparing the raw data of the three populations with Orkney samples showing lower glycan levels than Vis and Korčula (data not shown). The results are not totally unexpected in the light of the differences in methodology (explained in section 2.2.2) and the type of glycan structures present in each peak. On the one hand, IGG1 and IGG19 are themselves low intensity peaks and minimal integration inaccuracies in both quantification procedures could account for the differences observed. On the other hand, the plasma filtration step introduced before the isolation of IgG in the solution method reduced the non-specific binding of proteins other than IgG. As a consequence, peaks containing glycan structures present in proteins other than IgG would be expected to show decreased values in the solution method. This is the case of IGG16 and IGG20 that include glycan structures present not only in IgG but also in transferrin proteins which were likely to be eliminated during plasma filtration.

The correlation between the measures obtained with the two methods was further computed and correlation coefficients above 0.7 were obtained for the majority of the peaks (represented as red lines in Figure 10). Peaks showing a correlation coefficient lower than 0.45 mainly correspond to those peaks presenting a lower range of values with the solution method. However, it has been argued that the correlation coefficient is a misleading measure of the agreement between two clinical measurement methods and that alternative measures and graphical techniques should be used instead (Bland & Altman, 2003). A first problem pointed out is the fact that correlation depends on the range of the variables, i.e., it will vary if different group of subjects with different measures are selected. A second issue is that correlation looks at the degree of association between two variables, not the agreement between them. In other words, a good correlation is achieved if both measurements lie along any straight line while a good agreement is obtained only if data is distributed along the line of equality. In the comparison of the IgG glycan quantification measures, the lines of equality are similar to the correlation regression lines for peaks with high correlation values except for IGG2 (represented as green lines in Figure 10). The lines of equality for IGG2 and peaks with low correlation coefficients show a clear bias of the data points to lie on the right of the line of equality which confirms the tendency for the gel method to exceed the solution method for these peaks.

Plotting the difference between the measurements by the two methods against their mean has been proposed as a more informative alternative to the simple scatter plot of one method against the other in assessing between-method differences (Bland & Altman, 1999). Such a plot allows the examination of the relationship between the error measurement (estimated as the difference of values) and the true value (estimated as the average of values). An increase in the differences of the two IgG quantification methods as the magnitude of the glycan measurement increases is noticeable for the IgG peaks where a bias for the gel method to have higher values than the solution method was previously observed. A similar behaviour is shown by the IGG11 peak which, however, did not display significant differences between the range values of the two methods. For the rest of the peaks the differences did not vary in any systematic way over the range of measurements.

A certain lack of agreement between the gel and solution IgG glycan quantification methods is inevitable since they differ in the glycan preparation procedures and in some steps of the quantification analysis. The comparison analysis carried out suggested that in the majority of peaks there is a good agreement between the IgG quantification methods. Nonetheless, some peaks show a tendency to have higher measurements with the gel method than with the solution

method with these differences increasing with the increase in the magnitude of measurements. Such dissimilarities between methods should be taken into account when interpreting results derived obtained in the analyses of IgG glycans. In future studies, perhaps it would be advisable to seek for a formula that could enable the transformation of values between the two methods and allow a more accurate comparison of results.

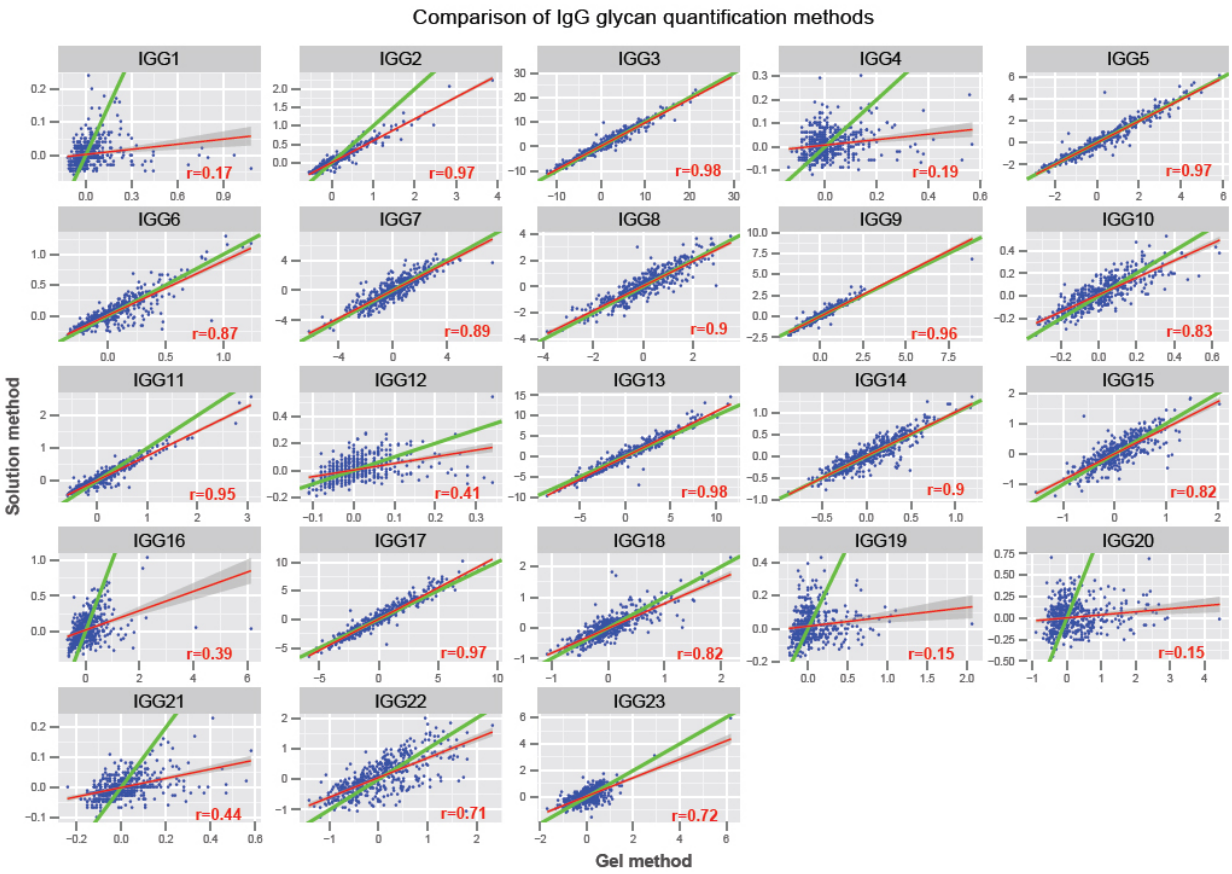


Figure 10. IgG glycan quantification measurements by solution method versus gel method. The correlation regression line is displayed in red and the corresponding correlation coefficient annotated on the bottom right corner of each graph. The line of equality is displayed in green. For the peaks having high correlation between the two methods, the correlation regression line and the line of equality are close to each other and even overlap in some cases. For the remaining peaks, the line of equality indicates a bias for the gel method to present higher values than the solution method.

3.2. Common aberrations from the normal human plasma N-glycan profile

Glycan profiles are rather similar in the majority of individuals; however, deviations from this normal glycan profile might occur due to patho-physiological conditions. Individuals having significantly different glycan profiles than the so called “normal profile” were identified while

analysing the plasma N-glycan profiles of 1991 individuals from Vis and Korčula cohorts. Six major outlying glycan features were observed; an example of the normal glycan profile and five of the aberrant profiles are shown in Figure 11.

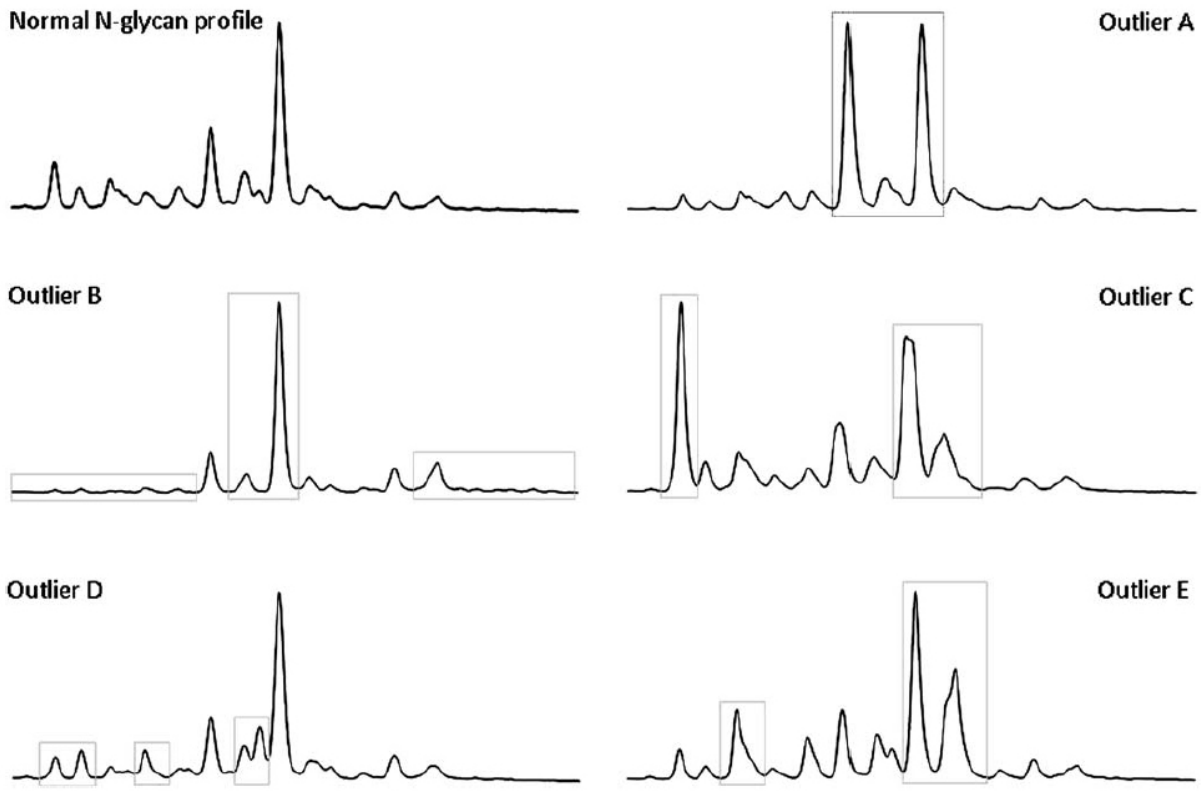


Figure 11. Normal and aberrant plasma N-glycan profiles. Examples of deviations from the normal glycan profile: outlier A, individual with elevated A2G2S1 glycans; outlier B, individual with glycan changes that mirror premature aging; outlier C, individual with elevated biantennary nongalactosylated glycans; outlier D, individual with elevated biantennary monosialylated glycans; outlier E, individual with increased core fucosylated glycans. Adapted from Pucic et al. (2010).

In order to investigate the possible causes leading to these aberrant profiles, limited size groups of individuals sharing the same profile characteristics as the outliers were formed and their phenotypic characteristics compared. While four of the outliers presented the same profile and were treated as a group, the manual inspection and simultaneous comparison of glycan profiles for the other outliers would be an impracticable task. In this case, computational methods were used to identify the nearest neighbours of these outliers, i.e., the individuals showing the most

similar glycan profiles (as previously described in section 2.6.1). The performed analyses were published in Pucic *et al.* (2010).

Subsequently, each of the six groups was analysed for the presence of common phenotypic characteristics among the individuals. In some groups the individuals shared certain clinical conditions, like renal problems, whereas in other groups the individuals were apparently healthy, demonstrating the existence of specific glyco-phenotypes that in some cases might represent risk factors for the development of specific diseases (Pucic *et al.*, 2010).

These groups were subjected to further analyses intended to explore the contribution of the genotype to group structuring. For the purpose, PCA, discriminant analysis of principal components and Random Jungle methods were applied to genotype data. The PCA and discriminant analysis of principal components did not reveal distinct clusters corresponding to the groups (data not shown). This lack of structure was further confirmed by the poor classification of groups (error of 84%) achieved with the Random Jungle algorithm. Altogether, the genotype data appears not to be able to discriminate the groups.

3.3. Analysis of clustering patterns inside populations

The structure of the population cohorts was examined for the existence of clusters of individuals based on the glycan profiles and phenotypes. Affinity propagation algorithm was used to perform the clustering analysis with plasma and IgG glycan profiles and phenotypes taken separately as predictor variables. The clustered data was visually represented in the form of a heatmap expressing the feature levels of samples arranged by cluster. This data representation was intented to facilitate the comparison of clusters and, consequently, the identification of cluster-specific characteristics.

With the purpose of exploring the internal structure of the populations, affinity propagation clustering was applied to each population individually. The number of clusters obtained varied between 80 and 100 when input preferences were set to the median of the similarity matrix and between 3 and 8 when set to the minimum of the similarity matrix for all feature data sets in the three populations. Visual inspection of the similarity matrices used as input in the affinity propagation clustering showed that the small number of large clusters better reproduced the data. In general, the data patterns displayed by these small cluster structures were similar between populations for all feature data sets (Supplementary figure 2 presents the results for the Vis

which are illustrative of the results obtained for Korčula and Orkney). In the clusters obtained with plasma glycans, GP7, GP9, DG5, DG6, Monosialo and some of the structural glycans appeared as the most distinct features among clusters. The cluster division based on IgG profiles showed main differences on the level of IGG3, IGG13, IGG43 and IGG55 and also in some of the Charged and Neutral Derived features. The cluster structure obtained for phenotypes basically divided the individuals according to the levels of BMI and waist-hip-related features.

In order to verify whether the affinity propagation algorithm would be able to separate the three populations into three clusters, the pooled data of all populations was considered for clustering. For all feature data sets, the pooled data was divided into several clusters containing a small number of samples. The fact that the algorithm failed to discriminate the populations was not surprising in the light of the results previously obtained for each population separately which showed cluster similarities across populations. Visual inspection of the corresponding similarity matrices revealed that a number of clusters between 2 and 4 would better fit the data. Thus, affinity propagation was run with the number of clusters set beforehand to 2, 3 and 4 for each feature data set. The clusters obtained were formed of approximately the same number of individuals from each population and revealed similar tendencies to those observed in the individual populations. While the most appropriate structure for phenotypes was composed of 3 clusters (Figure 12), for plasma and IgG glycans the most correct division of samples was difficult to establish (Figure 13, Supplementary figure 3). For instance, in the case of plasma glycans, the division into 2 clusters showed opposite levels of several glycan peaks such as GP9, Monosialo, BAMS, BADS and C.FUC (Figure 13A), while the division into 3 clusters besides these differences also revealed an emerging cluster with high levels of GP7 and G2 (Figure 13B). The specific data patterns presented by the two cluster structures can be concurrently acceptable and equally valid to describe the data.

Additionally, for each clustering experiment, the heatmaps of the other two feature data sets were also displayed so as to verify the existence of associations between the three feature data sets. In none of the cases did the other two feature data sets present a cluster specific pattern meaning that the division into clusters depends on the type of feature.

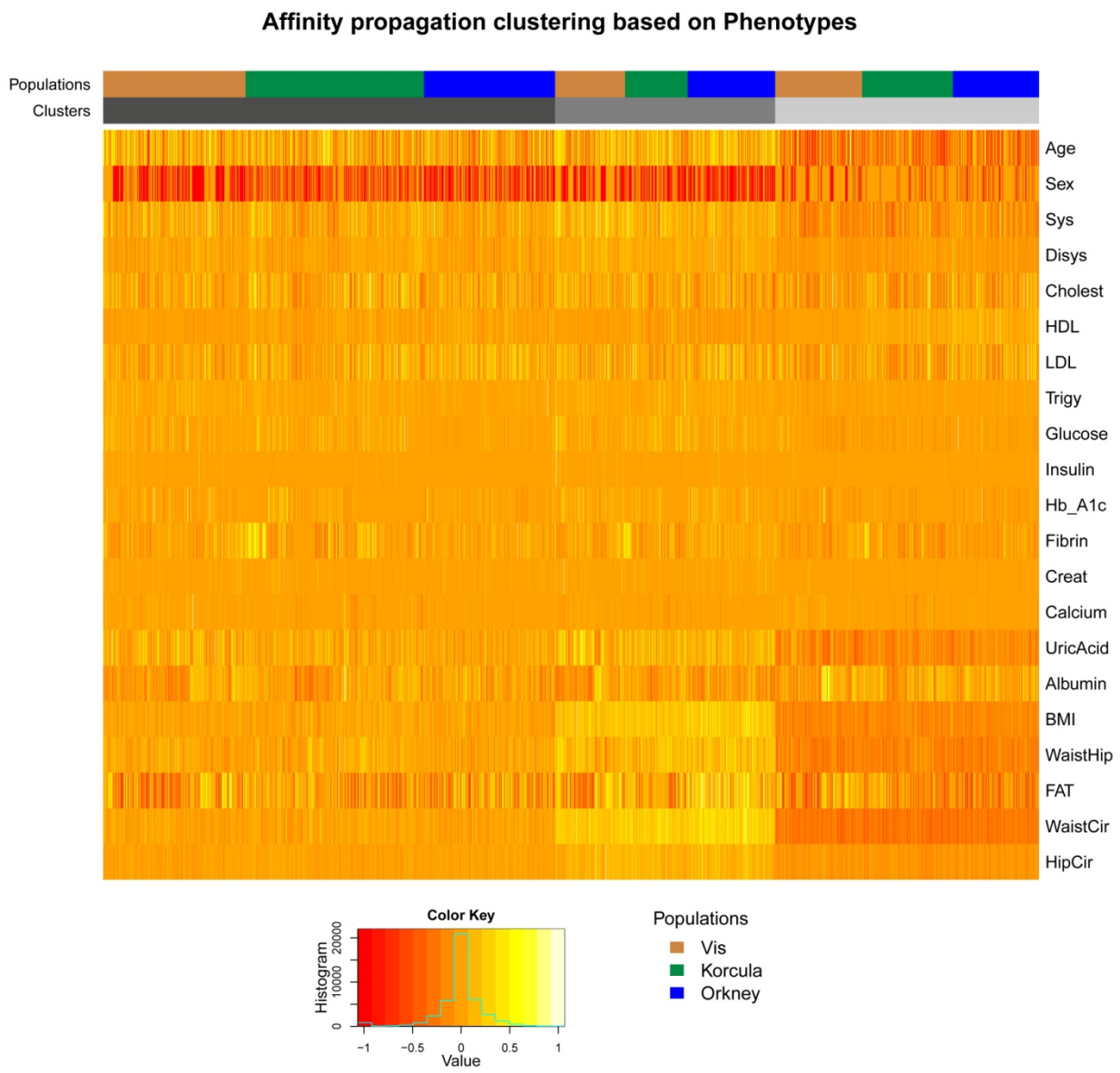


Figure 12. Affinity propagation clustering results based on phenotype data for the pooled data of populations. The structural division into 3 clusters was the one that best fitted the phenotype data. The clusters were comprised of approximately the same number of individuals from each population and presented differences at the level of uric acid, BMI, waist circumference and hip circumference. The heatmap represents the levels of each phenotypic feature (rows) for the samples in each cluster (columns); the key colour of the heatmap varies from red to yellow corresponding to low and high values, respectively. The bars above the heatmap depict the cluster division in different shades of grey and the population division coloured as gold for Vis, green for Korčula and blue for Orkney.

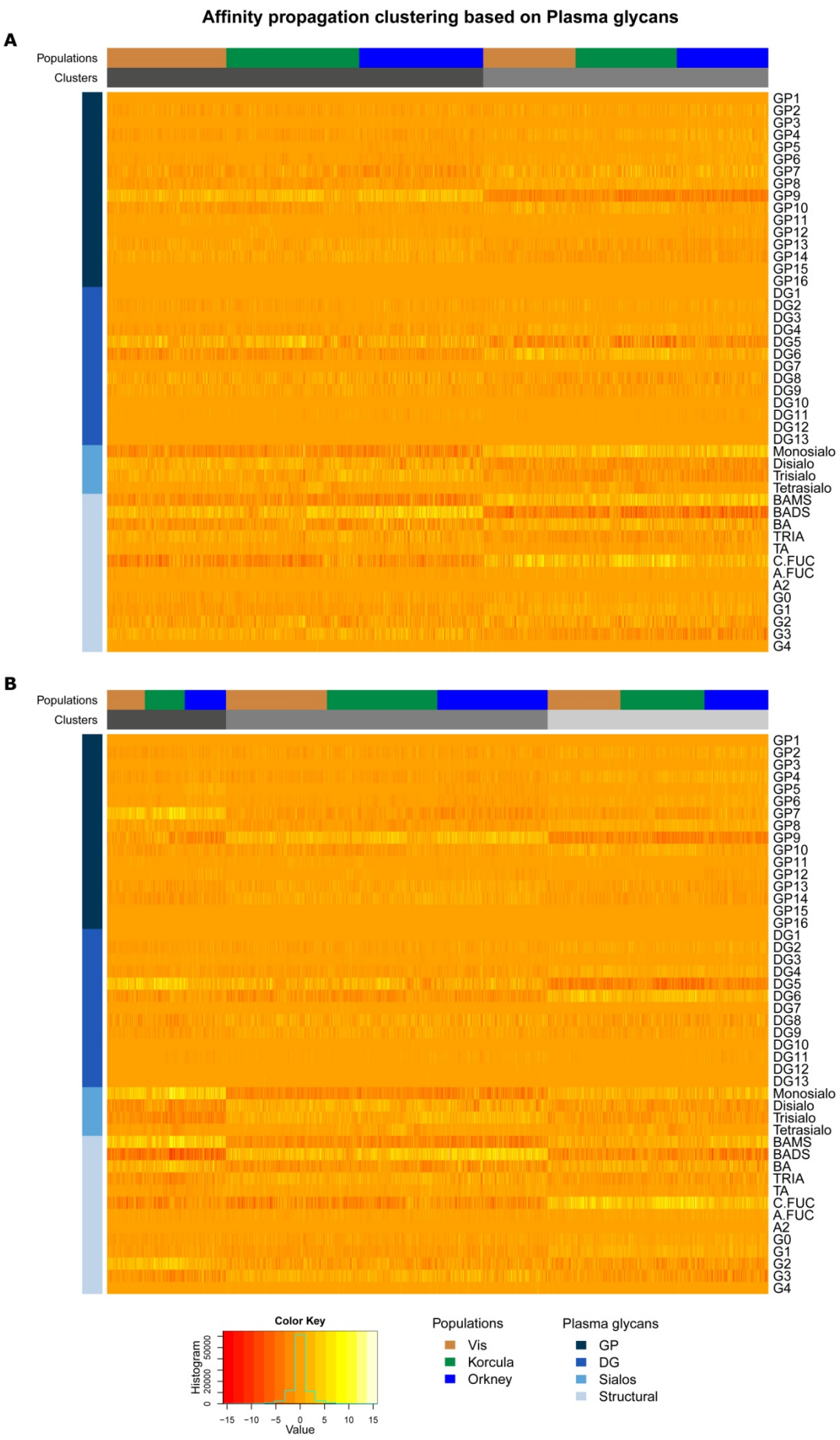

Affinity propagation clustering based on Plasma glycans
A
Populations
Clusters
GP1
GP2
GP3
GP4
GP5
GP6
GP7
GP8
GP9
GP10
GP11
GP12
GP13
GP14
GP15
GP16
DG1
DG2
DG3
DG4
DG5
DG6
DG7
DG8
DG9
DG10
DG11
DG12
DG13
Monosialo
Disialo
Trisialo
Tetrasialo
BAMS
BADS
BA
TRIA
TA
C.FUC
A.FUC
A2
G0
G1
G2
G3
G4
B
Populations
Clusters
GP1
GP2
GP3
GP4
GP5
GP6
GP7
GP8
GP9
GP10
GP11
GP12
GP13
GP14
GP15
GP16
DG1
DG2
DG3
DG4
DG5
DG6
DG7
DG8
DG9
DG10
DG11
DG12
DG13
Monosialo
Disialo
Trisialo
Tetrasialo
BAMS
BADS
BA
TRIA
TA
C.FUC
A.FUC
A2
G0
G1
G2
G3
G4
Color Key
Histogram
0
20000
50000
−15
−10
−5
0
5
10
15
Value
Populations
Vis
Korcula
Orkney
Plasma glycans
GP
DG
Sialos
Structural

Figure 13. Affinity propagation clustering results based on plasma glycan profiles for the pooled data of populations. The clustering results of affinity propagation algorithm run with K=2 (A) and K=3 (B) are presented to illustrate the difficulty in establishing the most reliable clustering structure. In the case of the 2 cluster division, opposite levels of glycan features such as GP9, Monosialo, BAMS, BADS and C.FUC are clearly observed. These main differences are retained for a part of the samples in the 3 cluster division which additionally reveals a cluster with high levels of GP7 and G2. The heatmap represents the levels of each glycan (rows) for the samples in each cluster (columns); the key colour of the heatmap varies from red to yellow corresponding to low and high values, respectively. The bars above the heatmap depict the cluster division in different shades of grey and the population division coloured as gold for Vis, green for Korčula and blue for Orkney. The bar on the left side of the heatmap indicates the four groups of plasma glycans: GP (dark blue), DG (blue), Sialos (medium blue) and Structural (light blue).

3.4. Correlation between N-glycome and phenotypic traits

The correlation between the plasma and IgG glycan profiles and the set of available phenotypes was carried out to identify environmental determinants that are likely to affect glycans. The analysis aimed to find the extent to which the correlations can be replicated across all population cohorts, in order to be able to identify general patterns of association and to provide evidence of possible population-specific correlations that could be related to the geographical and lifestyle separation of the populations.

The plasma and IgG data of the analysed populations presented patterns of correlation with age which have been previously described such as the decrease of core-fucosylation, galactosylation and sialylation and the increase of structures with bisecting GlcNAc (Huhn *et al.*, 2009). Additionally, the effect of age on both glycan sets in the populations of Korčula and Orkney replicated the findings reported for the Vis population confirming the age-dependency of certain glycans structures (Supplementary figure 4 and Supplementary figure 5)(Knezevic *et al.*, 2009; Pucic *et al.*, 2011). To remove the effects of aging and gender upon the associations between glycans and phenotypes, plasma and IgG data sets were subjected to age and sex correction, as described in section 2.4.4. All subsequent analyses were performed on the corrected data.

The correlation coefficients between glycans and phenotypes were higher for plasma glycans ranging from approximately -0.3 to 0.3 than for IgG glycans varying from approximately -0.17 to 0.17. Overall, the tendency of the glycan-phenotype associations was similar across populations for both plasma and IgG glycans although with slightly different magnitudes (Supplementary figure 6 and Supplementary figure 7).

In plasma glycans, statistically significant correlations present in all three populations were mainly found for body fat parameters and lipid-related measures ($p<0.000362$; Figure 14). DG10 and BADS peaks were positively correlated with BMI, waist circumference and hip circumference, while GP5, GP8 and BAMS were negatively correlated with these same phenotypes. DG10 was also positively correlated with cholesterol and LDL and DG8 with triglycerides. Particular correlations were observed for Vis between tretraantennary structures (TRIA and G3) and cholesterol and for Korčula between GP0, GP14, G1 and G3 and insulin.

In IgG glycans, despite the fact that the majority of strong correlations were consistent in all populations, statistically significant correlations were sparse and mainly shown for Orkney ($p<0.000216$; Figure 15). An interesting association pattern which did not pass the threshold of significance is the one displayed by the population of Orkney between the glycan structures of the IgG Neutral derived group and calcium (Supplementary figure 7). Positive correlations were observed for IgG glycan features containing bisecting N-acetylglucosamine (GlcNAc) whereas negative correlations were found for structures without bisecting GlcNAc. Bisecting GlcNAc structures are synthesized as a result of a transfer of a GlcNAc residue to the mannose residue at the base of the core of the N-glycan and are known to have important effects on the IgG protein function.

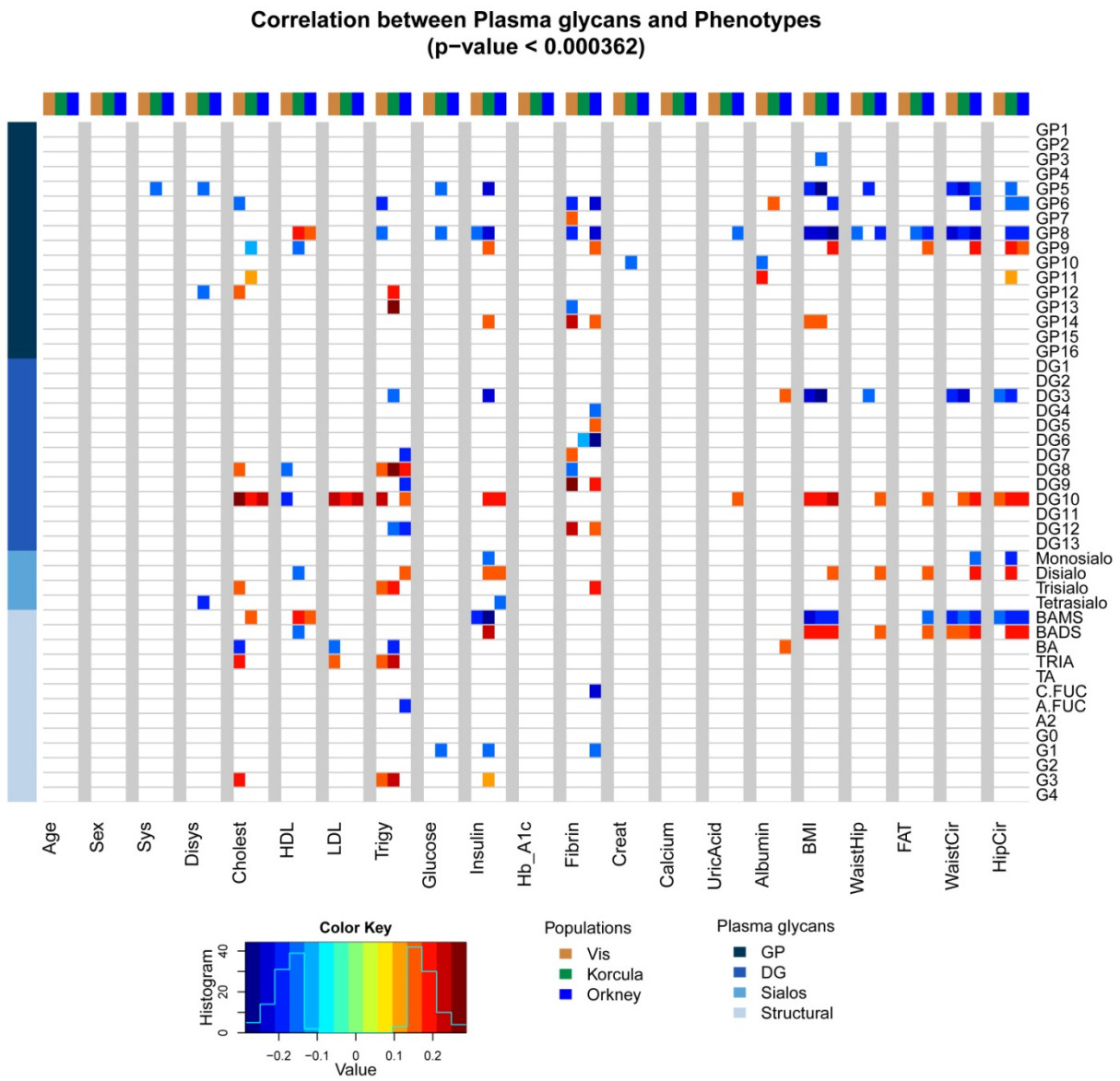


Figure 14. Statistically significant correlations between plasma glycans and phenotypes for all populations. The heatmap depicts the level of correlation between each plasma glycan feature (rows) and the phenotypes for each population (columns); correlation coefficients range from -0.3 (dark blue) to 0.3 (dark red). The bar above the heatmap indicates the population to which the three columns of each phenotype correspond to: gold for Vis, green for Korčula and blue for Orkney. The bar on the left side of the heatmap indicates the four groups of plasma glycans: GP (dark blue), DG (blue), Sialos (medium blue) and Structural (light blue). The significance level was set to 0.000362 to account for the multiple testing (46 plasma features and 3 populations).

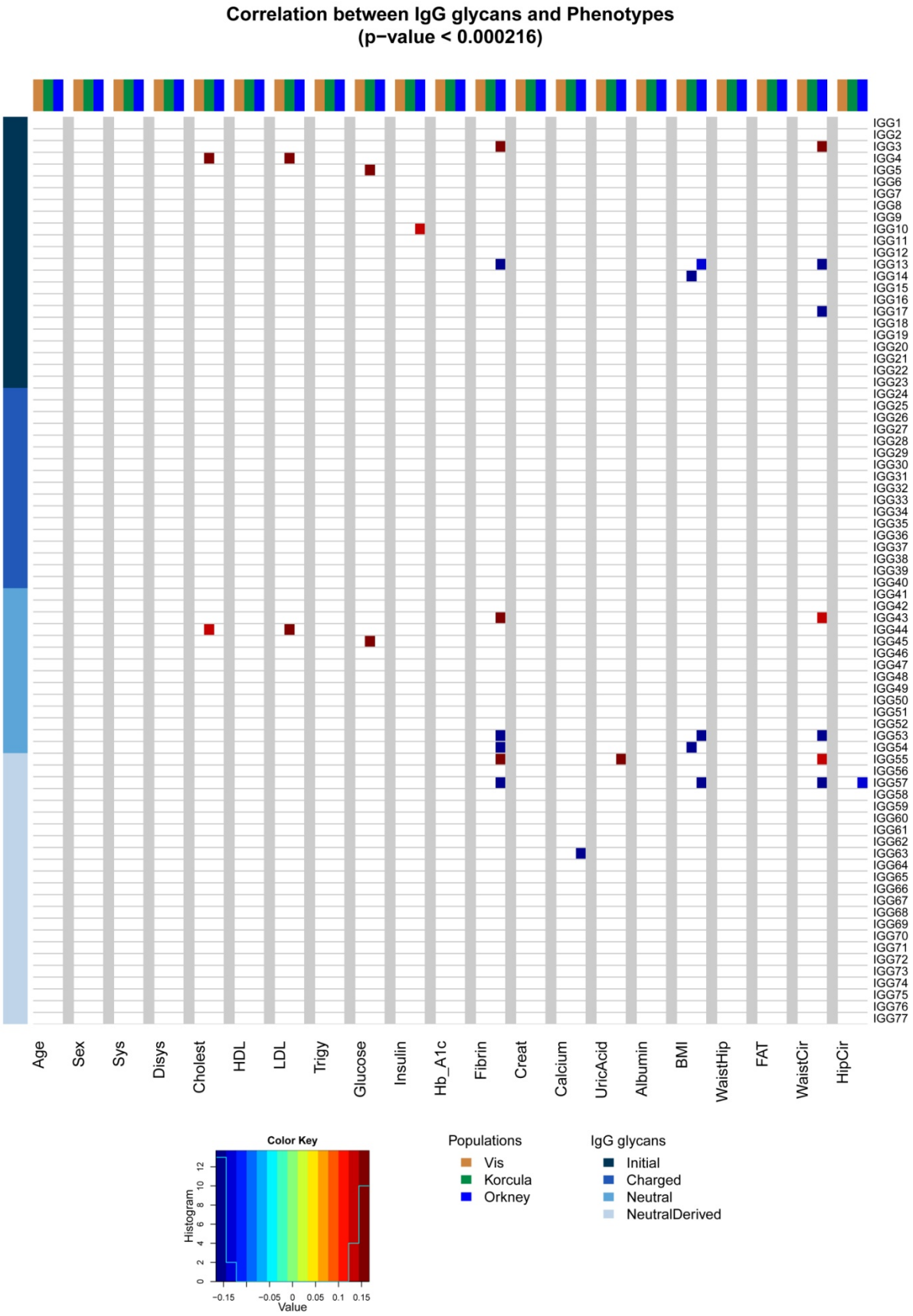

Correlation between IgG glycans and Phenotypes
(p−value < 0.000216)
IGG1
IGG2
IGG3
IGG4
IGG5
IGG6
IGG7
IGG8
IGG9
IGG10
IGG11
IGG12
IGG13
IGG14
IGG15
IGG16
IGG17
IGG18
IGG19
IGG20
IGG21
IGG22
IGG23
IGG24
IGG25
IGG26
IGG27
IGG28
IGG29
IGG30
IGG31
IGG32
IGG33
IGG34
IGG35
IGG36
IGG37
IGG38
IGG39
IGG40
IGG41
IGG42
IGG43
IGG44
IGG45
IGG46
IGG47
IGG48
IGG49
IGG50
IGG51
IGG52
IGG53
IGG54
IGG55
IGG56
IGG57
IGG58
IGG59
IGG60
IGG61
IGG62
IGG63
IGG64
IGG65
IGG66
IGG67
IGG68
IGG69
IGG70
IGG71
IGG72
IGG73
IGG74
IGG75
IGG76
IGG77
Age
Sex
Sys
Disys
Cholest
HDL
LDL
Trigy
Glucose
Insulin
Hb_A1c
Fibrin
Creat
Calcium
UricAcid
Albumin
BMI
WaistHip
FAT
WaistCir
HipCir
Color Key
Histogram
0
2
4
6
8
10
12
−0.15
−0.05
0
0.05
0.1
0.15
Value
Populations
Vis
Korcula
Orkney
IgG glycans
Initial
Charged
Neutral
NeutralDerived

Figure 15. Statistically significant correlations between IgG glycans and phenotypes for all populations. The heatmap depicts the level of correlation between each IgG glycan feature (rows) and the phenotypes for each population (columns); correlation coefficients range from -0.17 (dark blue) to 0.17 (dark red). The bar above the heatmap indicates the population to which the three columns of each phenotype correspond to: gold for Vis, green for Korčula and blue for Orkney. The bar on the left side of the heatmap indicates the four groups of IgG glycans: Initial (dark blue), Charged (blue), Neutral (medium blue) and Neutral derived (light blue). The significance level was set to 0.000216 to account for the multiple testing (77 IgG features and 3 populations).

While the ensemble of plasma glycans analysed contains N-glycans attached to a variety of proteins, the IgG glycans are a filtered subset containing only N-glycans attached to the IgG protein. This fact allows establishing a correspondence between the main IgG peaks and certain plasma peaks containing the same N-glycan structures (Supplementary table 5). The existence of such correspondence was considered to determine whether the associations with phenotypes found in plasma glycans could be captured by the corresponding IgG glycans component.

The IgG Initial group peaks (GP1-GP24) were combined into 11 plasma peaks (GP1-GP11) for the pooled data of all populations. The correlation of these IgG combined peaks with phenotypes was computed and compared to the original correlation pattern of plasma peaks (Figure 16). Plasma and IgG peaks presented a quite similar pattern of correlation with the strongest and most stable correlations being between GP5, GP6 and GP8 peaks and both body fat parameters and lipid-related measures. Opposite correlation tendencies were found for GP9 and waist and hip circumferences and for GP11 and albumin. In both cases, evident positive correlation was observed for plasma data and almost non-existing negative correlation was found for the corresponding IgG data. The same analysis preformed on the individual populations produced comparable results to those obtained and described above for the pooled data. The agreement shown between plasma and IgG peaks can be viewed as a reinforcement of the existence of associations between certain glycan structures and phenotypes.

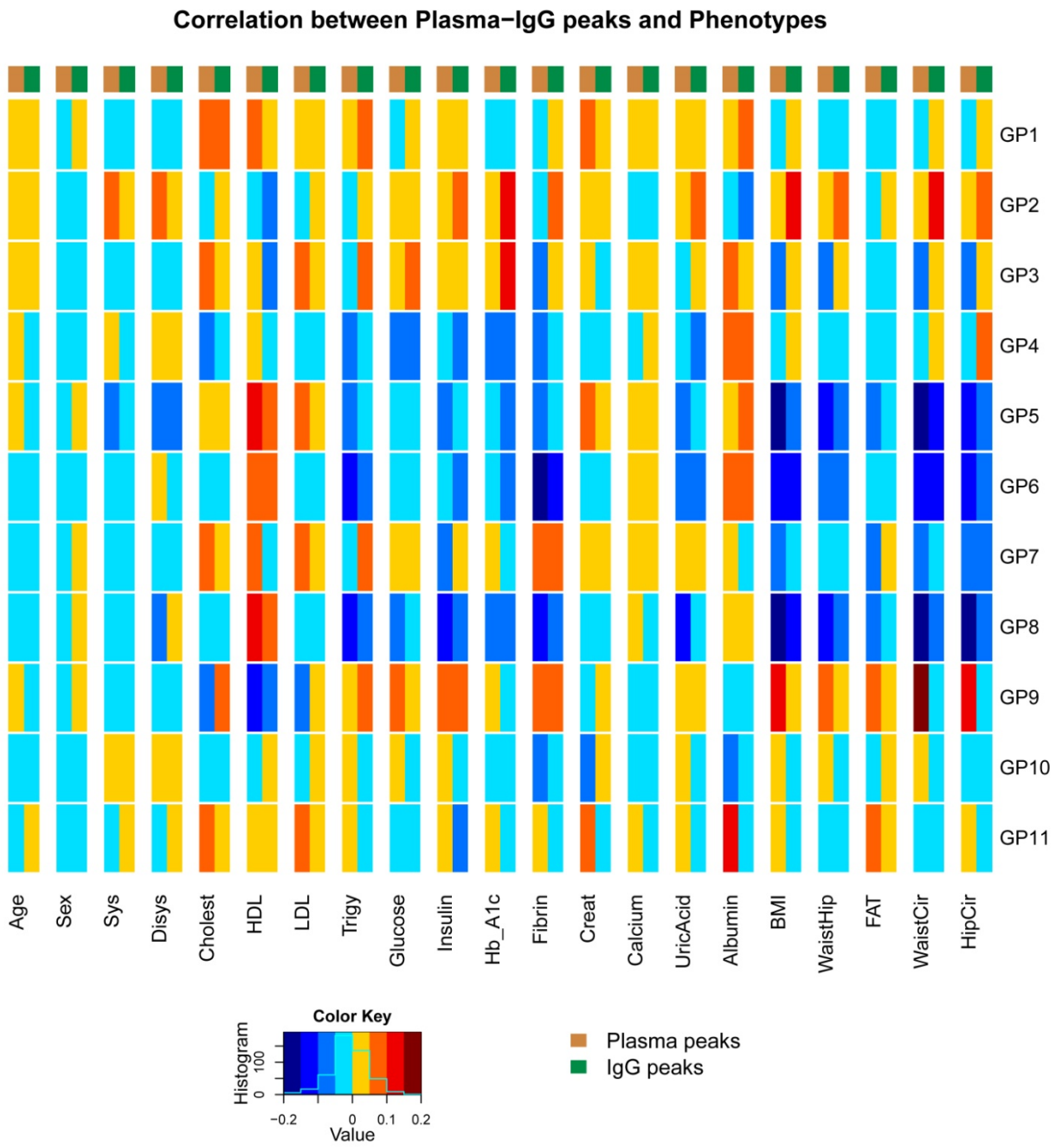


Figure 16. Correlation of 11 plasma peaks and their corresponding IgG peaks with phenotypes. The heatmap depicts the level of correlation between each of the 11 glycan peaks (rows) and the phenotypes (columns) for the pooled data of all populations; correlation coefficients range from -0.2 (dark blue) to 0.2 (dark red). Each phenotype comprises two columns corresponding to plasma (gold) and IgG (green) peaks as indicated by the bar above the heatmap.

3.5. Comparison of feature profiles from diabetes groups

The potential of glycans as a biomarker for diseases was evaluated for the particular case of diabetes. For the purpose, several computational methods were applied to a total of 1577 samples divided into 3 groups (non-diabetic, pre-diabetic and diabetic) in an attempt to detect possible glycan, phenotypic and genotypic features characteristic of each group status.

Parallel coordinates plot were used to display the glycan and phenotype profiles of the three groups in a visually clear manner and, in this way, facilitate the comparison of features across groups (Figure 17). The plasma and IgG profiles did not reveal any particular features that were clearly distinct across the three groups. On the other hand, the phenotype profiles showed marked differences at the level of HbA1c and glucose with the diabetic group having higher levels and the non-diabetic having lower levels. Less pronounced differences were found for age, systolic blood pressure, BMI, waist-to-hip ratio and waist circumference.

The nonparametric Wilcoxon sum rank statistical test was employed to further assess pairwise differences in levels of glycans and phenotypes between the three groups. The nonexistence of differences between groups in the case of plasma and IgG glycans was confirmed by the absence of statistically significant hits in these two feature data sets ($p<0.00109$ for plasma; $p<0.00065$ for IgG). Regarding the phenotypes, the majority of statistically significant differences were obtained for the pairwise comparisons of the non-diabetic group with the other two groups ($p<0.00238$; Figure 18). Differences in systolic blood pressure, glucose and HbA1c were considered statistically significant across the three groups.

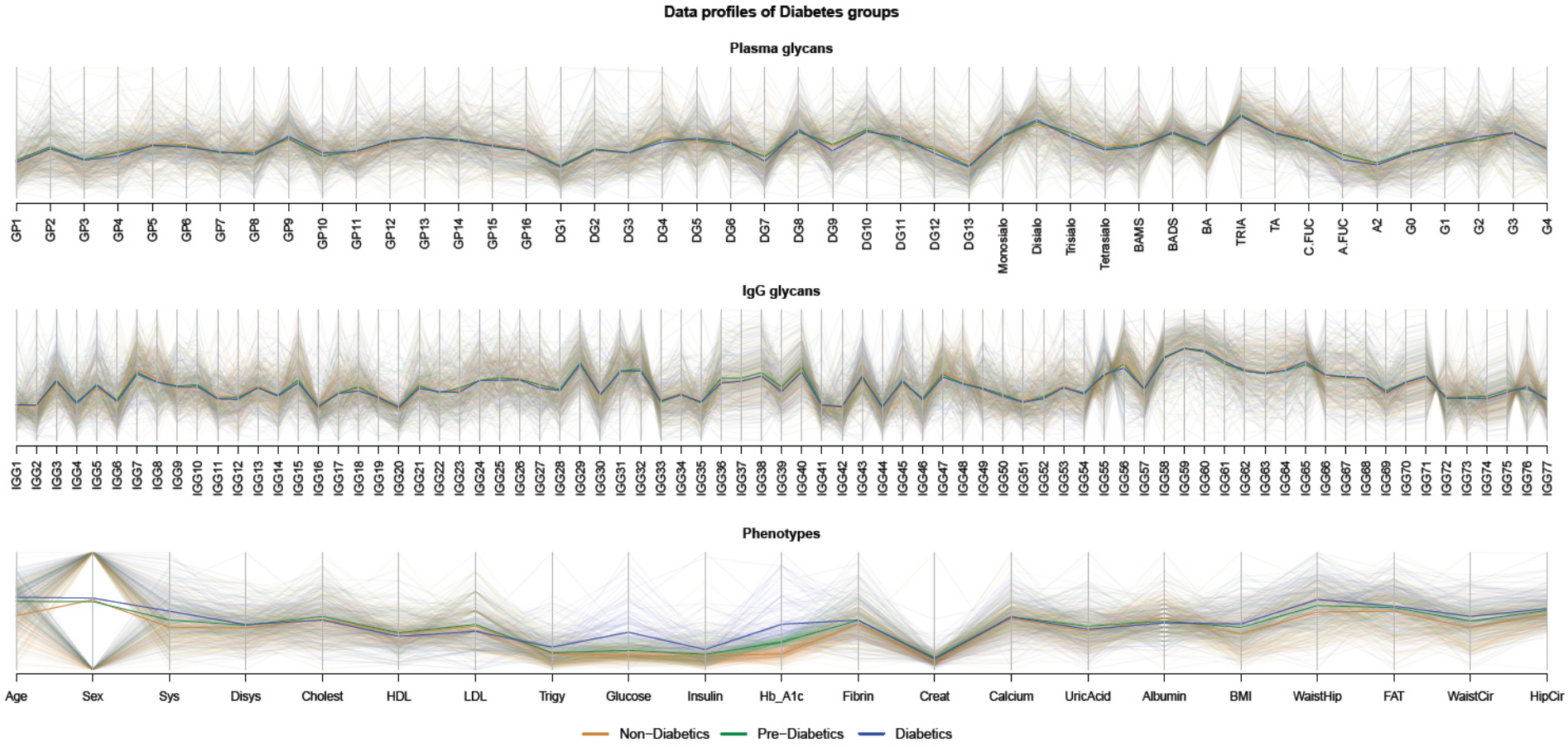


Figure 17. Parallel coordinates plots of plasma glycan, IgG glycan and phenotype profiles for non-diabetic, pre-diabetic and diabetic groups. The plasma and IgG profiles do not show any differences between groups while the phenotype profiles show, among others, high values of HbA1c and glucose for the diabetic group and low values for the non-diabetic group. The median values of the features for each group are highlighted. Sex is represented as 0 (males) and 1 (females). Non-diabetic group samples are represented by gold lines, pre-diabetic by green lines and diabetic by blue lines.

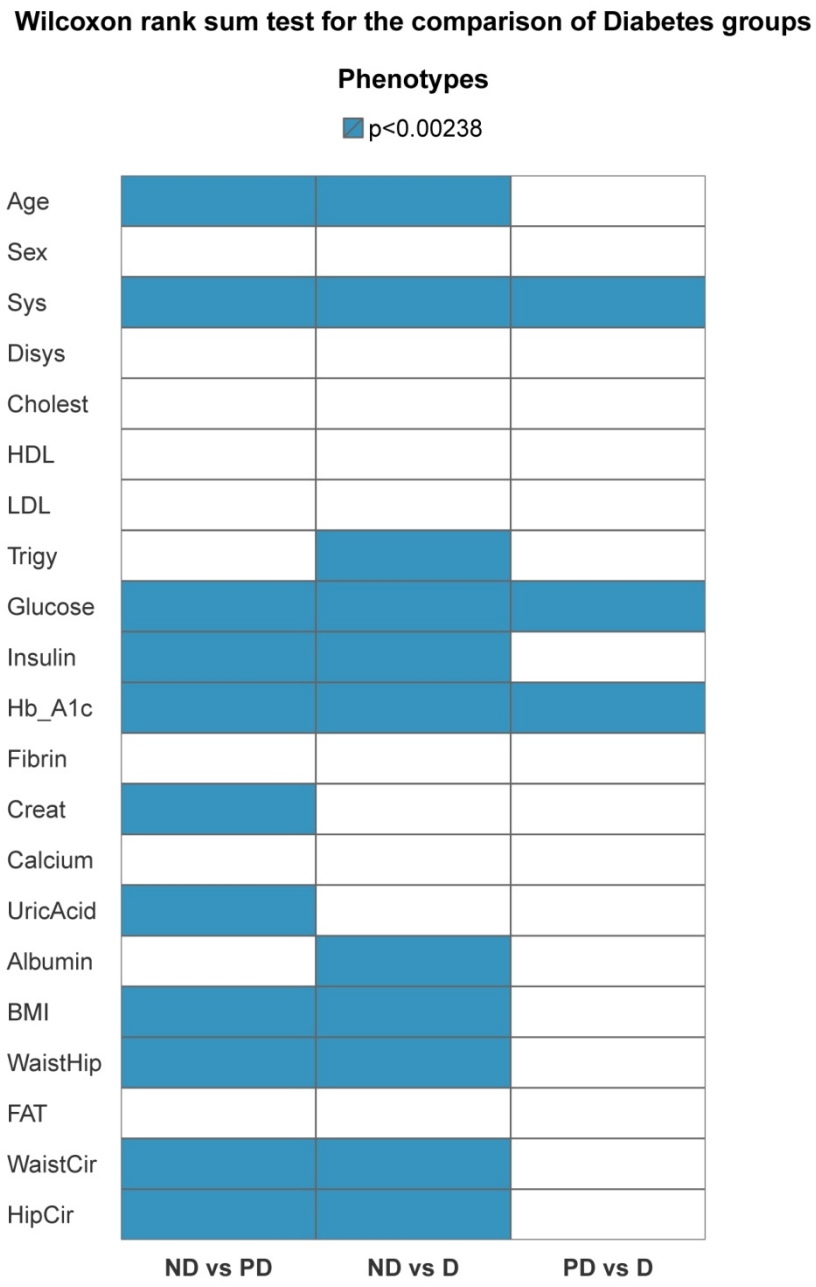


Figure 18. Wilcoxon sum rank test p-values for the pairwise comparison of the diabetes groups with regard to phenotype data. Statistically significant hits are highlighted in blue; the significance level was set to 0.00238 to account for multiple testing in each pairwise comparison. The names of the groups are abbreviated as ND for non-diabetics, PD for pre-diabetics and D for diabetics.

Following the initial overview of the data, PLS-DA and PCA methods were used to try to separate the groups based on the glycan profiles and the phenotypes. The analyses were carried out for the entire group of individuals and for the individuals divided by populations with similar outcomes produced. The analysis of plasma and IgG glycan profiles using both methods was not

able to differentiate the groups whether full glycan profiles or individual glycan groups were considered as could be anticipated from the data visualisation (Supplementary figure 8 and Supplementary figure 9). PLS-DA applied to the phenotype data set yielded a separation of the three groups with HbA1c and glucose emerging as the variables most contributing for the division as expected from the role they play in diabetes (Figure 19, left panel). PCA in its turn was not able to capture these phenotypic differences (Figure 19, right panel). Instead, PCA detected general phenotypic patterns already observed in the clustering analysis evidenced by the fact that the 3 features identified as the most contributing for PCA (uric acid, waist and hip circumferences) were among the ones being more distinct between clusters. Additionally, the top 10 contributing features from each data set were selected and jointly used as a new input for both PLS-DA and PCA. The results produced were identical to the ones reported for the phenotype data set due to the stronger influence of phenotype features over glycan ones.

Despite the unsatisfactory results obtained with PCA and PLS-DA analyses, the Random Forest (RF) method was employed with the purpose of obtaining quantification measures regarding the prediction process of the groups. The classification using either plasma or IgG glycan profiles as predictor variables (considering all peaks simultaneously or each group of peaks individually) was poor and almost all individuals were placed into the non-diabetic group (Table 4A and B). The classification using phenotypes as predictor variables had a much better performance assigning the majority of the testing samples to the correct group with an estimated error rate of 4.5% for a 10-fold cross-validation (Figure 20, Table 4C). The classification of groups based on the combination of the 10 most important variables from each feature data set (30 features in total) had a comparable performance to the classification using phenotypes similarly to what occurred in the PLS-DA and PCA analyses (Figure 20, Table 4D).

The results obtained with PLS-DA, PCA and RF analyses suggest that plasma and IgG glycans might not have enough predictive power for this data set while confirming the fact that HbA1c and glucose are good indicators of the diabetes status.

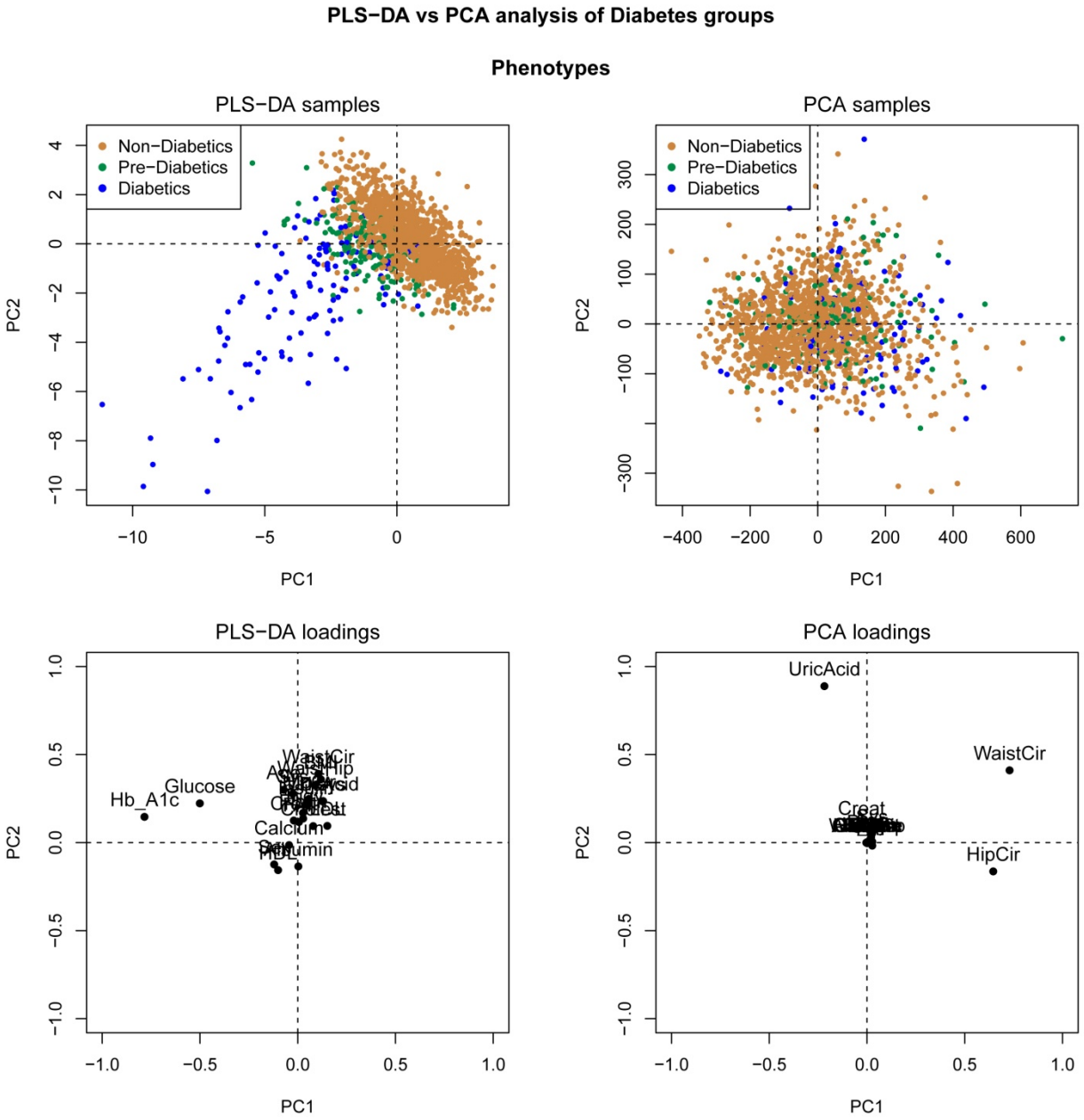


Figure 19. PLS-DA and PCA analysis of the diabetes groups using phenotype data. Although PLS-DA achieved a separation of the groups and HbA1c and glucose were indicated as the most important factors for the separation, PCA was not able to distinguish the groups. The score plots representing the data samples by the two first principal components (PC1 on the x-axis and PC2 on the y-axis) are shown on the upper panels; groups are coloured as gold for non-diabetic, green for pre-diabetic and blue for diabetic. The corresponding loading plots establishing the relative contributions of each phenotype feature to the overall variation in the groups are shown on the lower panels.

Table 4. Random Forest confusion matrices for the classification of the diabetes groups. The confusion matrices show the performance of the Random Forest algorithm in classifying testing samples using all plasma glycans (A), all IgG glycans (B), phenotypes (C) and the set of 30 most important features of all data sets (D). The plasma and IgG glycans have a poor classification performance with almost all testing samples assigned to the non-diabetic group. The phenotypes and the set of 30 most important features show similar performance with the non-diabetic group having the lower classification error per group. Each row of the matrix represents the instances in the actual group, while each column represents the instances in a predicted class (0 for non-diabetics, 1 for pre-diabetics and 2 for diabetics). The "Error" column indicates the test set errors for the classification of each group and for the overall classification of the testing samples.

A) Plasma glycans (all)

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	391	0	0	0
1	41	0	0	100
2	41	0	0	100
				17.34

B) IgG glycans (all)

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	393	0	1	0.3
1	42	0	0	100
2	37	0	0	100
				24.12

C) Phenotypes

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	394	0	0	0
1	9	35	4	27.1
2	5	0	26	16.1
				3.81

D) Set of 30 most important features

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	397	0	1	0.2
1	11	29	1	29.3
2	4	4	26	23.5
				4.44

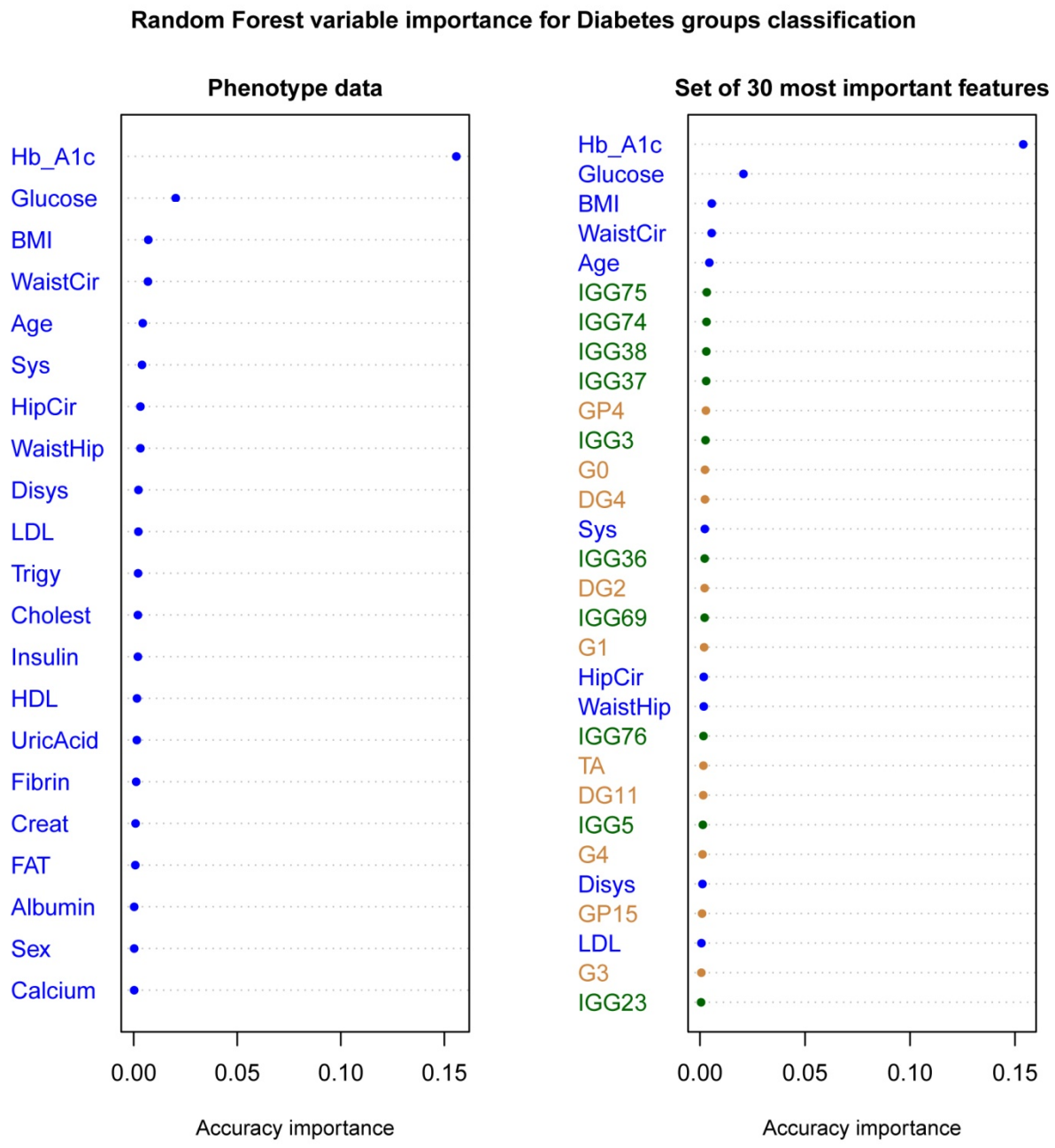


Figure 20. Random Forest variable importance for the classification of the diabetes groups. The variable importance plots correspond to the phenotype data (left panel) and to the set of 30 most important features of all data sets (right panel). The measure of variable importance presented is the mean decrease in accuracy estimated by comparing the accuracy in classification without and with permutation of the values of each predictor variable. When a given variable has little predictive power, its permutation will not cause substantial difference in accuracies, therefore a higher decrease in accuracy is indicative of a more important variable. Variable labels are coloured as gold for plasma glycans, green for IgG glycans and blue for phenotypes.

The possibility of classifying the samples according to the three groups based on the genotypic information was analysed by employing the Random Jungle (RJ) algorithm. RJ was not able to correctly classify the samples with the majority of them being assigned to the non-diabetics group both when using the all SNPs and glycan-related SNPs sets. Since the non-diabetic group was circa 5 times larger than the other two groups, RJ was performed several times with random subsets of 200 samples from the non-diabetics group to verify whether or not the unequal size of the groups could be affecting the results. The RJ performance did not improve when using the subsets of samples (error rate between 55-60%) indicating that the number of samples is not a critical factor for the classification.

The genetic contribution to disease and the possibility of prioritizing biomarkers in the case of diabetes was also explored using the correlation adjusted scores method. Among the top SNPs emerging as important with the correlation adjusted scores approach are several SNPs located in genes or regions around genes which have been directly related to diabetes or linked to the regulation of insulin secretion and retinal degeneration such as NCS1, CDK19, DCLK1, GLCCI1, CCR and DCC. The top 30 SNPs potentially associated with diabetes condition are listed in Supplementary table 6 and the genetic context of some of these SNPs is represented in Figure 21 and in Supplementary figure 10.

Genetic context of polymorphisms possibly associated with Diabetes

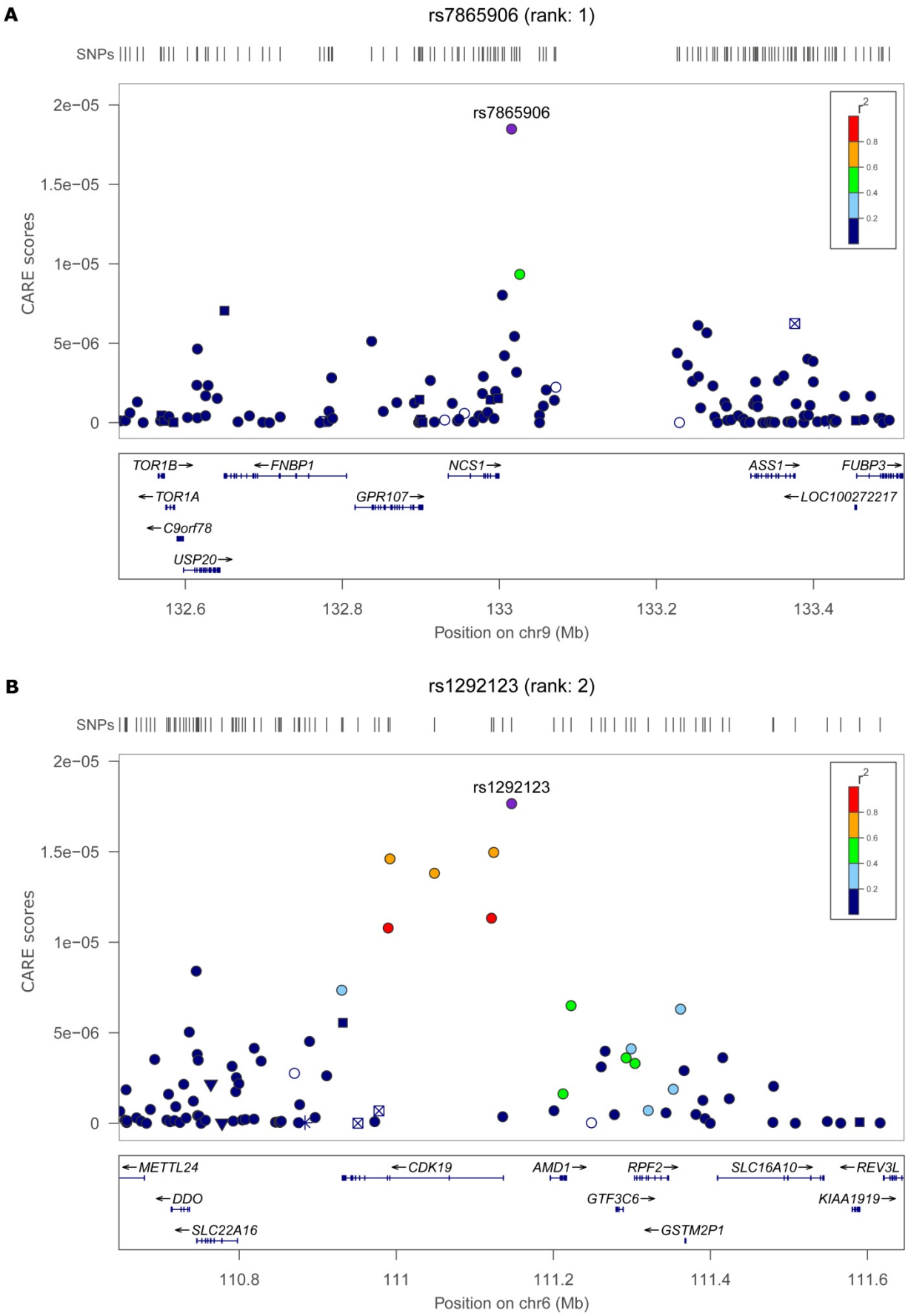


Figure 21. Genetic context of polymorphisms possibly associated with the diabetes condition. The regional association plots show the correlation adjusted scores (CARE scores) for SNPs distributed in a genomic region centred on variants rs1292123 on chromosome 6 (A) and rs6563348 on chromosome 13 (B). The flanking region extends 0.5Mb both upstream and downstream of the reference SNP which is labelled and shown in purple. The colour intensity of the other SNPs within the region represents the extent of their linkage disequilibrium (r2) with the reference SNP: red (r2 ≥ 0.8), orange (0.6 ≤ 0.8), green (0.4 ≤ 0.6), light blue (0.2 ≤ 0.4) and dark blue (r2 ≤ 0.2). The locations of known genes in the region are depicted below the association plot.

3.6. Comparison of feature profiles from isolated populations

Since the population cohorts studied represent isolated populations coming from different geographic regions, they are likely to present their own characteristics. In order to seek for population-specific patterns capable of differentiating between populations, the glycan and phenotype profiles were compared and the genotyped data analysed using diverse computational tools and algorithms.

Parallel coordinates plots were used to display the glycan and phenotype profiles of the three populations in concise yet descriptive manner (Figure 22). The plasma profiles do not show visible differences between groups, the IgG profiles show slight differences for several of the glycans and the phenotype profiles show differences for sex.

The statistical analysis was performed using the nonparametric Wilcoxon sum rank test which assessed the statistical differences in the levels of glycans and phenotypes for each pairwise comparison between populations (Figure 23). In the case of plasma glycans, differences in BADS between Vis and Orkney and in GP6 between Korčula and Orkney were considered significant ($p<0.00109$). Regarding the IgG glycans, several features emerged as significant in the pairwise comparisons of Orkney with either Vis or Korčula, while none of the features reached the threshold of significance in the pairwise comparison between Vis and Korčula ($p<0.00065$). As for the phenotypes, statistically significant differences were obtained for gender between Korčula and Orkney and for LDL in the pairwise comparisons of Orkney with Vis and Korčula ($p<0.00238$).

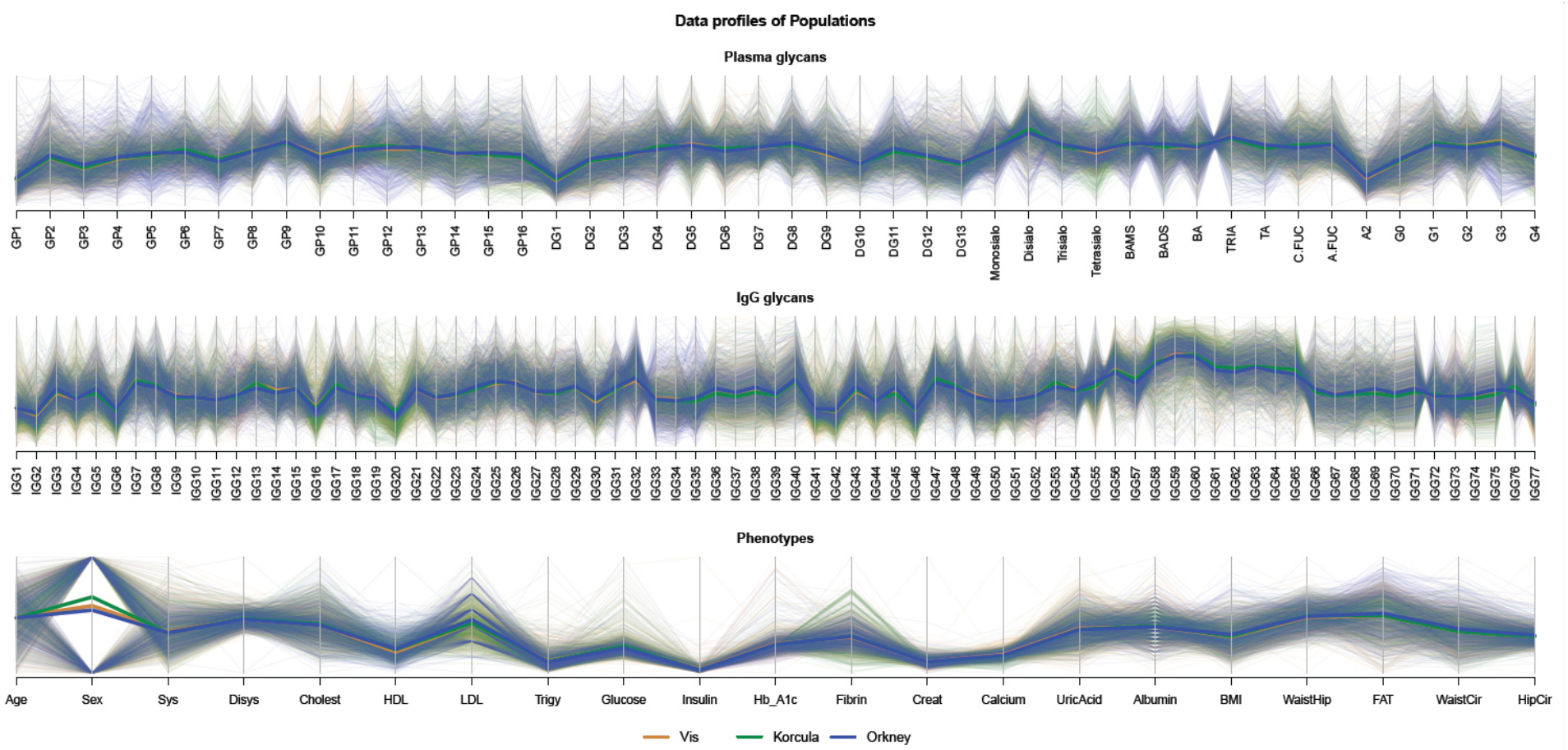


Figure 22. Parallel coordinates plots of plasma glycan, IgG glycan and phenotype profiles for Vis, Korčula and Orkney populations. The plasma profiles do not show visible differences between groups, the IgG profiles show slight differences for several of the glycans and the phenotype profiles show differences for gender ('Sex'). The median values of the features for each group are highlighted. Sex is represented as 0 (males) and 1 (females). Vis samples are represented by gold lines, Korčula by green lines and Orkney by blue lines.

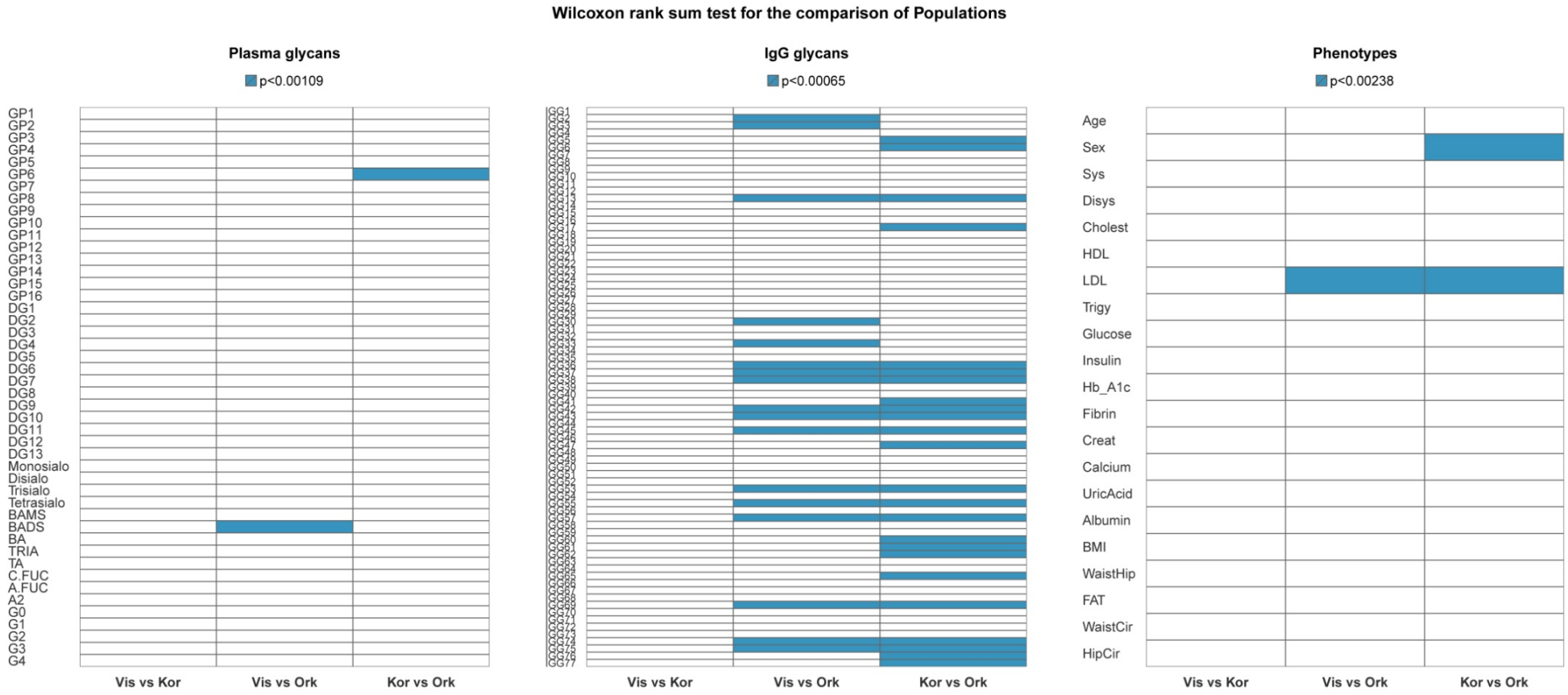


Figure 23. Wilcoxon sum rank test p-values for the pairwise comparison of the population cohorts with regard to plasma, IgG and phenotype data. Statistically significant hits are highlighted in blue; the significance level was set to 0.00109 for plasma glycans, 0.00065 for IgG glycans and 0.00238 for phenotypes to account for multiple testing in each pairwise comparison. The names of the groups are abbreviated as Vis for Vis, Kor for Korčula and Ork for Orkney.

PLS-DA and PCA methods were applied to each of the three feature profiles to investigate whether the patterns of dissimilarity established by the statistical analysis would be able to discriminate the populations. Surprisingly, none of the feature data sets, independently of the used method, yielded a separation of the populations (Figure 24 for plasma glycans, Supplementary figure 11 for IgG glycans and Supplementary figure 12 for phenotypes). The patterns of contribution of plasma, IgG and phenotype features to the overall data variation shown by the PCA analysis resemble the ones obtained when analysing the diabetes groups (Figure 24, lower right panel; see section 3.5 for comparison with diabetes data).

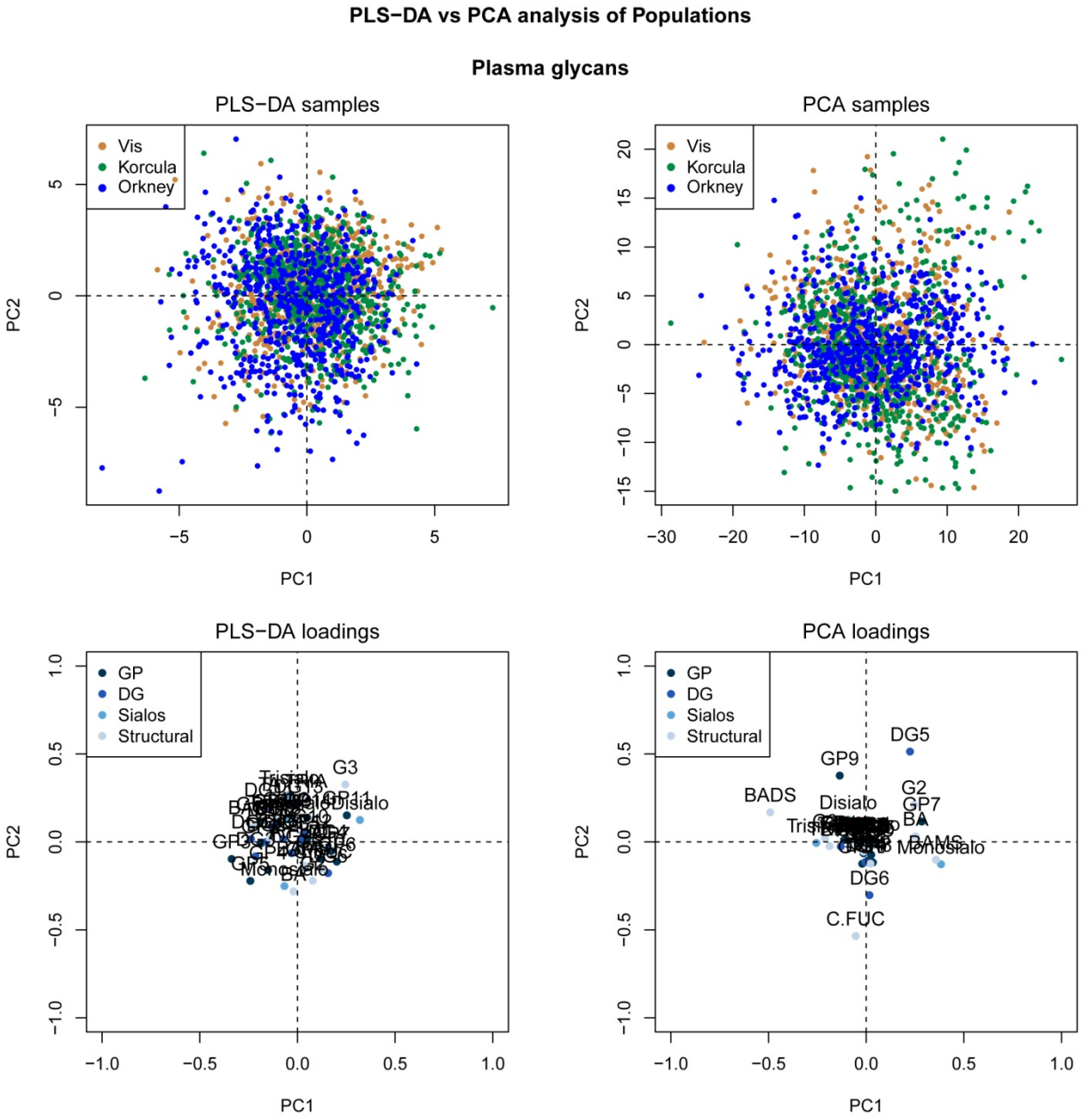


Figure 24. PLS-DA and PCA analysis of the population cohorts using plasma glycans data. None of the methods achieved a separation of the populations based on the profiles of all plasma glycans. The score plots

representing the data samples by the two first principal components (PC1 on the x-axis and PC2 on the y-axis) are shown on the upper panels; the populations are coloured as gold for Vis, green for Korčula and blue for Orkney. The corresponding loading plots establishing the relative contributions of each plasma feature to the overall variation in the populations are shown on the lower panels; glycans are coloured according to their group: GP (dark blue), DG (blue), Sialos (medium blue) and Structural (light blue).

Random Forest was conducted to investigate the performance of a classifier on the task of predicting populations based on glycan profiles (all peaks simultaneously and each group of peaks individually), phenotypes and a set of 30 features combining the top 10 contributing features from each data set. The best classification was achieved for the set of 30 mixed features and the worst for the IgG glycans. The 10-fold cross-validation estimated classification errors were 11%, 16%, 19% and 25% for the set of 30 features, phenotypes, plasma and IgG glycans, respectively (the confusion matrices for the classification of test set samples are shown in Table 5).

In the case of plasma glycans, the entire set of glycans produced the best results having an estimated classification error of 19% for a 10-fold cross-validation. The highest measures of variable importance belong to two GP glycans (Figure 25) explaining the fact that the GP group presented the lowest classification error rate among the individual groups of plasma glycans: a 27% error for GP compared to more than 40% for the other groups, estimated for a 10-fold cross-validation.

Regarding the classification using IgG glycans, it was also the entire set of glycans that achieved the best performance and the Initial group yielded the best classification among the individual groups of IgG glycans with estimated errors of 25% and 26% for a 10-fold cross-validation, respectively. The similarity in these two classification error values suggests that the Initial group of IgG glycans holds the most significant information for the overall classification. In fact, 5 glycans belonging to the Initial group were among the most important variables when the entire set of IgG glycans was used for classification (Figure 25).

The phenotype data had a similar classification performance as the plasma glycans.

Combining the top 10 most important features from each data set decreased the estimated error rate for a 10-fold cross-validation to 11%. This improvement in the overall classification indicates that each feature data set holds different types of information which complement each other (Figure 25).

Noteworthy that, in all these classification scenarios with different predictor variables, the Orkney population had invariably a significantly lower classification error rate when compared to Vis and Korčula populations (Table 5). Samples from Vis and Korčula were mostly mistaken among these two populations and to a lower extent with Orkney.

Although statistically significant differences were found in the levels of certain glycans and phenotypes across populations, these differences were not captured in the PLS-DA and PCA analyses. On the contrary, RF algorithm achieved a satisfactory classification of populations for a combined set of glycan and phenotype features and was able to yield a good separation of Orkney from the other two populations for all feature data sets. The different results obtained might be due to the distinct nature of the algorithms employed.

Table 5. Random Forest confusion matrices for the classification of the population cohorts. The confusion matrices show the performance of Random Forest algorithm in classifying testing samples using all plasma glycans (A), all IgG glycans (B), phenotypes (C) and the set of 30 most important features of all data sets (D). The set of 30 features yielded the best classification, plasma glycans and phenotypes had a slightly worse performance and IgG glycans presented the higher error rate. In all cases, Orkney population showed the lowest classification error per group. Each row of the matrix represents the instances in the actual population, while each column represents the instances in a predicted population (0 for Vis, 1 for Korčula and 2 for Orkney). The "Error" column indicates the test set errors for the classification of each group and for the overall classification of the testing samples.

A) Plasma glycans (all)

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	163	27	4	16
1	46	154	5	24.9
2	9	8	181	8.6
				16.58

B) IgG glycans (all)

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	112	54	11	36.7
1	55	144	9	30.8
2	9	6	197	7.1
				24.12

C) Phenotypes

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	163	18	9	14.2
1	33	162	21	25
2	6	12	173	9.4
				16.58

D) Set of 30 most important features

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	162	22	5	14.3
1	28	180	11	17.8
2	5	7	177	6.3
				13.07

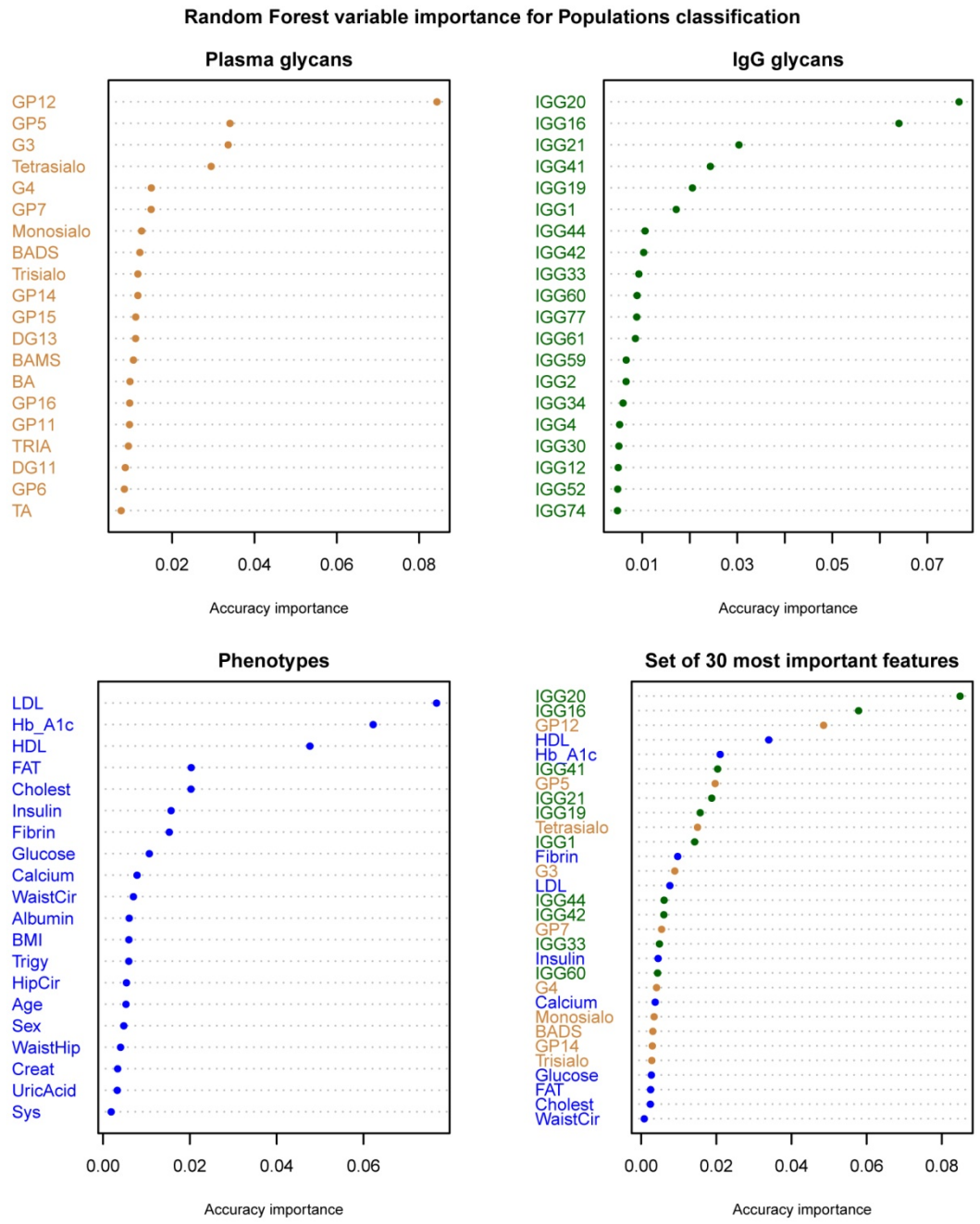


Figure 25. Random Forest variable importance for the classification of the population cohorts. The variable importance plots correspond to plasma glycans (upper left panel), IgG glycans (upper right panel), phenotypes (lower left panel) and the set of 30 most important features of all data sets (lower right panel). For plasma and IgG glycans only the top 30 features are displayed. The measure of variable importance presented is the mean decrease in accuracy estimated by comparing the accuracy in classification without and with permutation of the values of each predictor variable. When a given variable has little predictive power, its permutation will not cause substantial difference in accuracies, therefore a higher decrease in accuracy is indicative of a more important variable. Variable labels are coloured as gold for plasma glycans, green for IgG glycans and blue for phenotypes.

In an attempt to model the genetic structure of the populations, three different approaches – Random Jungle (RJ), discriminant analysis of principal component (DAPC) and correlation adjusted scores (CARE) – were applied to both sets of SNPs and their results compared.

RJ showed a better performance for the set of all SNPs with an error rate of 11% and the most important SNPs mainly located on chromosome 2. The classification based on the set of glycan-related SNPs had an error rate of 26% with the chromosome 6 harbouring the most important SNPs. In both cases, Vis and Korčula samples tended to be mixed among them whereas Orkney samples were better differentiated showing the lowest classification error per group (the confusion matrices for the classification are shown in Table 6).

The DAPC analysis was able to separate the populations using both set of SNPs with the set of all SNPs achieving a more defined separation (Figure 26, upper panels). In both cases, the first component differentiated Orkney from the other two populations while the second component differentiated Vis from Korčula. Since the set of all SNPs yielded better results, the results of its analysis will be considered below. SNPs with the largest contributions to the first discriminant component were mainly localized on chromosomes 2, 4 and 6, while SNPs contributing to the second discriminant component apparently did not share a preferential location. Regarding their genotype frequency, the SNPs related to the first component showed similar genotype patterns for Vis and Korčula while exhibiting a different genotypic profile for Orkney (Figure 27). The SNPs related to the second component showed less pronounced differences in the genotype frequencies which were present either for Vis or Korčula (Supplementary figure 13).

In order to verify the results obtained with RJ and DAPC while accounting for any bias in the data and controlling for the possibility of overfitting of the models, the analyses were repeated using randomised groups, i.e. randomly assigning each sample to one of the populations. For the randomised data in both set of SNPs, the RJ classification error greatly increased and the DAPC analysis showed a single cluster as opposed to the cluster arrangements observed for the non-randomised data (Table 6 and Figure 26, lower panels). This fact indicates that an underlying population structure based on genotype data exists and is lost when populations are randomised. Moreover, this genetic structure was to some extent captured by SNPs related to glycosylation.

Table 6. Random Jungle confusion matrices for the classification of population cohorts based on genotype data. The confusion matrices show the performance of Random Jungle algorithm in classifying the population cohorts using the set of all SNPs (A) and the set of glycan-related SNPs (B) for both non-randomised and randomised populations. The set of all SNPs yielded a quite satisfactory classification achieving a much smaller error than the set of glycan-related SNPs. In both cases, the Orkney population showed the lowest classification error per group. The classification performance was greatly diminished when the population classes were randomized suggesting the disruption of a certain population structure. Each row of the matrix represents the instances in the actual population, while each column represents the instances in a predicted population (0 for Vis, 1 for Korčula and 2 for Orkney). The "Error" column indicates the test set errors for the classification of each group and for the overall classification.

A) Set of all SNPs

True populations

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	473	158	6	25.7
1	41	658	6	6.7
2	0	7	641	1.1
				11

Randomised populations

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	150	311	176	76.5
1	141	366	198	48.1
2	143	333	172	73.5
				65.4

B) Set of glycans-related SNPs

True populations

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	386	182	69	39.4
1	88	547	70	22.4
2	42	72	534	17.6
				26.3

Randomised populations

	Predicted outcome			**Error (%)**
	0	**1**	**2**	
0	142	339	156	77.7
1	148	366	191	48.1
2	159	355	134	79.3
				67.7

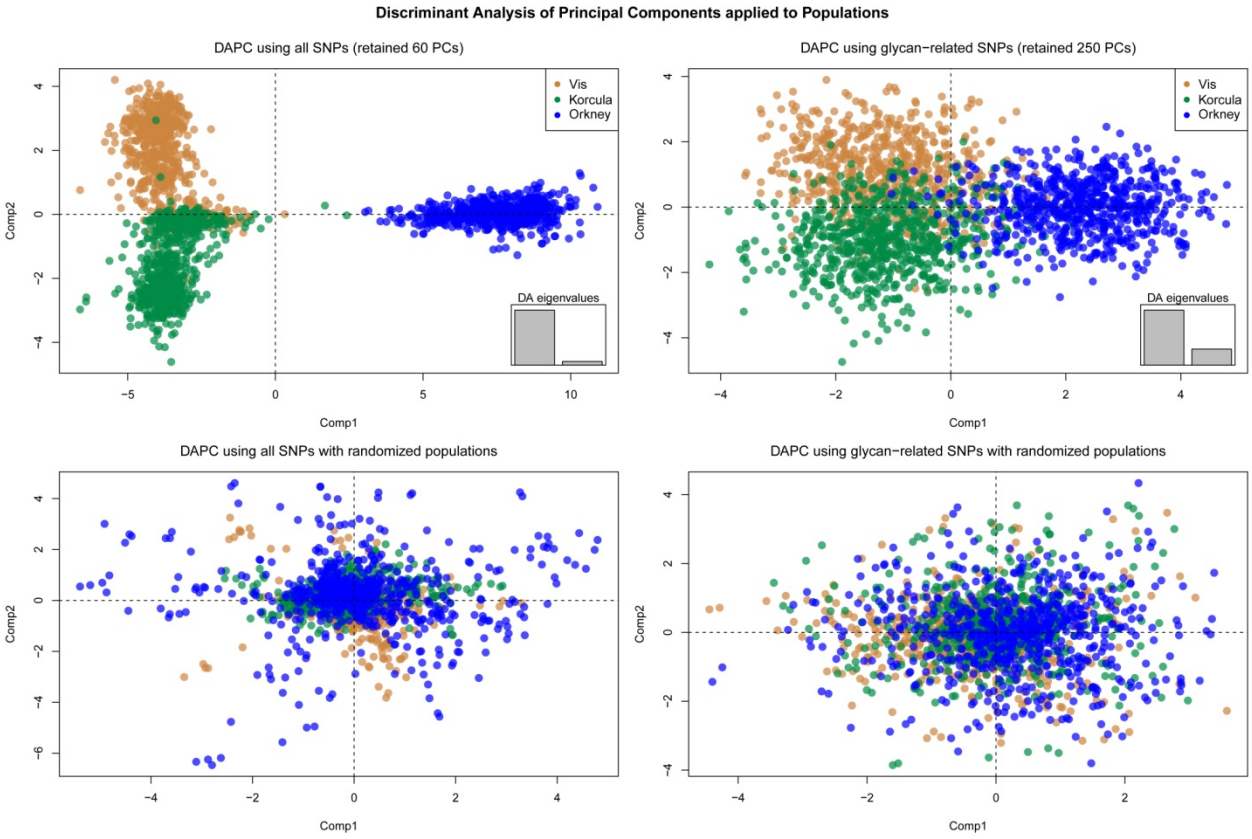


Figure 26. DAPC analysis of the population cohorts using genotype data. The score plots representing the data samples by the two first discriminant components (Comp1 on the x-axis and Comp2 on the y-axis) are shown for the set of all SNPs (left panels) and the set of glycan-related SNPs (right panels). In the analysis of the non-randomised data, the first discriminant component differentiated Orkney from the other two populations, while the second discriminant component differentiated Vis from Korčula (upper panels). The observed genetic structure of populations was lost when the population classes were randomised (lower panels). Populations are coloured as gold for Vis, green for Korčula and blue for Orkney.

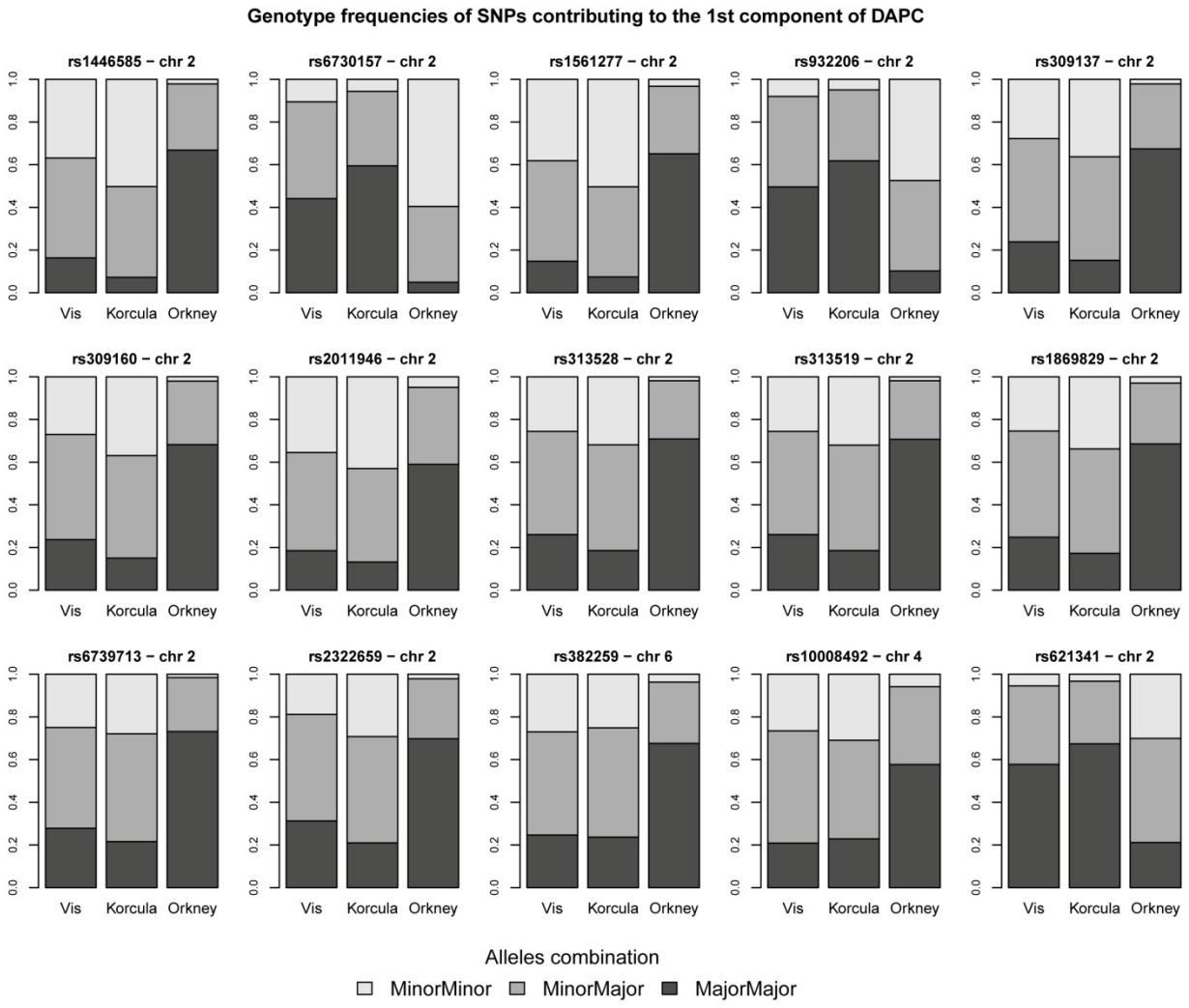


Figure 27. Genotype frequencies of the 15 SNPs most contributing to the first discriminant component of the DAPC analysis of the population cohorts. SNPs are mainly located on chromosomes 2, 4 and 6 and show different genotype patterns for Orkney when compared to Vis and Korčula which exhibit similar patterns. Genotypes are coded in grey shades with light grey corresponding to minor-minor allele combination, medium grey corresponding to minor-major allele and dark grey corresponding to major-major allele.

A common finding in the SNP selection analyses carried out by all three methods was the fact that among the top 100 most important SNPs identified when analysing the set of all SNPs were present some of the most important SNPs identified when using the set of glycan-related SNPs (Table 7). The rs494620 variant found in the solute carrier family 44 gene (SLC44A4) in chromosome 6 was the only SNP assigned as important by all three methods.

Table 7. Glycan-related SNPs present among the most contributing SNPs for the genetic structure of populations. List of glycan-related SNPs found among the top 100 most important SNPs for the analysis of population differentiation based on the set of all SNPs. Information regarding the chromosome harbouring the SNP, the SNP alleles, the associated gene and the gene description are displayed. The SNP selection methods which detected each of the SNPs are also indicated: RJ stands for Random Jungle, DAPC for discriminant analysis of principal components and CARE for correlation adjusted scores. n.a.: description from Ensembl database was not available.

SNP	Chr	Alleles	Gene	Gene description	Methods
rs494620	6	G/A	SLC44A4	solute carrier family 44, member 4	RJ, DAPC, CARE
rs644827	6	T/C	SLC44A4	solute carrier family 44, member 4	RJ, DAPC
rs2242665	6	C/T	SLC44A4	solute carrier family 44, member 4	RJ, DAPC
rs660550	6	C/A	SLC44A4	solute carrier family 44, member 4	RJ, DAPC
rs651970	2	A/G	MGAT5	mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase	RJ
rs9267649	6	A/G	NEU1	sialidase 1 (lysosomal sialidase)	RJ
rs845739	5	G/T	AC012603.1	n.a.	RJ
rs3901856	6	A/G	SLC35F1	solute carrier family 35, member F1	RJ
rs2301010	5	T/C	MAN2A1	mannosidase, alpha, class 2A, member 1	CARE

The relative performance of the three SNP selection methods and the extent of agreement of their results were assessed by selecting the SNPs consistently detected by all methods within the top 100 and comparing their ranking position with each method. A total of 35 SNPs were found in common in the analysis involving the set of all SNPs while 27 SNPs were commonly identified in the analysis involving the set of glycan-related SNPs. The comparison of SNP ranks given by the three methods revealed a fairly good agreement of results across the methods particularly for the best positioned SNPs in the set of all SNPs (Figure 28). The set of 35 SNPs (Supplementary table 7) are mainly located on chromosome 2 in a region spanning genes related to mRNA processing, protein biosynthesis and trafficking and on chromosome 6 in a region comprehending genes involved in the immune system response, cell interactions and glycosylation-related processes (Figure 29).

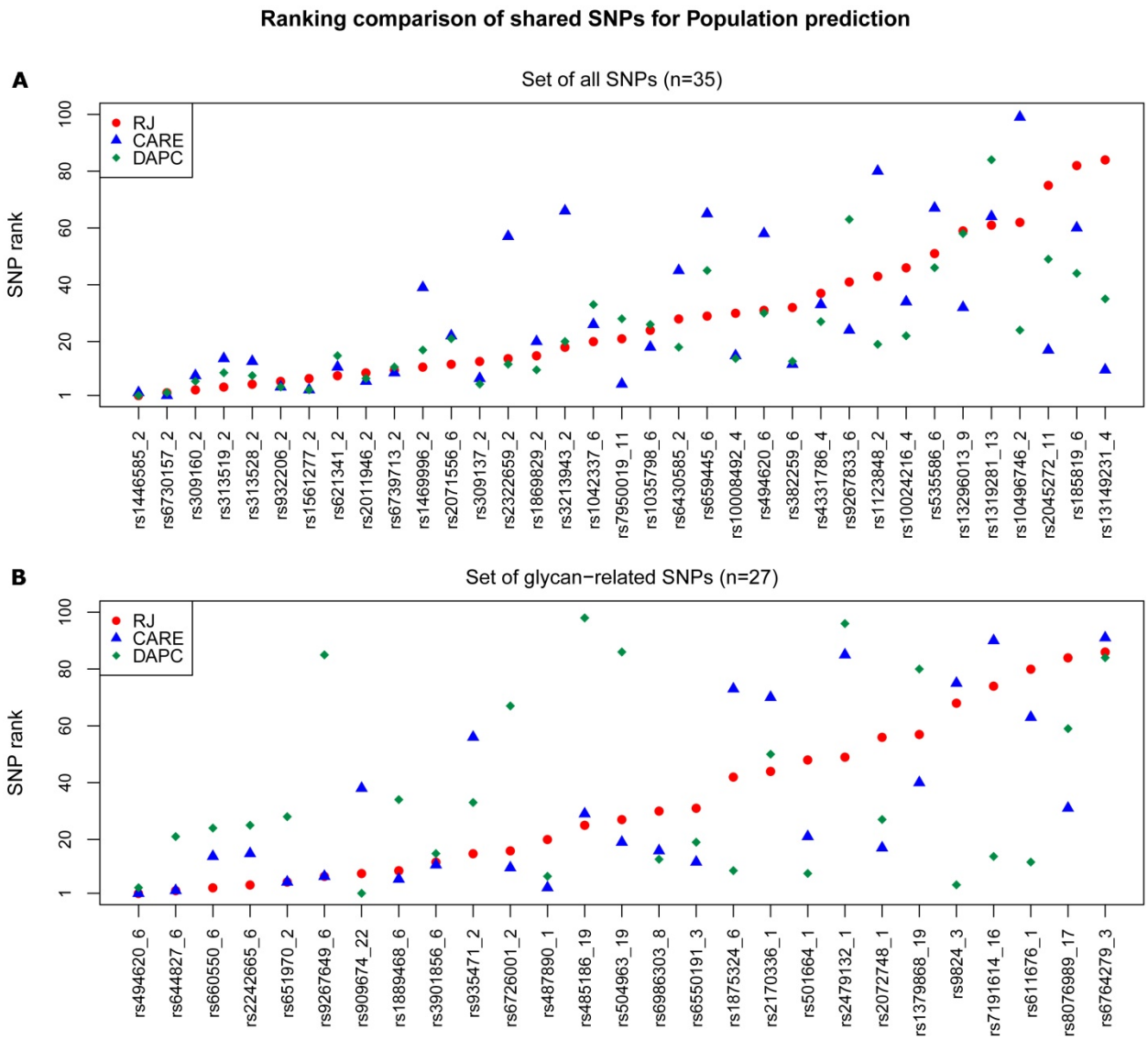


Figure 28. Ranking comparison of the most important SNPs consistently identified by the three investigated approaches in the analysis of the genetic structure of populations. A total of 35 SNPs were commonly detected among the top 100 SNPs by all methods in the analysis performed with the set of all SNPs (A) and 27 SNPs in the analysis performed with the set of glycan-related SNPs (B). The plots represent the ranks of the selected SNPs according to each method: Random Jungle (RJ; red circles), correlation adjusted scores (CARE; blue triangles) and discriminant analysis of principal components (DAPC; green diamonds). The first SNPs in the set of all SNPs showed the best ranking agreement between all three methods. The SNPs were ordered using the ranking obtained by RJ.

Genetic context of important polymorphisms for Population prediction

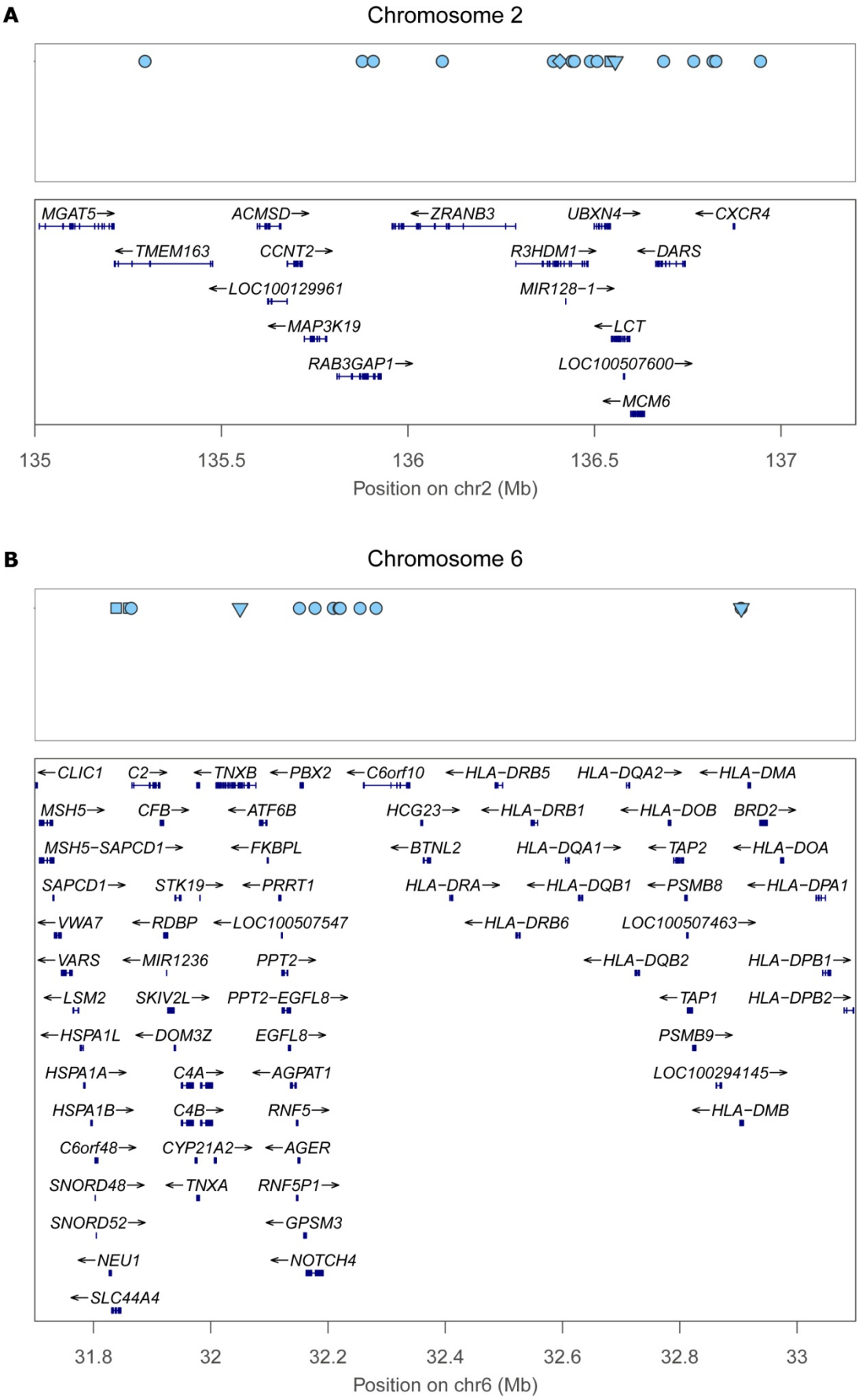


Figure 29. Genetic context of the most important SNPs consistently identified by the three investigated approaches in the analysis of the genetic structure of populations. In the analysis performed with the set of all SNPs, the majority of SNPs detected among the top 100 SNPs by all methods are mainly distributed along the chromosomes 2 (A) and 6 (B). The annotation categories for SNPs are: non-synonymous (inverted triangle), synonymous or UTR (square), MCS44 Placental (diamond) and no annotation (filled circle).

3.7. N-glycome association studies

Possible associations between glycan levels and SNPs were examined to get more insights into the genetic background influencing the glycosylation process.

Since the individuals from the three population cohorts showed similarities of their glycan profiles (as previously discussed in section 3.3), glycome-association analyses were performed in the pooled data of all populations. In this way, more information is included in the analyses and the derived results are not population-specific but can be interpreted in the light of general population trends.

Three methods – Random Jungle (RJ), correlation adjusted scores (CARE) and bayesian sparse linear mixed model (GEMMA) – were applied for the selection of SNPs associated with plasma glycans, IgG glycans and phenotypes. The SNPs were used as explanatory variables and each glycan or phenotype feature was used as a single quantitative response variable in the analysis. In order to assess the agreement between methods and obtain a consensus SNP selection, the results of the three methods were combined for each feature by selecting the SNPs consistently detected by all methods within the top 100 SNPs.

For plasma, 64 SNPs were found in common by all methods and 4 of them had an association with more than one trait. Only five of the glycan traits did not present SNP associations which were fetched by all methods. For IgG, 94 SNPs were identified by all methods with 13 of them being first-ranked by the three methods. The methods did not present shared associations for 12 of the glycan traits and 17 SNPs showed an association with two or more glycan traits. For phenotypes, the methods detected in common 11 SNPs associated with only 6 of the traits, namely systolic pressure, HDL, triglycerides, insulin, calcium and uric acid. The lists of associations found together with the annotation relative to the genes overlapping the variation or the nearest gene to it, the chromosome where the variation is located as well as the rank position of the variation by each method are presented in Supplementary table , Supplementary table 9 and Supplementary table 10 for plasma, IgG and phenotypes, respectively. Overall, CARE and

GEMMA appear to perform better than RJ as evidenced by the higher ranks presented for the majority of associations (Figure 30).

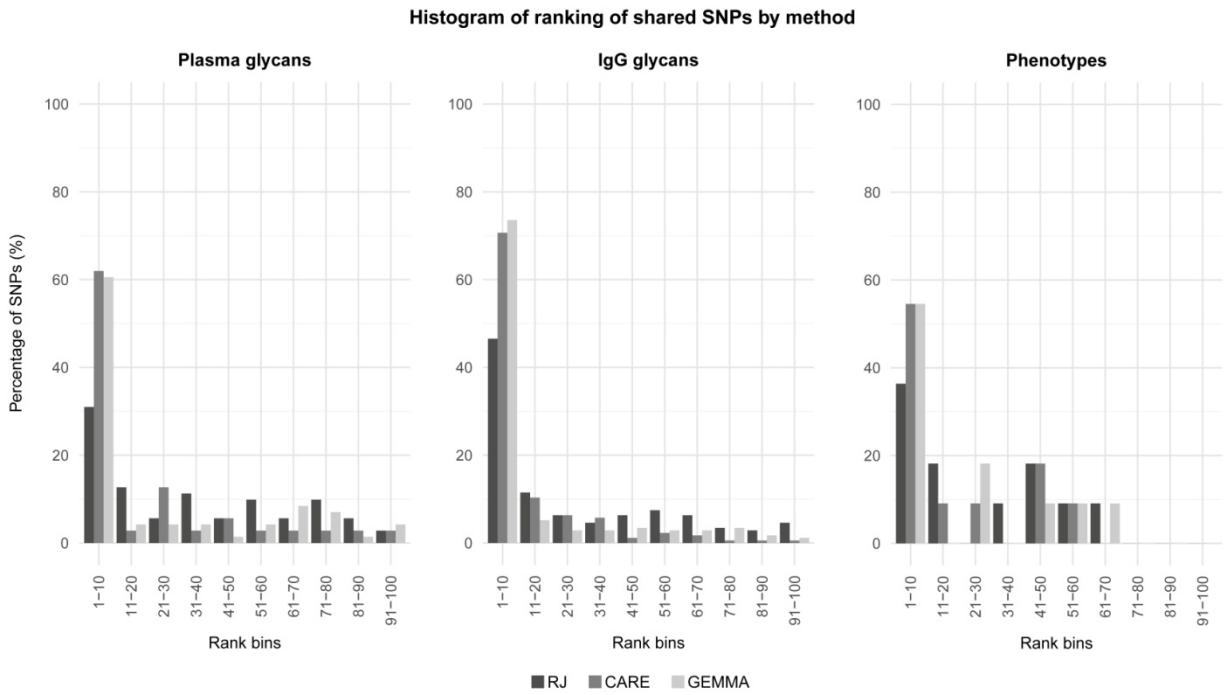


Figure 30. Histograms of SNPs ranking by method. The histograms represent the ranking by methdod of SNPs identified to be in association with the traits from each feature data set by all three methods. For all feature data sets, CARE and GEMMA tend to rank the SNPs in higher positions than the RF.

Additionally, GEMMA algorithm yielded two estimation measures of heritability: PVE and PGE representing the proportion of variance in the analysed traits explained by both small and large effect size SNPs and by the large effect size SNPs alone, respectively (refer to section 2.6.7 for an explanation of the measures).

In the case of plasma glycans, estimates of PVE indicate that between 20% and 45% of the heritability of the majority of glycan traits can be explained by the available SNPs (Figure 31A). Estimates of PGE were mainly below 20% and only three plasma traits showed estimates above 30% (Supplementary figure 14A). Regarding the IgG glycans, almost half of the traits had estimates of PVE between 40% and 60% with glycans belonging to the Neutral derived group showing the highest estimates (Figure 31B). Similarly to what was observed for plasma glycans, only few IgG traits had PGE values above 30% and glycans belonging to the Charged group presented the highest values (Supplementary figure 14B). For phenotypes, estimates of PVE

were above 40% for albumin, hip circumference and triglycerides (Figure 31C), while estimates of PGE were above 40% for HbA1c and insulin with the later having PGE of more than 65% (Supplementary figure 14C). As any other measures of heritability, these estimates can be influenced by environmental factors and their interpretation should take that fact into consideration.

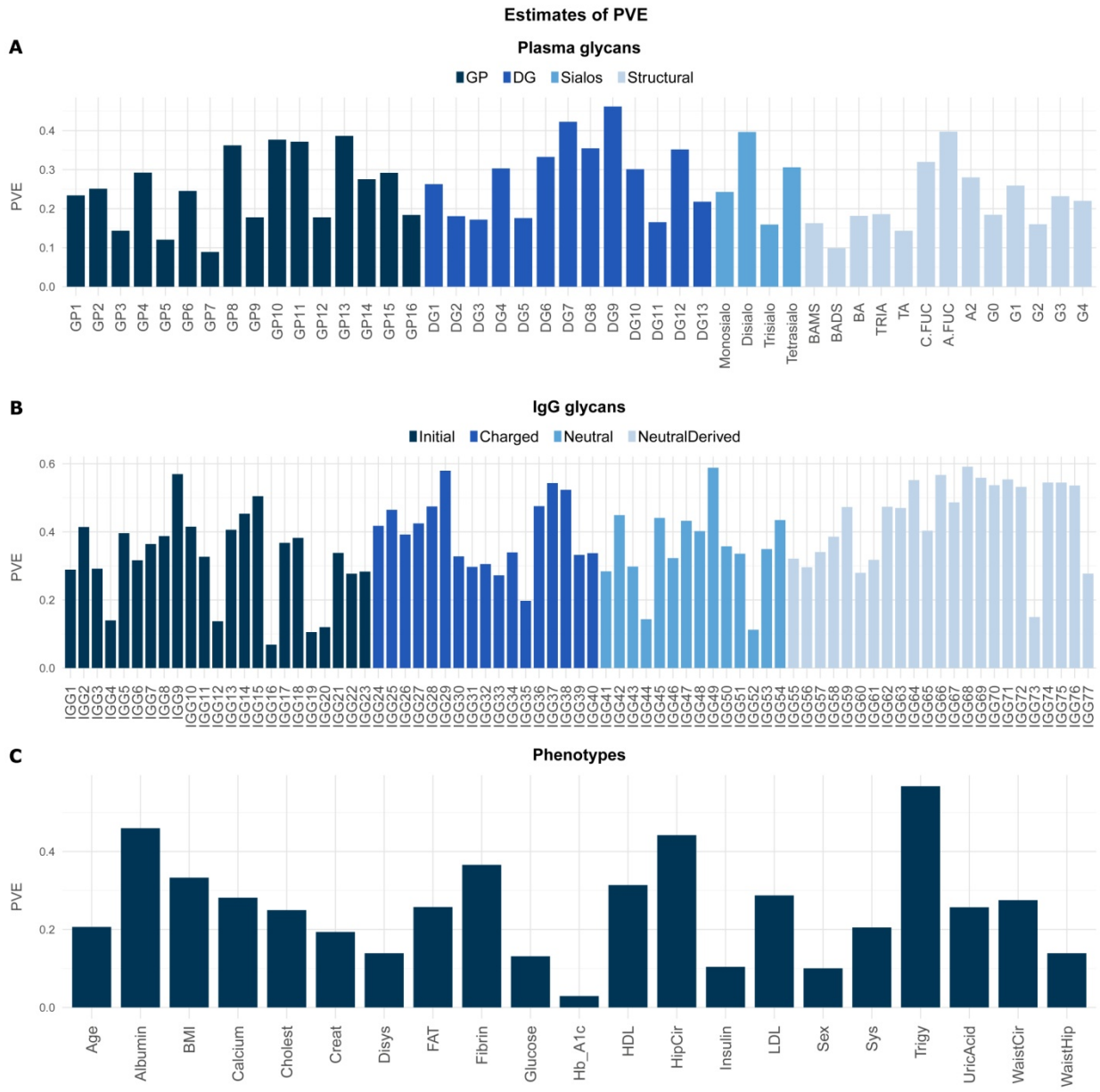


Figure 31. Proportion of the variance of all traits of the three feature data sets explained by genotype. The estimates of PVE by the Bayesian sparse linear mixed model are represented for each trait of plasma glycans (A), IgG glycans (B) and phenotypes (C). For plasma and IgG glycans the colours of the bars represent the glycan groups as indicated in the corresponding legend above the barplot.

4. DISCUSSION

4.1. Glyco-phenotypes in the general population

A rather similar glycan profile with very little changes over time is shown by the majority of people and can be referred to as the "normal profile". Due to genetic background and/or environmental conditions, deviations from this normal profile might occur and their significance should be examined.

The association of certain glycan alterations with biochemical traits and clinical conditions was studied after the identification of several individuals exhibiting a glycan profile that significantly differed from the general population. Using computational approaches, a total of six groups were formed based on particular glycan aberrations and a comprehensive analysis of several biochemical traits and medical data was able to identify some of the shared characteristics within each group. While some of the observed glycan aberrations were associated with serious conditions, other glycan changes apparently did not reflect any peculiar medical status.

Although the results only suggest a possible association between plasma glycan patterns and (patho)physiological conditions, they revealed the existence in the general population of glyco-phenotypes which might represent risk factors for the development of specific diseases. Moreover, the observed deviations from the normal plasma glycan profiles in the six groups of individuals were much more pronounced than changes reported to occur in common diseases and the incidence of these deviations in the studied population was much lower than the incidence of any common disease (Pucic *et al.*, 2010). Together, these facts indicate that these common aberrations from the normal plasma might originate due to rare mutations and/or rare combinations of common mutations instead of being a result of altered physiological conditions.

The genotype influence upon the group structuring and its association with the particular glycan aberrations presented by each group were analysed by three methods. On the one hand, the poor performance of all employed methods suggests a lack of genetic structure behind the glyco-phenotype groups. On the other hand, the small number of samples in each group might not be sufficient to select representative genotypes containing significant information that would allow a proper discrimination of the groups.

The validity of these glyco-phenotypes findings and the exploration of a possible genetic background effect should involve the examination of a larger number of individuals with the identified aberrations.

4.2. Internal clustering structure of isolated populations

Finding particular glyco-phenotypes in the general population motivated the examination, at the level of single population, of the presence of larger clusters that could be characterised by certain glycan or phenotype features. Comparing the characteristics of clusters with similar consensus profiles can bring new insights into the causes of glycosylation and phenotypic alterations.

A first approach in this direction has been described in detail in a previous work investigating the clustering structure of individuals based on their plasma and IgG glycan profiles in four isolated population cohorts (Klarić, 2012). The study aimed to determine the optimal number of clusters for a population and to assess the stability of the constituted clusters. Briefly, the k-means algorithm was used to perform the clustering of the samples and the consensus clustering approach was applied to obtain a characterization of the clusters robustness. The consensus clustering consisted in the repetition of k-means algorithm with different subsets of the data and the construction of a summary matrix – consensus matrix – where each element is defined as the ratio between the number of times two samples clustered together and the number of times the same samples were selected for the k-means clustering (Monti *et al.*, 2003). Despite the fact that clustering was performed on glycan raw data (i.e. without age, sex and batch correction), the analysis showed a certain level of internal structure of the populations with relatively robust clusters varying in number from 3 to 6 according to the population cohort.

In the present thesis, to further address the subject of glyco-phenotypes, the feature-specific profiles of each cluster were inspected and the analyses were extended to the phenotype feature data set. However, in the adopted approach the data used was corrected by age, sex and for batch effects and the clustering algorithm employed was the novel affinity propagation clustering method.

The affinity propagation algorithm yielded a number of clusters in the same range as the ones previously obtained and the inspection of the cluster profiles revealed the existence of similar data patterns across populations. Due to these cluster similarities between populations, the analysis of the pooled data of all populations did not produce a cluster for each population;

however, it did replicate the cluster profiles observed for the individual populations which will be discussed below.

For plasma glycans, two main cluster profiles were observed. In the first profile, GP9, DG5 and BADS which mainly refer to biantennary and digalactosylated glycans structures (A2G2) showed opposite pattern to DG6, Monosialo, BAMS and C.FUC which mainly represent fucosylated biantennary and digalactosylated glycans (FA2G2). In the second profile, peaks GP7, DG5, Monosialo, BAMS and G2 which have in common biantennary digalactosylated glycans with one sialic acid (A2G2S1) presented a contrary pattern to BADS and C.FUC which refer to biantennary digalactosylated glycans with two sialic acids and fucose.

For IgG glycans, a division into two clusters presented a clustering profile showing opposite tendency of structures with core fucose and without galactose (IGG3, IGG43 and IGG55) and of structures with core fucose and two galactoses (IGG13, IGG17, IGG56 and IGG57). Increasing the number of clusters resulted in the further separation of the above clusters according to the level of sialylation (IGG24-IGG27) and the presence of bisecting GlcNAc (IGG62-IGG69).

The phenotype data produced a cluster structure with samples divided according to three levels of BMI, waist circumference and waist-to-hip ratio. In the cluster with the lowest levels of these measures, the majority of individuals are female and also present low values of uric acid.

The fact that groups of individuals from geographically separated populations display similar characteristics suggests the presence of an underlying structure based on glycans and phenotypes that is shared between populations. Particularly, the presence of groups of samples with different combinations of glycan structural features might reflect biological interactions at the level of glycosylation. Together, these findings could be used to attempt to identify the SNPs responsible for such specific feature signatures by analysing the genotype profiles of each cluster and, in this way, address the study of the association between SNPs and glycans and phenotypes from a different angle.

4.3. Association between N-glycans and phenotypes

Besides the genotype effect, environmental factors also seem to influence glycan structures to a certain degree. The changes in glycosylation due to common biochemical and lifestyle parameters (herein designated by phenotypes) have been previously analysed for all glycans in Vis and for the structural glycosylation features in Korčula (Knezevic *et al.*, 2010; Knezevic *et*

al., 2009). However, such analysis has not yet been conducted regarding the IgG glycans. Here, the analysis of the association between plasma glycans and phenotypes was complemented by extending it to the Orkney population and a comprehensive analysis of the association between IgG glycans and phenotypes was performed for the first time.

In plasma glycans, statistically significant correlations present in all three populations were mainly observed for phenotypic traits linked with obesity and unhealthy lifestyle such as BMI, waist and hip circumferences and to lower extent cholesterol, LDL and triglycerides. Associations with certain phenotypes were found to be more uneven across populations while being exhibited by only some of the populations such as cholesterol in Vis, insulin in Korčula and fibrinogen in Vis and Orkney. Although these patterns can be a consequence of population predisposition and be regarded as population-specific, it should be taken into consideration the fact that these phenotypes have been shown to be subjected to large intra-individual variation (Demacker *et al.*, 1982).

In IgG glycans, only few associations were shown to be statistically significant. This could be explained by the strict threshold of significance resulting from the criteria used for the correction of multiple testing, a subject where the best approach to follow is still a matter of debate. Nonetheless, some patterns of correlation which did not pass the threshold of significance were similar in all three populations and shall be briefly commented on.

Body fat parameters, triglycerides, glucose, insulin, HbA1c, fibrinogen and uric acid were directly correlated with agalactosylated structures (IGG55) and inversely correlated with digalactosylated structures (IGG57), indicating that galactosylation is decreased in conditions where these parameters are elevated. A shared characteristic to these phenotypes is their known connection to a certain degree with a pathological status such as obesity, diabetes or cardiovascular problems.

IgG with reduced content of galactose has been reported as a common feature in a number of autoimmune diseases which are known to be characterised by inflammatory conditions (Ciric *et al.*, 2005; Huhn *et al.*, 2009). Obesity not only presents a chronic low-grade inflammation as it has been implicated in the susceptibility to autoimmune diseases such as diabetes (Golay & Ybarra, 2005; Kahn *et al.*, 2006; Procaccini *et al.*, 2011). Although the obesity-autoimmune relationship is still not thoroughly understood, it appears to be the result of complex interactions between several factors and conditions where hormones and neural mediators may have an important role (Steinman *et al.*, 2003).

Fibrinogen is a protein involved in blood clot formation and elevated levels of fibrinogen have been identified as major risk factor for cardiovascular diseases with possible mechanisms through which fibrinogen might operate being suggested (Cook & Ubben, 1990; Danesh *et al.*, 2005; Stec *et al.*, 2000). Besides its clotting factor role, fibrinogen seems to function as a signalling molecule in the inflammatory response and has been recently linked with diseases presenting an inflammatory component like multiple sclerosis, Alzheimer's disease and rheumatoid arthritis (Davalos & Akassoglou, 2012).

These associations found between IgG glycosylation patterns and phenotypes which are related to pathological conditions suggest that glycosylation might be involved in the intricate interplay of factors existent in the pathways leading to these disorders.

Another interesting pattern which was only observed for the Orkney population is the positive association between calcium and glycan structures containing bisecting GlcNAc and the negative association with glycan structures without bisecting GlcNAc. While the presence of bisecting GlcNAc on IgG increases its effector functions (Takahashi *et al.*, 2009), calcium signalling has an important role in immune function by participating in diverse mechanisms of the immune system (Diamantstein & Odenwald, 1974). Moreover, N-glycans on T-cell glycoproteins are found to be involved in triggering T-cell functions (Walzel *et al.*, 2006).

Due to the fact that the IgG glycans are a filtered set of the plasma N-glycans, the main IgG chromatographic peaks can be combined into 11 plasma peaks. The comparison of the associations of both plasma peaks and their corresponding IgG peaks with phenotypes revealed a good agreement between the majority of peaks. The fact that IgG glycans captured the associations from the pool of all plasma N-glycans suggests that these associations might actually be connected to IgG protein. The few associations shown to be contradictory between peaks might reflect their specificity to a particular protein.

The ensemble of N-glycans in human plasma originates from a variety of glycoproteins which differentially contribute to the general glycan composition observed. Establishing the individual glyco-contribution of each protein and exploring the specific association of their glycan structures with other features might bring more detailed information about the influence of specific glycans in protein function as well as their connection to pathophysiologic states.

4.4. Diabetes: a case example

Diabetes affects a high number of people worldwide and has been the continuous focus of many research studies intended to understand the wide range of mechanisms leading to disease. The goal of such studies is the discovery of biomarkers that can reliably detect the presence or susceptibility to diabetes at an initial stage since an early diagnosis can help in disease prevention.

The diabetes status of a subset of samples from the three population cohorts was available and the samples assigned into one of three groups: non-diabetic, pre-diabetic and diabetic. The three groups were compared with respect to plasma, IgG, phenotype and genotype profiles using graphical, statistical and classification methods.

Plasma glycan profiles did not present significant differences between the three analysed groups and their use as predictor variables did not discriminate between groups. Studies performed to examine the alteration of N-glycans in the serum glycoproteins in diabetes are scarce and a first finding reports the elevated levels of glycoprotein fucose in diabetes (McMillan, 1972). The analysis of the N-glycans in the serum of the model mice of type 2 diabetes with obesity also revealed an increased fucosylation of N-glycans but the same modification in human serum was found to be small (Itoh *et al.*, 2007). Although pointing to a possible increase of fucose content in diabetes and suggesting its association with the pathophysiology of diabetes, these studies should be viewed and interpreted with caution since they were based on small sample sizes and further studies are necessary to confirm the findings. The apparent absence of significant differences between the plasma glycan profiles of the three groups analysed in the present work could arise from the fact that the N-glycans analysed originate from an ensemble of proteins which might conceal potential differences of specific proteins.

Analysing the changes in glycosylation patterns of single proteins in diabetes could give precise insights about which proteins are more prone to be targets of abnormal glycosylation in the disease. The comparison of IgG glycan profiles in the three groups did not reveal statistically significant differences and similarly to plasma glycans showed a poor predictive power in classifying the samples. The investigation of N-glycan structures of an acute-phase protein showed an increase in fucosylated glycans which was significant in individuals with inflammation but not in individuals with type 2 diabetes (Higai *et al.*, 2003). Furthermore, fucosylation of IgG was found to be significantly increased in patients with rheumatoid arthritis

(Gornik *et al.*, 1999). Perhaps the alterations to which IgG is subjected in diabetes differ from those presented by chronic inflammatory conditions.

Non-enzimatic glycosylation or glycation, the addition of glucose to proteins without the controlling action of enzymes, has an increased rate of occurrence in diabetics and has been indicated to contribute to several long-term complications in diabetes. Many proteins are known to be glycated to a much higher degree in diabetics than in normal individuals including the well studied case of glycated hemoglobin (HbA1c), lens crystallins, basic myelin protein, collagen and also IgG. The levels of glycated IgG were found to be significantly higher in diabetic than in normal individuals suggesting that non-enzimatic glycosylation of IgG might be associated with changes in its function (Bunn, 1981; Vlassara *et al.*, 1986). The relative impact of structural changes in N-glycans and of non-enzimatic glycosylation in the pathophysiology of diabetes and their target proteins requires more comprehensive studies concerning single proteins.

Contrary to plasma and IgG profiles, the phenotype data showed noticeable differences at the level of HbA1c and glucose and less marked differences for age, systolic blood pressure, BMI, waist-to-hip ratio and waist circumference. In all cases, the diabetic group presented higher levels and the non-diabetic lower levels of the features. HbA1c is routinely used for monitoring long term glycemic control in people with diabetes and is currently the most commonly used marker in the diagnosis of diabetes. High blood glucose is a sign of diabetes or that a person is at high risk for developing the disease, although it is considered to be an insufficient indicator of diabetes. The other altered phenotypes between groups are more likely to be presented by individuals with diabetes than healthy individuals and so are considered either as predisposing factors or symptoms of diabetes. Thus, it was not surprising that both HbA1c and glucose were found to markedly vary across groups while the remaining phenotypes varied to a lesser extent. The association of these phenotypes with diabetes was further verified by their high importance and predictive power in the separation and classification of the groups.

The analysis of the genotype contribution to the diabetes status revealed SNPs harboured in genomic regions comprising genes directly or indirectly related to diabetes as briefly described below.

The first most important SNP is located in a region near the neuronal calcium sensor-1 gene (NCS1), a calcium binding protein involved in the molecular mechanisms of calcium and metabolic signalling by which cells are able to adjust insulin secretion in response to glucose stimulation (Gromada *et al.*, 2005).

Several SNPs overlap the cyclin-dependent kinase 19 gene (CDK19) whose potential function in the nervous system has been suggested upon the results of functional analyses of its closest orthologue in Drosophila, CDK8. CDK8 was shown to regulate dendritic development and to play a major role in the development of peripheral sensory organs including the eye (Mukhopadhyay *et al.*, 2010). Different changes of dendritic structures have been reported in glaucoma, a condition that has diabetic retinopathy as a possible cause.

The initial portion of doublecortin-like kinase 1 gene (DCLK1) harbours four high-ranked SNPs. DCLK1, a neuronal function-related gene highly expressed in the ganglion cell layer of retina, is shown to be downregulated in rat model of diabetes suggesting its association with reduction of synapses observed in diabetic rat retinas (Brucklacher *et al.*, 2008; Van Kirk *et al.*, 2011). Also, DCKL1 regulates microtubule polymerization and stabilization and has been found to be a marker for the identification of pancreatic stem cells which could be used in cell replacement therapies such as in type 1 diabetes (Mwangi & Srinivasan, 2010).

The glucocorticoid induced transcript 1 gene (GLCCI1), located in close proximity to the Islet Cell Autoantigen 1 gene (ICA1), also contains important SNPs. A region flanked by these two genes has been identified as a glaucoma susceptibility locus due to the presence of a common variant associated with elevated intraocular pressure (BMES & WTCCC2, 2013). Both genes have been shown to be expressed in the human eye and are plausible candidates in the determination of intraocular pressure, a major risk factor for the development and progression of glaucoma. On the one hand, ICA1 has been indicated as an auto-antigen in insulin-dependent diabetes mellitus and glaucoma is a well known eye problem in people with diabetes. On the other hand, since glucocorticoids increase the risk of glaucoma by raising the intraocular pressure, GLCCI1 could be implicated in intraocular pressure via its response to endogenous cortisol.

SNPs located in the upstream region and in a non-coding region of the deleted in colorectal carcinoma gene (DCC) were also identified among the most important SNPs. The region containing the DCC gene was suggested to be associated with autoimmune diseases in a large study comprising families with type 1 diabetes, multiple sclerosis and rheumatoid arthritis (Merriman *et al.*, 2001).

The gene cluster of chemokine receptors (CCR) is a highly enriched area for chemokine receptor genes and harbours some SNPs of interest. Small signaling proteins secreted by cells called chemokines and their corresponding receptor genes induce calcium signaling in cells and are

involved in processes of the immune system. Up- or down-regulation of both chemokine and chemokine receptor gene expression has been observed in a wide range of inflammatory and autoimmune diseases (Navratilova, 2006). Also, functional polymorphisms in chemokine-related genes have been implicated in the pathogenesis of type 1 diabetes and its microvascular complications (Yang *et al.*, 2004).

In the upstream region of the cardiac ryanodine receptor calcium release channel gene (RYR2), a single SNP isolates itself from all the other neighboring SNPs. RYR2 gene is central to the heartbeat cycle while regulating the calcium homeostasis responsible for the heart muscle cell contractions. Alteration of calcium signalling was found to be present in diabetic cardiomyopathy and to be related with partial loss of RYR2 function (Yaras *et al.*, 2005). Furthermore, mutations in this gene have been reported to cause arrhythmias of the right ventricle in a condition known as arrhythmogenic right ventricular cardiomyopathy (Milting *et al.*, 2006). Other SNPs in the flanking region of RYR2 were linked with heart failure conditions in association studies (information retrieved from the NCBI dbSNP database website, see http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=4659764).

The agreement with previous findings in some cases and the suggestive link with diabetes in other cases present evidence in favour of the use of alternative methods to univariate regression for SNP selection. Nonetheless, a thorough inspection of other high-ranked SNPs and their flanking regions should be carried out in order to investigate whether other novel associations with diabetes can arise.

4.5. Population-specific patterns

Isolated populations derived from factors like geographic or cultural isolation present a level of genetic discontinuity. Such differentiated cohorts have shown to be of valuable importance for the mapping of rare genetic diseases as well as for unravelling the genetics of common complex diseases (Vitart *et al.*, 2006).

The three studied cohorts are themselves isolated populations and the analysis not only of their genetic data but also of biochemical traits might on one side reveal characteristics specific to a population, and on the other side give reliable information about conserved associations in the general population.

Significant differences were obtained in all feature data sets with all of them being observed between Orkney and the other two populations. The IgG glycans set presented the highest number of differences compared with plasma glycans and phenotypes where only two traits in each data set achieved a statistical significance. The differences seen for IgG glycans could be an experimental artifact due to the fact that IgG glycans in Orkney were quantified using a different method than in Vis and Korčula. However, it was show in section 3.1 that the measurements of most IgG glycans correlated well between methods despite the differences in magnitude.

Despite their significance, these differences were not powerful enough to yield discrimination between populations when performing the PLS-DA and PCA analyses. Nonetheless, it should be noted that the importance of the traits for the principal components in the PCA analysis of each feature data set reproduced the findings observed in the clustering analysis (see sections 3.3 and 4.2). For instance, in the case of plasma glycans, GP9, DG5 and BADS show opposite contributions to DG6, Monosialo, BAMS and C.FUC, a tendency that was observed to occur across clusters. For IgG glycans, the features IGG3, IGG43 and IGG55 (referring to agalactosylated glycan structures) are grouped as they have similar contributions and they also presented different patterns across clusters. In the case of phenotype data, uric acid, waist and hip circumferences have the highest contributions and were also shown to differ between clusters. Altogether, these results suggest that a common phenotypic background exists between populations which is independent of the geographic location of the individuals and might be related with certain shared lifestyle habits.

The classification of populations by RF based on glycan and phenotype profiles yielded more satisfactory results. The best performance in the classification of populations with a relatively low error was achieved for a combined set of glycans and phenotypes indicating that the traits complement each other by introducing additional information about the populations. For all feature data sets, Orkney is always better separated from the other two populations which are consistently mistaken. Although considered isolated populations, the fact that Vis and Korčula are much closer geographically to each other than to Orkney means that they are under more similar biological pressures which might be the cause of the observed results.

The different results obtained with PLS-DA and PCA analyses and with RF are related with the distinct nature of the algorithms. While PLS-DA and PCA search for linear combinations of features that can explain the variability of the data, RF method is a more flexible and nonlinear approach. The fact that RF achieved better results than PLS-DA and PCA suggests that

population-specifc patterns of glycans and phenotypes might arise from nonlinear combinations of features.

Concerning the analysis of the genetic structure of the populations, the ensemble of SNPs was able to separate the populations as could be expected from the fact that the populations are isolated and hold their own genetic signatures. Moreover, Orkney could be clearly separated from the other two populations which partially overlap to a small extent reflecting the geographical proximity of Vis and Korčula relative to Orkney. Further investigation of the SNPs most contributing for the genetic structure of the populations yielded a set of 35 SNPs consistently detected by the three different SNP selection methods employed. These SNPs are mainly located in two regions: one on chromosome 2 comprising genes related to mRNA processing, protein biosynthesis and trafficking, and another on chromosome 6 comprehending genes involved in the immune system response, cell interactions and glycosylation-related processes. The population-specific characteristics arising from the differences in these SNPs and from possible functional alteration upon genes affected by them is yet to be elucidated and would require a throughout examination of patho-physiologic differences present between populations. Given the functions of the genes harbouring and flanking these SNPs, a plausible explanation would be that these SNPs might reflect a predisposition of a population for certain diseases or conditions.

4.6. Association between N-glycans and genotypes

Understanding the influence of the genomic background upon a trait or disease is of extreme importance to expand the knowledge about the pathways leading to those phenotypes and, consequently, help in the development of more accurate diagnostic tests and treatment solutions.

A first approach to elucidate possible relationship between the glycan and genotype profiles in two of the populations included in the present study was the subject of a previous work (Tica, 2011). In the study, genotypes were used to calculate estimates of pairwise identity-by-descent, a measure that is useful for detecting pairs of individuals who look more similar to each other than it would be expected by chance in a random sample. As such, the pairwise identity-by-descent estimations were taken as a measure of the distance between pairs of individuals and were subjected to hierarchical clustering. The clusters obtained based on genotypes were characterised regarding their glycan profile. Enrichment of certain glycan features was observed for some clusters suggesting the presence of a link between glycans and genotypes.

In the present thesis, the association between glycans and genotypes was approached with a similar rationale to that of GWAS but using different modeling methods. Three multivariate methods which considered the interaction among SNPs in their algorithms were applied to address the problem of SNP selection in the glycosylation context. The glycome-wide analysis carried out intended to assess whether these methods were responsive to this data as well as to attempt to unravel novel associations.

The principal associations previously reported for both plasma and IgG glycans were captured by the used methods (Huffman *et al.*, 2011; Lauc *et al.*, 2010a; Lauc *et al.*, 2013). Additional associations for plasma glycans with three glycosyltransferases are implied for DG7 and disialylated structures and with two genes from the solute carrier family are implied for GP13, G3 and disialylated structures (Table 8). For the disialylated structures, the variation rs9847446 is flanked by two members of the solute carrier family, namely SLC9C1 and SLC35A5. The solute carrier SLC9A9 has been reported before to be associatiated with tetrasialylated structures (Huffman *et al.*, 2011). In the case of IgG glycans, several traits were found to be associated with the variant rs6764279 in the ST6GAL1 gene and the majority of these associations was ranked in the first place by all methods which is strongly suggestive of its influence upon the traits in question (Supplementary table 9). A careful examination of the genetic context of SNPs which have not yet been reported to be linked with glycosylation should be carried out as they may contain additional information.

Table 8. Genetic variants implied to be associated with plasma glycan traits. List of SNPs consistently identified by the three methods to be associated with the presented glycan traits within the top 100 SNPs. Genes overlapping with the SNP are annotated without asterisk and neighbour genes of the SNP are annotated with asterisk. The variation number in parenthesis following the name of the method indicates the rank position achieved by the SNP with that particular method. RJ: Random Jungle; CARE: correlation adjusted scores; GEMMA: bayesian sparse linear mixed model.

Trait	SNP	Chr	Genes	Methods (rank)
GP13	rs13107325	4	SLC39A8	RJ(5); CARE(2); GEMMA(6)
DG7	rs315081	1	ST6GALNAC3	RJ(17); CARE(2); GEMMA(3)
DG7	rs4569731	4	GALNTL6	RJ(72); CARE(86); GEMMA(66)
Disialo	rs9847446	3	RP11-231E6.1*	RJ(3); CARE(7); GEMMA(3)
Disialo	rs759602	3	ST6GAL1	RJ(38); CARE(46); GEMMA(92)
G3	rs13107325	4	SLC39A8	RJ(6); CARE(1); GEMMA(1)

The association analysis between phenotypes and genotypes was also performed. Associations detected by all methods were achieved for systolic pressure, HDL, triglycerides, insulin, calcium and uric acid. While some of the SNPs are located in genomic regions that can be related with the corresponding trait, some apparently have no link.

The variation rs10507382, implied to be associated with systolic blood pressure, is located on chromosome 13 overlapping the Fms-Related Tyrosine Kinase 1 gene (FLT1). This gene encodes a protein member of the vascular endothelial growth factor receptor family and plays an important role in angiogenesis and vasculogenesis.

The variation rs159382, implied to be associated with triglycerides, is located in a region of chromosome 5 close to the Phosphodiesterase 4D, CAMP-Specific gene (PDE4D) whose mutations have been associated with the levels of serum triglyceride (Sinha *et al.*, 2013). Moreover, neighbouring SNPs are implicated in other GWAS studies analysing cholesterol and triglycerides (information retrieved from the NCBI dbSNP database website, see http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=159382).

The variation rs6679047, implied to be associated with insulin levels is located on chromosome 1 upstream the Nuclear Receptor Subfamily 5, Group A, Member 2 gene (NR5A2) which is involved in the pancreatic function.

The variation rs7914270, implied to be associated with uric acid, is harboured in the Solute Carrier Family 2 (Facilitated Glucose Transporter), Member 9 gene (SLC2A9 or GLUT9) located on chromosome 10. This gene has been reported before as a modulator of uric acid levels and the results were replicate in several populations (Le *et al.*, 2008; Vitart *et al.*, 2008; Zemunik *et al.*, 2009).

The small number of associations found for phenotypes are a consensus of those associations consistently identified by all methods and thus should not be regarded as a result of the poor performance of the methods. The individual results of each method might possibly reveal other findings as well as confirming previous ones.

The results achieved demonstrate the ability of the employed SNP selection methods to reproduce recent results of GWAS applied to glycosylation traits and to suggest other potential associations. Despite the large number of GWAS conducted nowadays and their success in revealing important genetic factors underlying human diseases and traits, GWAS still faces challenges not only at the level of the rationale behind the analysis but also at the computational level. Most GWAS approaches test one SNP at a time and overlook potential multiple causal variants by disregarding the interdependencies between SNPs which occur in complex diseases and traits. Additionally, genome wide studies are usually computationally demanding and the traditional methods are becoming obsolete with the increasing in size of the data sets available for such type of analysis. Recently developed polygenic modelling methods implement more efficient algorithms capable of analysing a large number of SNPs while simultaneously incorporating dependencies among SNPs. This increase in efficiency is reflected in less computationally exhaustive algorithms which have the advantage of a reduction in the computational time required to perform the analysis, thus contributing to gains in terms of speed, time and also knowledge.

Heritability represents the proportion of the phenotype variance that can be attributed to genetic factors and is a recurrent analysis in any genetic study. GEMMA algorithm provides two such measures: PVE which estimates the proportion of variance in the analysed traits explained by both small and large effect size SNPs and PGE which is the proportion of variance in the trait explained by the large effect size SNPs alone.

Estimates of PVE were up to around 45%, 60% and 50% for plasma glycan, IgG glycan and phenotype traits, respectively, while estimates of PGE were above 30% only for few traits in all cases. Overall, the fact that estimates of PVE were higher than PGE suggests that small

polygenic effects might have a stronger contribution for most of the analysed traits, whereas a limited number of traits are influenced by large effect size SNPs. However, these estimates of heritability should be interpreted with caution since the presence of unmeasured environmental factors that influence the phenotype and are correlated with genotype can affect the estimates (Zhou *et al.*, 2013). Particular care should therefore be taken in the present case where the obtained estimates are based on a pooled data set of three populations subjected to different environments which can compromise the results. Further analysis considering each population separately should be performed in order to verify whether the different relative contributions of small and large effect size SNPs to glycan traits and phenotypes are conserved across populations or indeed vary according to the population.

5. REFERENCES

Adamczyk, B., Tharmalingam, T. & Rudd, P.M. 2012. Glycans as cancer biomarkers. Biochim. Biophys. Acta, 1820:1347-53.

Andreopoulos, B., An, A., Wang, X., et al. 2009. A roadmap of clustering algorithms: finding a match for a biomedical application. Brief Bioinform, 10:297-314.

Anthony, R.M. & Ravetch, J.V. 2010. A novel role for the IgG Fc glycan: the anti-inflammatory activity of sialylated IgG Fcs. J. Clin. Immunol., 30 Suppl 1:S9-14.

Anthony, R.M., Wermeling, F. & Ravetch, J.V. 2012. Novel roles for the IgG Fc glycan. Ann. N. Y. Acad. Sci., 1253:170-80.

Aoki-Kinoshita, K.F. 2008. An introduction to bioinformatics for glycomics research. PLoS Comput Biol, 4:e1000075.

Balzarini, J. 2007. Targeting the glycans of glycoproteins: a novel paradigm for antiviral therapy. Nat Rev Microbiol, 5:583-97.

Bates, D., Maechler, M., Bolker, B., et al. (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-4. Available from *http://CRAN.R-project.org/package=lme4*.

Bland, J.M. & Altman, D.G. 1999. Measuring agreement in method comparison studies. Stat. Methods Med. Res., 8:135-60.

Bland, J.M. & Altman, D.G. 2003. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet. Gynecol., 22:85-93.

BMES & WTCCC2 2013. Genome-wide association study of intraocular pressure identifies the GLCCI1/ICA1 region as a glaucoma susceptibility locus. Hum. Mol. Genet., 22:4653-60.

Bodenhofer, U., Kothmeier, A. & Hochreiter, S. 2011. APCluster: an R package for affinity propagation clustering. Bioinformatics, 27:2463-4.

Breiman, L. 2001. Random Forests. Mach Learn, 45:5-32.

Brucklacher, R.M., Patel, K.M., VanGuilder, H.D., et al. 2008. Whole genome assessment of the retinal response to diabetes reveals a progressive neurovascular inflammatory response. BMC medical genomics, 1:26.

Bunn, H.F. 1981. Nonenzymatic glycosylation of protein: relevance to diabetes. Am. J. Med., 70:325-30.

Bush, W.S. & Moore, J.H. 2012. Chapter 11: Genome-wide association studies. PLoS Comput Biol, 8:e1002822.

Campbell, M.P., Royle, L., Radcliffe, C.M., et al. 2008. GlycoBase and autoGU: tools for HPLC-based glycan analysis. Bioinformatics, 24:1214-6.

Chakrabarti, S., Cox, E., Frank, E., et al. 2009. Data Mining: Know It All. In: editors. Morgan Kaufmann Publishers Inc. USA. Available from *http://books.google.hr/books?id=WRqZ0QsdxKkC&dq=Data+Mining+Know+It+All&source=gbs_navlinks_s*.

Ciric, D., Milosevic-Jovcic, N., Ilic, V., et al. 2005. A longitudinal study of the relationship between galactosylation degree of IgG and rheumatoid factor titer and avidity during long-term immunization of rabbits with BSA. Autoimmunity, 38:409-16.

Cook, N.S. & Ubben, D. 1990. Fibrinogen as a major risk factor in cardiovascular disease. Trends Pharmacol. Sci., 11:444-51.

Cooper, G.M. & Shendure, J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet, 12:628-40.

Cordell, H.J. 2009. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet, 10:392-404.

Cummings, R.D. 2009. The repertoire of glycan determinants in the human glycome. Mol Biosyst, 5:1087-104.

Dallas, D.C., Martin, W.F., Hua, S., et al. 2013. Automated glycopeptide analysis-review of current state and future directions. Brief Bioinform, 14:361-374.

Danesh, J., Lewington, S., Thompson, S.G., et al. 2005. Plasma fibrinogen level and the risk of major cardiovascular diseases and nonvascular mortality: an individual participant meta-analysis. JAMA, 294:1799-809.

Davalos, D. & Akassoglou, K. 2012. Fibrinogen as a key regulator of inflammation in disease. Seminars in immunopathology, 34:43-62.

Dell, A., Galadari, A., Sastre, F., et al. 2010. Similarities and differences in the glycosylation mechanisms in prokaryotes and eukaryotes. International journal of microbiology, 2010:148178.

Demacker, P.N., Schade, R.W., Jansen, R.T., et al. 1982. Intra-individual variation of serum cholesterol, triglycerides and high density lipoprotein cholesterol in normal humans. Atherosclerosis, 45:259-66.

Diamantstein, T. & Odenwald, M.V. 1974. Control of the immune response in vitro by calcium ions. I. The antagonistic action of calcium ions on cell proliferation and on cell differentiation. Immunology, 27:531-41.

Ding, N., Nie, H., Sun, X., et al. 2011. Human serum N-glycan profiles are age and sex dependent. Age Ageing, 40:568-75.

Fiedler, K. & Simons, K. 1995. The role of N-glycans in the secretory pathway. Cell, 81:309-12.

Freeze, H.H. 2006. Genetic defects in the human glycome. Nat Rev Genet, 7:537-51.

Frey, B.J. & Dueck, D. 2007. Clustering by passing messages between data points. Science, 315:972-6.

Gamblin, D.P., Scanlan, E.M. & Davis, B.G. 2009. Glycoprotein synthesis: an update. Chem. Rev., 109:131-63.

Genome-wide Efficient Mixed Model Association algorithm (2013). Available from *http://home.uchicago.edu/xz7/software.html*.

Golay, A. & Ybarra, J. 2005. Link between obesity and type 2 diabetes. Best practice & research. Clinical endocrinology & metabolism, 19:649-63.

Gornik, I., Maravic, G., Dumic, J., et al. 1999. Fucosylation of IgG heavy chains is increased in rheumatoid arthritis. Clin. Biochem., 32:605-8.

Gornik, O., Pavic, T. & Lauc, G. 2012. Alternative glycosylation modulates function of IgG and other proteins - Implications on evolution and disease. Biochim. Biophys. Acta, 1820:1318-26.

Gornik, O., Wagner, J., Pucic, M., et al. 2009. Stability of N-glycan profiles in human plasma. Glycobiology, 19:1547-53.

Gromada, J., Bark, C., Smidt, K., et al. 2005. Neuronal calcium sensor-1 potentiates glucose-dependent exocytosis in pancreatic beta cells through activation of phosphatidylinositol 4-kinase beta. Proc. Natl. Acad. Sci. U. S. A., 102:10303-8.

Gu, J. & Taniguchi, N. 2008. Potential of N-glycan in cell adhesion and migration as either a positive or negative regulator. Cell adhesion & migration, 2:243-5.

Hayes, C.A., Nemes, S. & Karlsson, N.G. 2012. Statistical analysis of glycosylation profiles to compare tissue type and inflammatory disease state. Bioinformatics, 28:1669-76.

Helenius, A. & Aebi, M. 2001. Intracellular functions of N-linked glycans. Science, 291:2364-9.

Helenius, A. & Aebi, M. 2004. Roles of N-linked glycans in the endoplasmic reticulum. Annu. Rev. Biochem., 73:1019-49.

Higai, K., Azuma, Y., Aoki, Y., et al. 2003. Altered glycosylation of alpha1-acid glycoprotein in patients with inflammation and diabetes mellitus. Clin. Chim. Acta, 329:117-25.

Hirschhorn, J.N. & Daly, M.J. 2005. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet, 6:95-108.

Hounsell, E.F. & Davies, M.J. 1993. Role of protein glycosylation in immune regulation. Ann. Rheum. Dis., 52 Suppl 1:S22-9.

Huffman, J.E., Knezevic, A., Vitart, V., et al. 2011. Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults. Hum. Mol. Genet., 20:5000-11.

Huhn, C., Selman, M.H., Ruhaak, L.R., et al. 2009. IgG glycosylation analysis. Proteomics, 9:882-913.

Igl, W., Johansson, A. & Gyllensten, U. 2010. The Northern Swedish Population Health Study (NSPHS)--a paradigmatic study in a rural population combining community health and basic research. Rural Remote Health, 10:1363.

Igl, W., Polasek, O., Gornik, O., et al. 2011. Glycomics meets lipidomics--associations of N-glycans with classical lipids, glycerophospholipids, and sphingolipids in three European populations. Mol Biosyst, 7:1852-62.

Itoh, N., Sakaue, S., Nakagawa, H., et al. 2007. Analysis of N-glycan in serum glycoproteins from db/db mice and humans with type 2 diabetes. American journal of physiology. Endocrinology and metabolism, 293:E1069-77.

Jiang, R., Tang, W., Wu, X., et al. 2009. A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics, 10 Suppl 1:S65.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics, 24:1403-5.

Jombart, T. & Ahmed, I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics, 27:3070-1.

Jombart, T., Devillard, S. & Balloux, F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet., 11:

Kahn, S.E., Hull, R.L. & Utzschneider, K.M. 2006. Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature, 444:840-6.

Kaneko, Y., Nimmerjahn, F. & Ravetch, J.V. 2006. Anti-inflammatory activity of immunoglobulin G resulting from Fc sialylation. Science, 313:670-3.

Klarić, L. 2012. Clustering and profile analysis of human plasma and IgG-associated glycans. Graduation thesis, University of Zagreb.

Kleene, R. & Schachner, M. 2004. Glycans and neural cell interactions. Nature reviews. Neuroscience, 5:195-208.

Knezevic, A., Gornik, O., Polasek, O., et al. 2010. Effects of aging, body mass index, plasma lipid profiles, and smoking on human plasma N-glycans. Glycobiology, 20:959-69.

Knezevic, A., Polasek, O., Gornik, O., et al. 2009. Variability, heritability and environmental determinants of human plasma N-glycome. J Proteome Res, 8:694-701.

Kung, L.A., Tao, S.C., Qian, J., et al. 2009. Global analysis of the glycoproteome in Saccharomyces cerevisiae reveals new roles for protein glycosylation in eukaryotes. Mol Syst Biol, 5:308.

Lauc, G., Essafi, A., Huffman, J.E., et al. 2010a. Genomics meets glycomics-the first GWAS study of human N-Glycome identifies HNF1alpha as a master regulator of plasma protein fucosylation. PLoS Genet, 6:e1001256.

Lauc, G., Huffman, J.E., Pucic, M., et al. 2013. Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. PLoS Genet, 9:e1003225.

Lauc, G., Rudan, I., Campbell, H., et al. 2010b. Complex genetic regulation of protein glycosylation. Mol Biosyst, 6:329-35.

Lauc, G. & Zoldos, V. 2010. Protein glycosylation--an evolutionary crossroad between genes and environment. Mol Biosyst, 6:2373-9.

Le Cao, K.A., Gonzalez, I. & Dejean, S. 2009. integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics, 25:2855-6.

Le, M.T., Shafiu, M., Mu, W., et al. 2008. SLC2A9--a fructose transporter identified as a novel uric acid transporter. Nephrol. Dial. Transplant., 23:2746-9.

Lee, R.T., Lauc, G. & Lee, Y.C. 2005. Glycoproteomics: protein modifications for versatile functions. Meeting on glycoproteomics. EMBO Rep, 6:1018-22.

Li, J., Horstman, B. & Chen, Y. 2011. Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. Bioinformatics, 27:i222-9.

Liaw, A. & Wiener, M. 2002. Classification and Regression by randomForest. R News, 2:18-22.

Lowe, J.B. & Marth, J.D. 2003. A genetic approach to Mammalian glycan function. Annu. Rev. Biochem., 72:643-91.

Lux, A., Aschermann, S., Biburger, M., et al. 2010. The pro and anti-inflammatory activities of immunoglobulin G. Ann. Rheum. Dis., 69 Suppl 1:i92-96.

Maitra, S. & Yan, J. 2008 Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression Casualty Actuarial Society, 79-90.

Marino, K., Bones, J., Kattla, J.J., et al. 2010. A systematic approach to protein glycosylation analysis: a path through the maze. Nat. Chem. Biol., 6:713-723.

Marquardt, T. & Denecke, J. 2003. Congenital disorders of glycosylation: review of their molecular bases, clinical presentations and specific therapies. Eur. J. Pediatr., 162:359-79.

Marquardt, T. & Freeze, H. 2001. Congenital disorders of glycosylation: glycosylation defects in man and biological models for their study. Biol. Chem., 382:161-77.

McMillan, D.E. 1972. Elevation of glycoprotein fucose in diabetes mellitus. Diabetes, 21:863-71.

Merriman, T.R., Cordell, H.J., Eaves, I.A., et al. 2001. Suggestive evidence for association of human chromosome 18q12-q21 and its orthologue on rat and mouse chromosome 18 with several autoimmune diseases. Diabetes, 50:184-94.

Milting, H., Lukas, N., Klauke, B., et al. 2006. Composite polymorphisms in the ryanodine receptor 2 gene associated with arrhythmogenic right ventricular cardiomyopathy. Cardiovasc. Res., 71:496-505.

Mittag, F., Buchel, F., Saad, M., et al. 2012. Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. Hum. Mutat., 33:1708-18.

Monti, S., Tamayo, P., Mesirov, J., et al. 2003. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn, 52:91-118.

Moore, J.H., Asselbergs, F.W. & Williams, S.M. 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics, 26:445-55.

Moremen, K.W., Tiemeyer, M. & Nairn, A.V. 2012. Vertebrate protein glycosylation: diversity, synthesis and function. Nature reviews. Molecular cell biology, 13:448-62.

Mukhopadhyay, A., Kramer, J.M., Merkx, G., et al. 2010. CDK19 is disrupted in a female patient with bilateral congenital retinal folds, microcephaly and mild mental retardation. Hum. Genet., 128:281-91.

Mwangi, S.M. & Srinivasan, S. 2010. DCAMKL-1: a new horizon for pancreatic progenitor identification. American journal of physiology. Gastrointestinal and liver physiology, 299:G301-2.

Narimatsu, H. 2006. Human glycogene cloning: focus on beta 3-glycosyltransferase and beta 4-glycosyltransferase families. Curr. Opin. Struct. Biol., 16:567-75.

Navratilova, Z. 2006. Polymorphisms in CCL2&CCL5 chemokines/chemokine receptors genes and their association with diseases. Biomedical papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia, 150:191-204.

Ohtsubo, K. & Marth, J.D. 2006. Glycosylation in cellular mechanisms of health and disease. Cell, 126:855-67.

Papadias, D., Qiongmao, S., Yufei, T., et al. 2004 Group nearest neighbor queries Data Engineering, 2004. Proceedings. 20th International Conference on, 301-312.

The Perl Programming Language (2013). Available from *http://www.perl.org/*.

Procaccini, C., Carbone, F., Galgani, M., et al. 2011. Obesity and susceptibility to autoimmune diseases. Expert review of clinical immunology, 7:287-94.

Pucic, M., Knezevic, A., Vidic, J., et al. 2011. High throughput isolation and glycosylation analysis of IgG-variability and heritability of the IgG glycome in three isolated human populations. Mol Cell Proteomics, 10:M111 010090.

Pucic, M., Muzinic, A., Novokmet, M., et al. 2012. Changes in plasma and IgG N-glycome during childhood and adolescence. Glycobiology, 22:975-82.

Pucic, M., Pinto, S., Novokmet, M., et al. 2010. Common aberrations from the normal human plasma N-glycan profile. Glycobiology, 20:970-5.

Purcell, S., Neale, B., Todd-Brown, K., et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 81:559-75.

Queitsch, C., Carlson, K.D. & Girirajan, S. 2012. Lessons from model organisms: phenotypic robustness and missing heritability in complex disease. PLoS Genet, 8:e1003041.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available from *http://www.R-project.org/*.

Rambaruth, N.D. & Dwek, M.V. 2011. Cell surface glycan-lectin interactions in tumor metastasis. Acta Histochem., 113:591-600.

Random Jungle (2013). Available from *http://imbs-luebeck.de/imbs/de/node/227*.

Rasmussen, M., Guo, X., Wang, Y., et al. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science, 334:94-8.

Rotival, M., Zeller, T., Wild, P.S., et al. 2011. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet, 7:e1002367.

Royle, L., Campbell, M.P., Radcliffe, C.M., et al. 2008. HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. Anal. Biochem., 376:1-12.

Rudan, I., Biloglav, Z., Vorko-Jovic, A., et al. 2006. Effects of inbreeding, endogamy, genetic admixture, and outbreeding on human health: a (1001 Dalmatians) study. Croat. Med. J., 47:601-10.

Rudan, I., Campbell, H. & Rudan, P. 1999. Genetic epidemiological studies of eastern Adriatic Island isolates, Croatia: objective and strategies. Coll. Antropol., 23:531-46.

Rudan, I., Marusic, A., Jankovic, S., et al. 2009. "10001 Dalmatians:" Croatia launches its national biobank. Croat. Med. J., 50:4-6.

Saldova, R., Huffman, J.E., Adamczyk, B., et al. 2012. Association of medication with the human plasma N-glycome. J Proteome Res, 11:1821-31.

Schwarz, D.F., Konig, I.R. & Ziegler, A. 2010. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics, 26:1752-8.

Schwarz, F. & Aebi, M. 2011. Mechanisms and principles of N-linked protein glycosylation. Curr. Opin. Struct. Biol., 21:576-82.

Shi, T. & Horvath, S. 2006. Unsupervised Learning with Random Forest Predictors. Journal of Computational and Graphical Statistics, 15:118-138.

Sinha, E., Meitei, S.Y., Garg, P.R., et al. 2013. PDE4D gene polymorphisms and coronary heart disease: a case-control study in a north Indian population. J. Clin. Lab. Anal., 27:297-300.

Snider, J. (2013). Protein Glycosylation. Thermo Scientific Pierce Protein Biology Products, Rockford, IL Campus. Available from *http://www.piercenet.com/method/protein-glycosylation*.

Sparks, S.E. & Krasnewich, D.M. 2005. Congenital Disorders of Glycosylation Overview. In: Pagon RA, A.M., Bird TD, et al., editors. GeneReviews™ [Internet]. Seattle, WA. Available from *http://www.ncbi.nlm.nih.gov/books/NBK1332/*.

Stec, J.J., Silbershatz, H., Tofler, G.H., et al. 2000. Association of fibrinogen with cardiovascular risk factors and cardiovascular disease in the Framingham Offspring Population. Circulation, 102:1634-8.

Steinman, L., Conlon, P., Maki, R., et al. 2003. The intricate interplay among body weight, stress, and the immune response to friend or foe. J. Clin. Invest., 111:183-5.

The Structure, Function and Importance of Carbohydrates (2013). Available from *https://www.neb.com/sitecore/content/nebsg/home/tools-and-resources/feature-articles/the-structure-function-and-importance-of-carbohydrates*.

Stumpo, K.A. & Reinhold, V.N. 2010. The N-glycome of human plasma. J Proteome Res, 9:4823-30.

Svetnik, V., Liaw, A., Tong, C., et al. 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci., 43:1947-58.

Takahashi, M., Kuroki, Y., Ohtsubo, K., et al. 2009. Core fucose and bisecting GlcNAc, the direct modifiers of the N-glycan core: their functions and target proteins. Carbohydr. Res., 344:1387-90.

Takahashi, M., Tsuda, T., Ikeda, Y., et al. 2004. Role of N-glycans in growth factor signaling. Glycoconj. J., 20:207-12.

Thanabalasingham, G., Huffman, J.E., Kattla, J.J., et al. 2013. Mutations in HNF1A result in marked alterations of plasma glycan profile. Diabetes, 62:1329-37.

Thobhani, S., Yuen, C.T., Bailey, M.J., et al. 2009. Identification and quantification of N-linked oligosaccharides released from glycoproteins: an inter-laboratory study. Glycobiology, 19:201-11.

Thotakura, N.R. & Blithe, D.L. 1995. Glycoprotein hormones: glycobiology of gonadotrophins, thyrotrophin and free alpha subunit. Glycobiology, 5:3-10.

Tica, J. 2011. Computational analysis of plasma glycome and genotypes in human populations. Graduation thesis, University of Zagreb.

Tobias, R.D. 1995 An Introduction to Partial Least Squares Regression SUGI Proceedings,

Trombetta, E.S. 2003. The contribution of N-glycans and their processing in the endoplasmic reticulum to glycoprotein biosynthesis. Glycobiology, 13:77R-91R.

Van den Steen, P., Rudd, P.M., Dwek, R.A., et al. 1998. Concepts and principles of O-linked glycosylation. Crit. Rev. Biochem. Mol. Biol., 33:151-208.

Van Kirk, C.A., VanGuilder, H.D., Young, M., et al. 2011. Age-related alterations in retinal neurovascular and inflammatory transcripts. Mol. Vis., 17:1261-74.

Vanhooren, V., Dewaele, S., Libert, C., et al. 2010. Serum N-glycan profile shift during human ageing. Exp. Gerontol., 45:738-43.

Vanhooren, V., Laroy, W., Libert, C., et al. 2008. N-glycan profiling in the study of human aging. Biogerontology, 9:351-6.

Varki, A., Cummings, R.D., Esko, J.D., et al. 2009. Essentials of glycobiology. In: Varki A., C.R.D., Esko J.D., et al., editors. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, N.Y. Available from *http://www.ncbi.nlm.nih.gov/books/NBK1908/*.

Vitart, V., Biloglav, Z., Hayward, C., et al. 2006. 3000 years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia. Eur. J. Hum. Genet., 14:478-87.

Vitart, V., Rudan, I., Hayward, C., et al. 2008. SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. Nat. Genet., 40:437-42.

Vlassara, H., Brownlee, M. & Cerami, A. 1986. Nonenzymatic glycosylation: role in the pathogenesis of diabetic complications. Clin. Chem., 32:B37-41.

Walzel, H., Fahmi, A.A., Eldesouky, M.A., et al. 2006. Effects of N-glycan processing inhibitors on signaling events and induction of apoptosis in galectin-1-stimulated Jurkat T lymphocytes. Glycobiology, 16:1262-71.

Wei, X. & Li, L. 2009. Comparative glycoproteomics: approaches and applications. Brief. Funct. Genomic. Proteomic., 8:104-13.

Winham, S.J., Colby, C.L., Freimuth, R.R., et al. 2012. SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics, 13:164.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. Roy. Stat. Soc. Ser. B. (Stat. Method.), 73:3-36.

Wuhrer, M., Stam, J.C., van de Geijn, F.E., et al. 2007. Glycosylation profiling of immunoglobulin G (IgG) subclasses from human serum. Proteomics, 7:4070-81.

Yang, B., Houlberg, K., Millward, A., et al. 2004. Polymorphisms of chemokine and chemokine receptor genes in Type 1 diabetes mellitus and its complications. Cytokine, 26:114-21.

Yao, L., Zhong, W., Zhang, Z., et al. 2009. Classification tree for detection of single-nucleotide polymorphism (SNP)-by-SNP interactions related to heart disease: Framingham Heart Study. BMC proceedings, 3 Suppl 7:S83.

Yaras, N., Ugur, M., Ozdemir, S., et al. 2005. Effects of diabetes on ryanodine receptor Ca release channel (RyR2) and Ca2+ homeostasis in rat heart. Diabetes, 54:3082-8.

Yoshida, Y. 2003. A novel role for N-glycans in the ERAD system. J Biochem, 134:183-90.

Zemunik, T., Boban, M., Lauc, G., et al. 2009. Genome-wide association study of biochemical traits in Korcula Island, Croatia. Croat. Med. J., 50:23-33.

Zhao, Y.Y., Takahashi, M., Gu, J.G., et al. 2008. Functional roles of N-glycans in cell signaling and cell adhesion in cancer. Cancer Sci, 99:1304-10.

Zhou, X., Carbonetto, P. & Stephens, M. 2013. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet, 9:e1003264.

Zhou, X. & Stephens, M. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet., 44:821-4.

Zuber, V., Duarte Silva, A.P. & Strimmer, K. 2012. A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies. BMC Bioinformatics, 13:284.

Zuber, V. & Strimmer, K. 2009. Gene ranking and biomarker discovery under correlation. Bioinformatics, 25:2700-7.

Zuber, V. & Strimmer, K. 2011. High-Dimensional Regression and Variable Selection Using CAR Scores. Stat Appl Genet Mol, 10:

CURRICULUM VITAE

Ana Sofia Pedrosa Pinto was born in Coimbra, Portugal, in 1984.

She graduated from the University of Coimbra, Portugal, in Biomedical Engineering in 2007 with the thesis entitled *Home Monitoring for Obstructive Sleep Apnea Diagnosis in Children*.

She has been a researcher in the Bioinformatics Group of the University of Zagreb since 2007. She participated in a research project entitled *Computational Analysis of Exonic Splicing Regulators* in the Bioinformatics Group of the University of Zagreb. She was also involved in a project concerning the implementation of a database and the development of a web server which resulted in a published paper. While in this group she was a teaching assistant of Algorithms and Programming Course in the University of Zagreb from 2009 to 2011.

She enrolled the Interdisciplinary PhD Program in Biophysics from the University of Split, Croatia, in October 2008.

She participated in several conferences and workshops with posters and oral presentations regarding the thesis research topic as well as the first research project.

She is author and co-author of two publications:

Pinto, S., Vlahoviček, K. and Buratti, E. (2011) PRO-MINE: A Bioinformatics Repository and Analytical Tool for TARDBP Mutations. Human Mutation, 32: E1948–E1958.

Pučić, M., **Pinto, S.**, Novokmet, M., Knežević, A., Gornik, O., Polašek, O., Vlahoviček, K., Wei, W., Rudd, P. M., Wright, A. F., Campbell, H., Rudan, I., and Lauc, G. (2010) Common aberrations from normal human N-glycan plasma profile. Glycobiology 20:970-975.

SAŽETAK

Glikozilacija je jedna od najopsežnijih modifikacija proteina. Glikani utječu na strukturu i funkciju proteina na koje su vezani, a poznato je i da imaju važne uloge u fiziološkim i patološkim procesima. Sinteza glikana ne odvija se prema kalupu, kao što je to slučaj kod proteina, nego u njoj sudjeluje kompleksna mreža interakcija stotina različitih enzima i transkripcijskih faktora. Nedostatak univerzalnog koda za sintezu glikana zajedno sa tehnološkim poteškoćama kvantifikacije glikana razlozi su ograničenom razumijevanju procesa koji reguliraju njihovu sintezu. Značajni napretci u analitičkim postupcima omogućili su razvoj pouzdanih visoko-protočnih metoda za kvantifikaciju glikana, a time i prve studije plazmatskog N-glikoma velikog broja ljudi, što je glikomiku postavilo uz bok ostalim visokoprotočnim metodama. Ove cjelovite studije otkrile su različite oblike povezanosti genetske predispozicije i okolišnih faktora u glikozilaciji proteina.

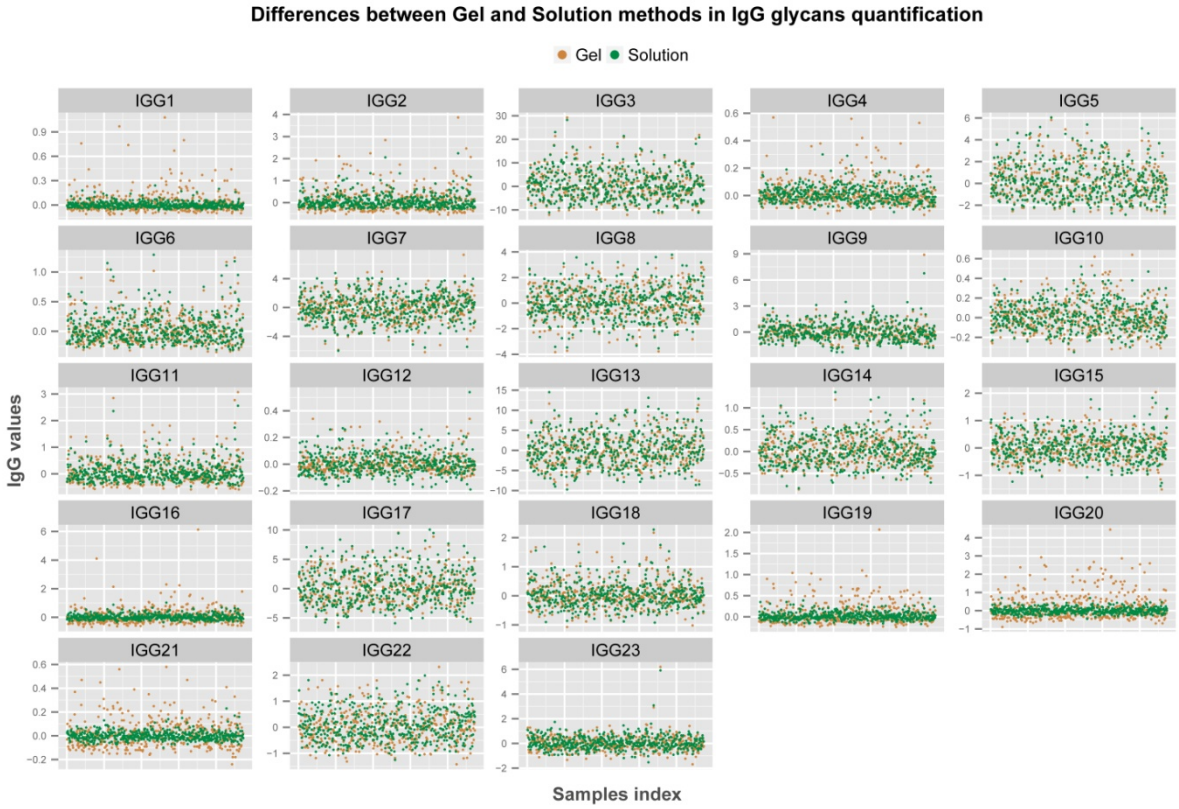
Kako bi se istražila genomska i i okolišna regulacija glikozilacije, u ovom su radu glikanski, fiziološki i biokemijski podaci, uz genotipove tri različite izolirane ljudske populacije analizirani serijom računalnih metodam. Također, predložen je općeniti obrazac za pripremu i obradu glikomskih podataka za daljnje analize. U općoj populaciji su identificirani specifični glikanski profili potencijalno povezani sa određenim patologijama i evaluiran je potencijal glikana kao biomarkera dijabetesa. Analizom unutarnjih struktura populacija pronađene su skupine čiji su profili slični među različitim populacijama. Osim toga, unatoč geografskoj i okolišnoj razdvojenosti populacija, otkriveno je nekoliko obrazaca povezanosti glikana i fenotipskih značajki koji se pojavljuju u svim populacijama. Genski polimorfizmi koji utječu na glikozilaciju su analizirani metodama više varijabli, zasnovanih na poligenskom modeliranju. Potvrda prethodnih otkrića i pronalazak novih potencijalnih poveznica sugeriraju da bi ove metode mogle postati alternativa tradicionalnim cjelogenomskim studijama zasnovanim na jednoj istraživanoj varijabli.

ABSTRACT

Glycosylation is one of the most extensive protein modifications. Glycans influence both structure and function of the proteins and they are known to have important roles in physiological and pathological processes. Glycan synthesis is not template driven but encoded within a complex network involving the interaction of hundreds of enzymes and transcriptional factors. The inexistence of a universal glycan structure code and technological restrictions in glycan quantification analysis have hindered the knowledge about the processes involved in the regulation of glycan assembly. Major breakthroughs in analytical procedures have allowed the quantification of glycans in a high-throughput manner and motivated the first large-scale studies on human plasma N-glycome which put glycomics on the same par as other *omics* approaches. These first comprehensive studies reported a diverse contribution of genetic background and environmental factors to glycosylation.

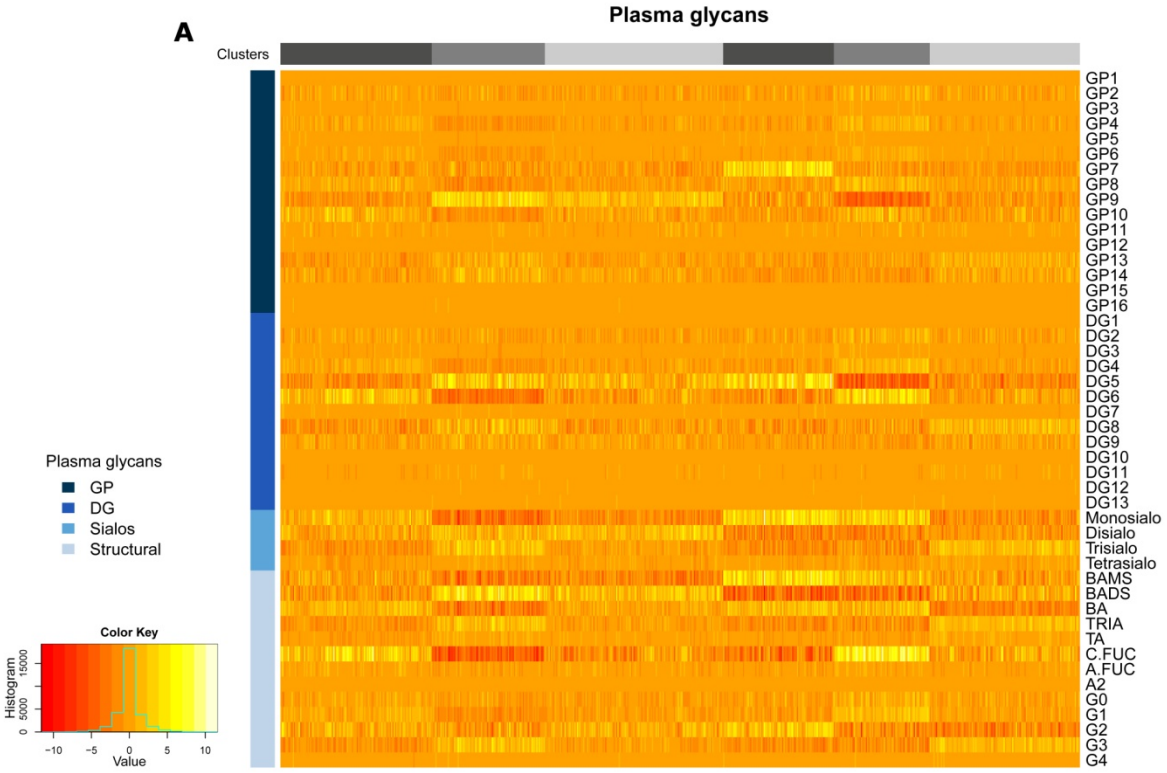
In order to explore the genomic and environmental regulation of glycosylation, different computational methods were employed to the integrated analysis of glycan, physiological/biochemical and genotype data in three isolated population cohorts. A general data processing pipeline to treat and pre-process glyco-related data prior to analysis was established. Specific glyco-phenotypes possibly related to pathologies were identified in the general population and the potential use of glycan modifications as biomarkers was evaluated for the particular case of diabetes. The analysis of the internal structure of populations revealed the presence of cluster profiles similar between populations. Additionally, several patterns of associations between glycans and phenotypes were shared across populations despite their geographical and environmental separation. Multivariate methods based on a polygenic modelling were used to investigate genetic polymorphisms affecting glycosylation. Confirmation of previous findings and the identification of possible novel links suggest that these efficient methods could provide an alternative to traditional univariate genome-wide association studies.

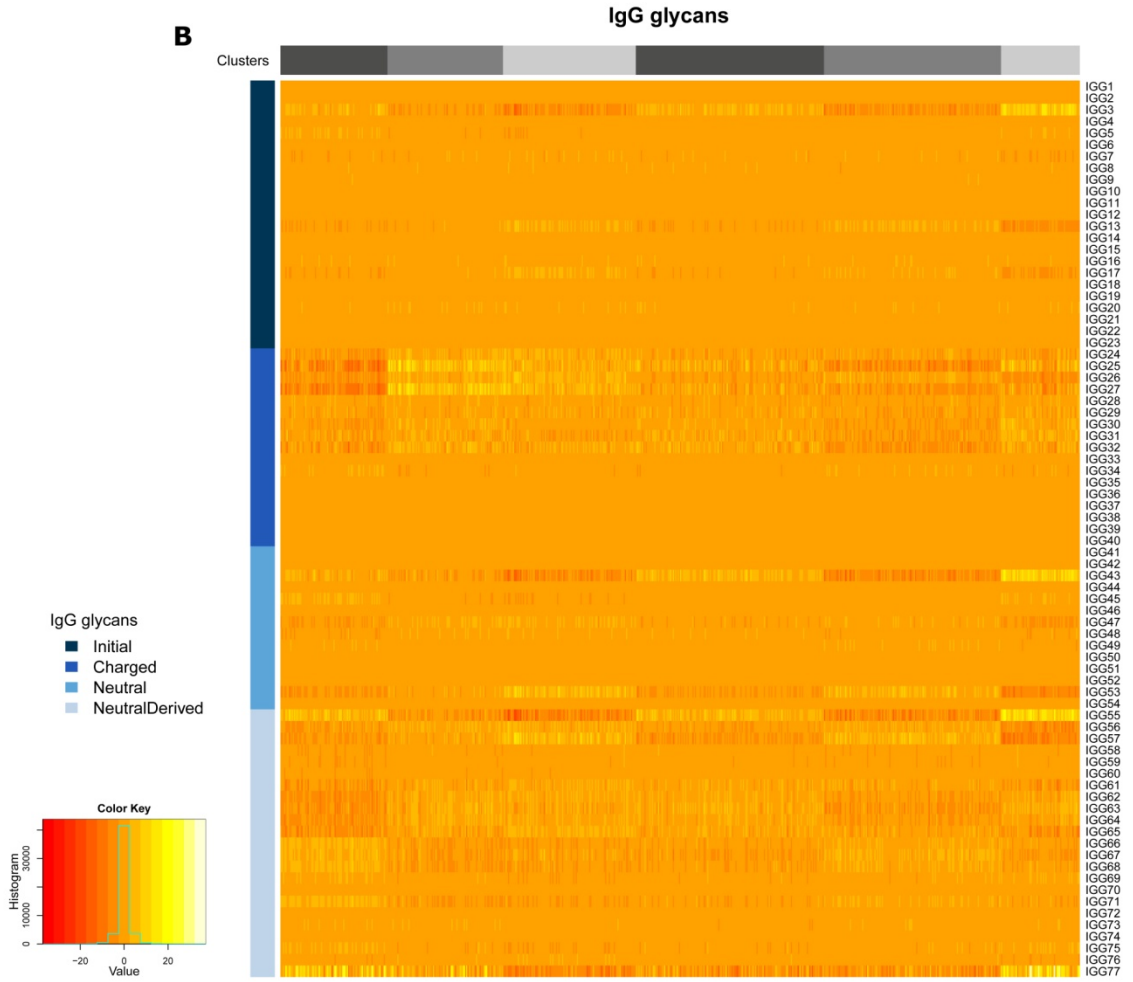
APPENDIX A. Supplementary Figures

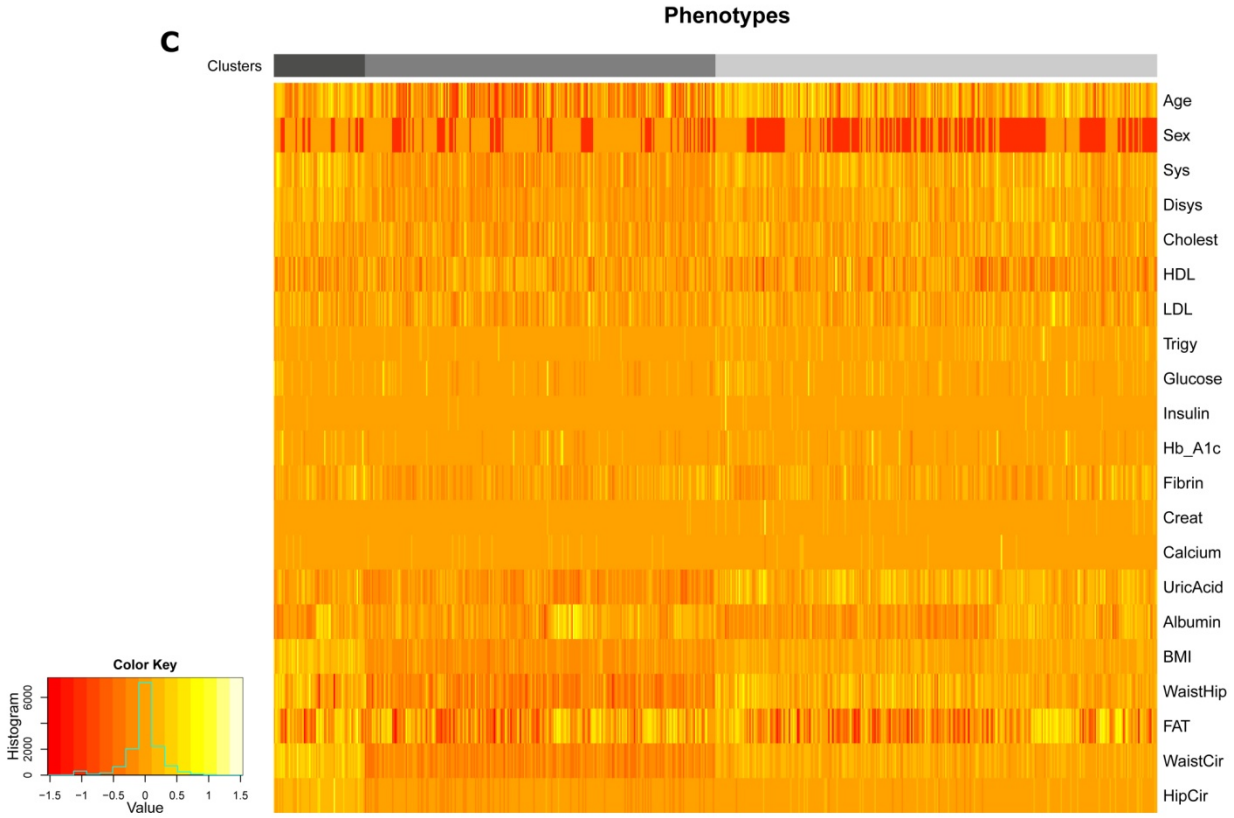


Supplementary figure 1. Differences between gel and solution methods for IgG glycan quantification. Scatterplots showing the IgG values for each sample as measured with the gel (golden points) and the solution methods (green points).

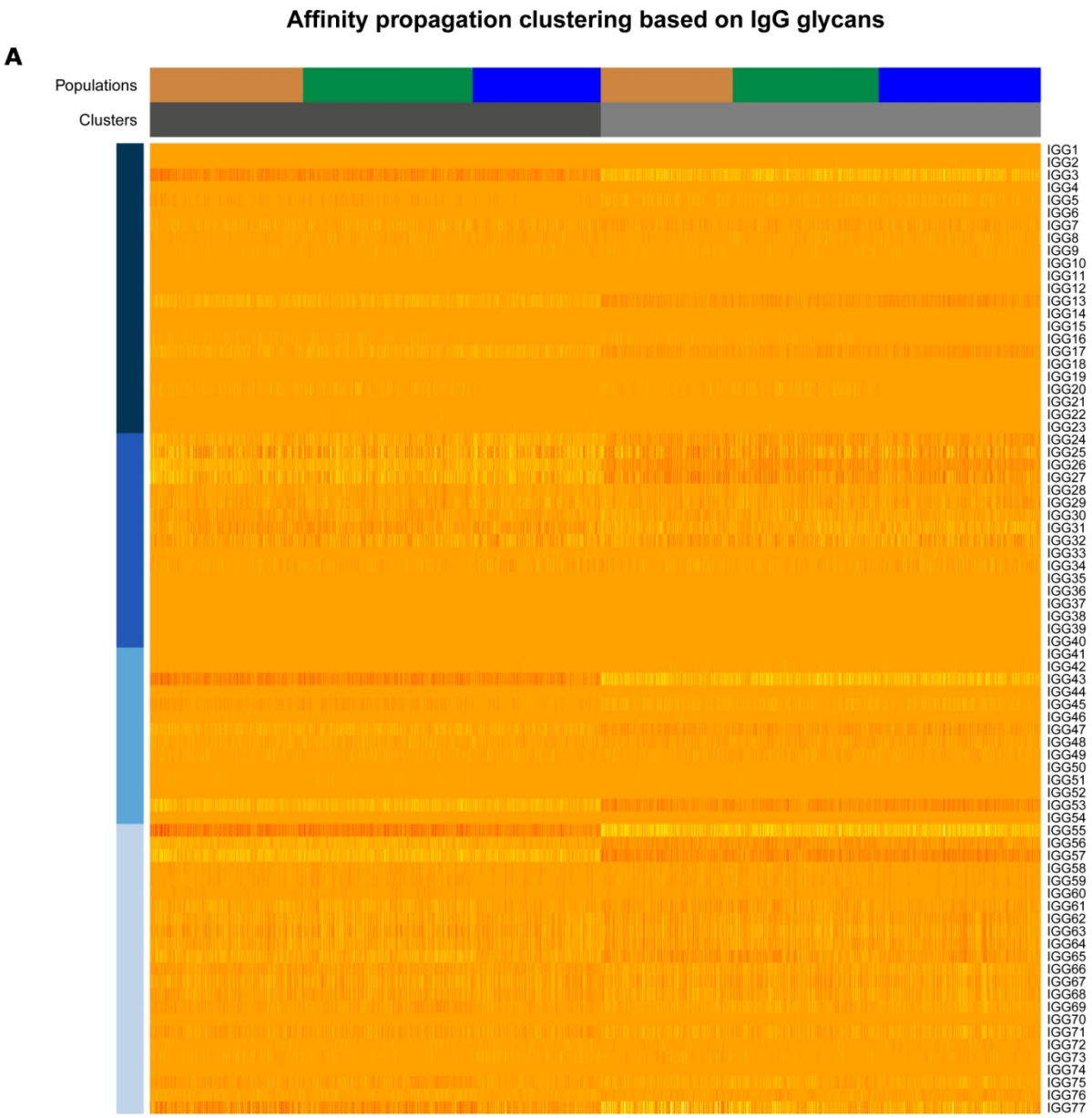
Affinity propagation clustering for Vis cohort

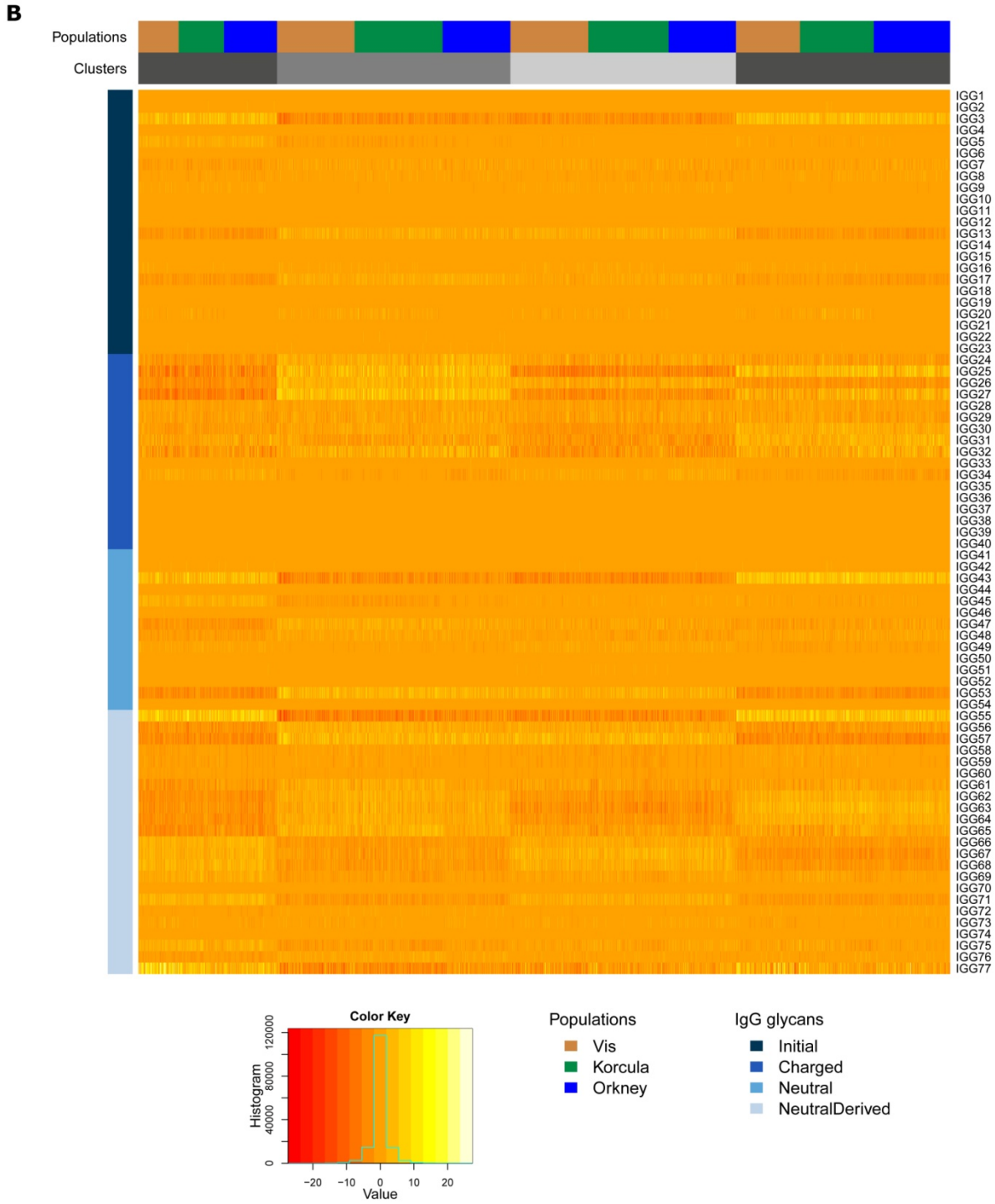






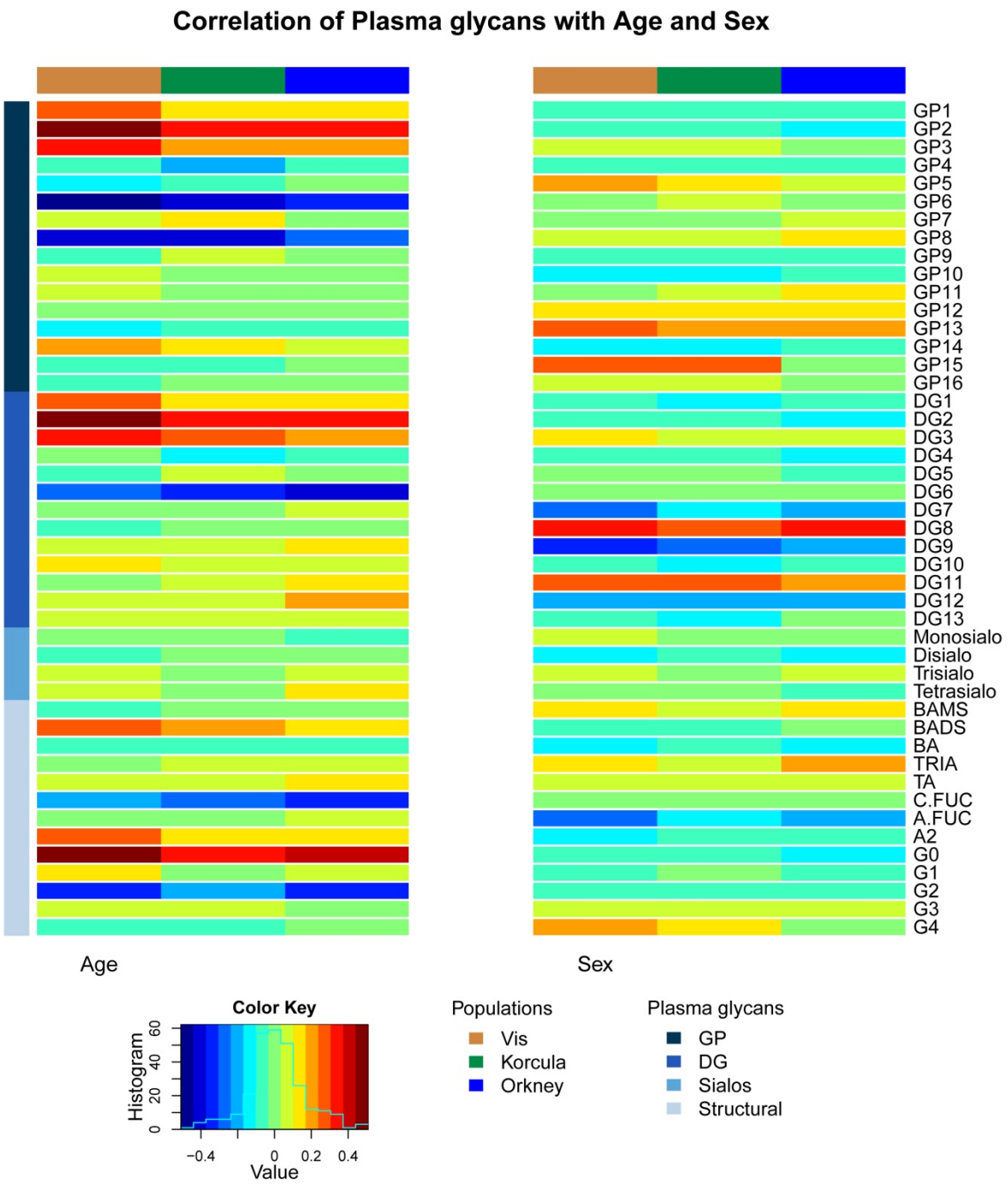
Supplementary figure 2. Affinity propagation clustering results for Vis cohort. Clustering results of affinity propagation algorithm for the Vis population based on plasma glycans (A), IgG glycans (B) and phenotypes (C). The results obtained for Korčula and Orkney population are similar to the ones obtained for Vis and, thus, were not presented. The heatmap represents the levels of each feature (rows) for the samples in each cluster (columns); the key colour of the heatmap varies from red to yellow corresponding to low and high values, respectively. The bar above the heatmap depicts the cluster division in different shades of grey. The bar on the left side of the heatmaps of plasma and IgG glycans represents the corresponding glycan groups as indicated in the legend.


Affinity propagation clustering based on IgG glycans
A
Populations
Clusters
IGG1
IGG2
IGG3
IGG4
IGG5
IGG6
IGG7
IGG8
IGG9
IGG10
IGG11
IGG12
IGG13
IGG14
IGG15
IGG16
IGG17
IGG18
IGG19
IGG20
IGG21
IGG22
IGG23
IGG24
IGG25
IGG26
IGG27
IGG28
IGG29
IGG30
IGG31
IGG32
IGG33
IGG34
IGG35
IGG36
IGG37
IGG38
IGG39
IGG40
IGG41
IGG42
IGG43
IGG44
IGG45
IGG46
IGG47
IGG48
IGG49
IGG50
IGG51
IGG52
IGG53
IGG54
IGG55
IGG56
IGG57
IGG58
IGG59
IGG60
IGG61
IGG62
IGG63
IGG64
IGG65
IGG66
IGG67
IGG68
IGG69
IGG70
IGG71
IGG72
IGG73
IGG74
IGG75
IGG76
IGG77

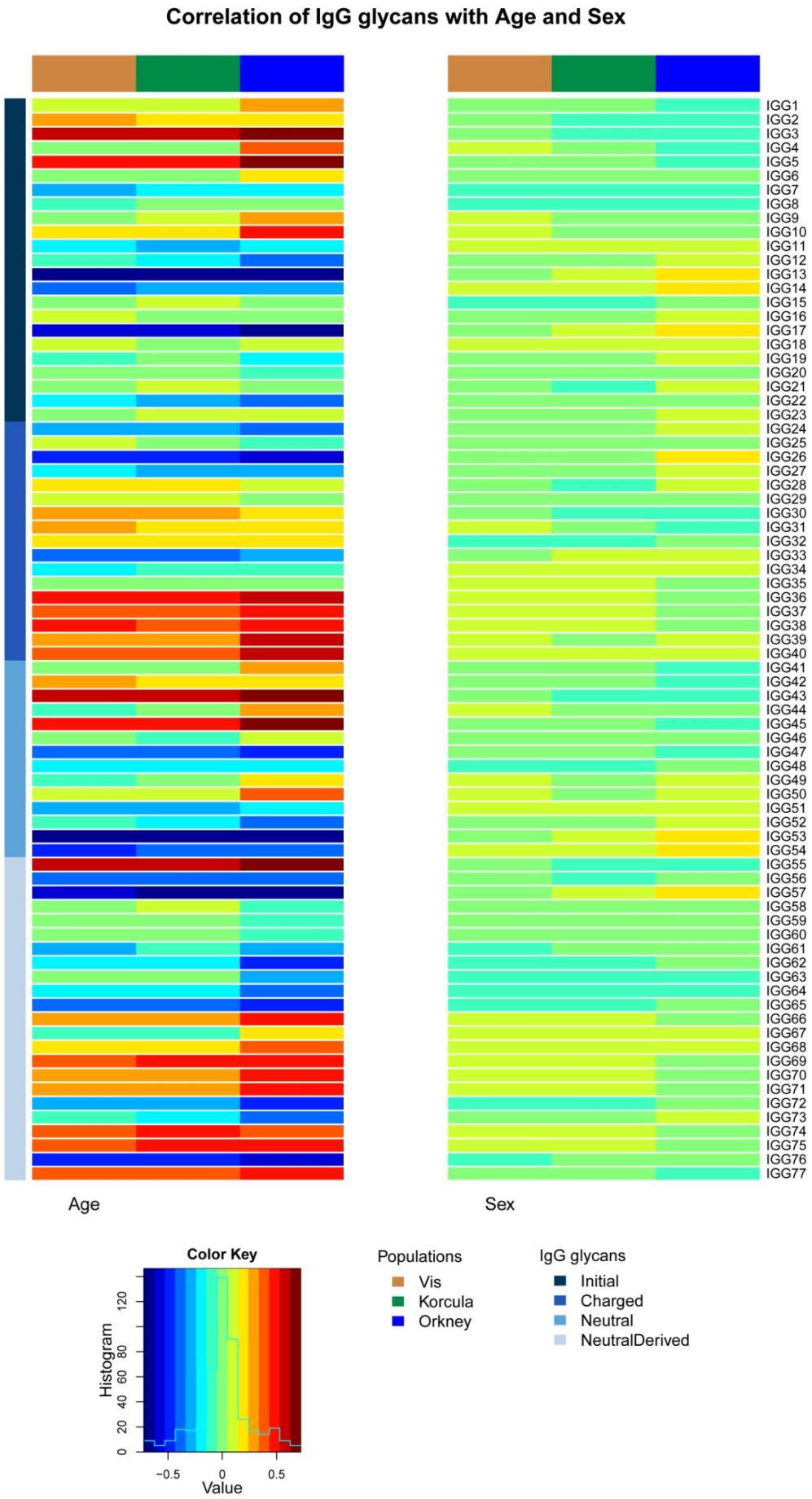


Supplementary figure 3. Affinity propagation clustering results based on IgG glycan profiles for the pooled data of populations. The clustering results of affinity propagation algorithm run with K=2 (A) and K=4 (B) are presented to illustrate the difficulty in establishing the most reliable clustering structure. In the case of the 2 cluster division, opposite levels of glycan features such as IGG3, IGG17, IGG43, IGG55 and IGG57 among others are clearly observed. In the case of the 4 cluster division, additional differences in IGG24-IGG27 and IGG62-IGG69 glycan features are revealed. The heatmap represents the levels of each glycan (rows) for the samples in each cluster (columns); the key colour of the heatmap varies from red to

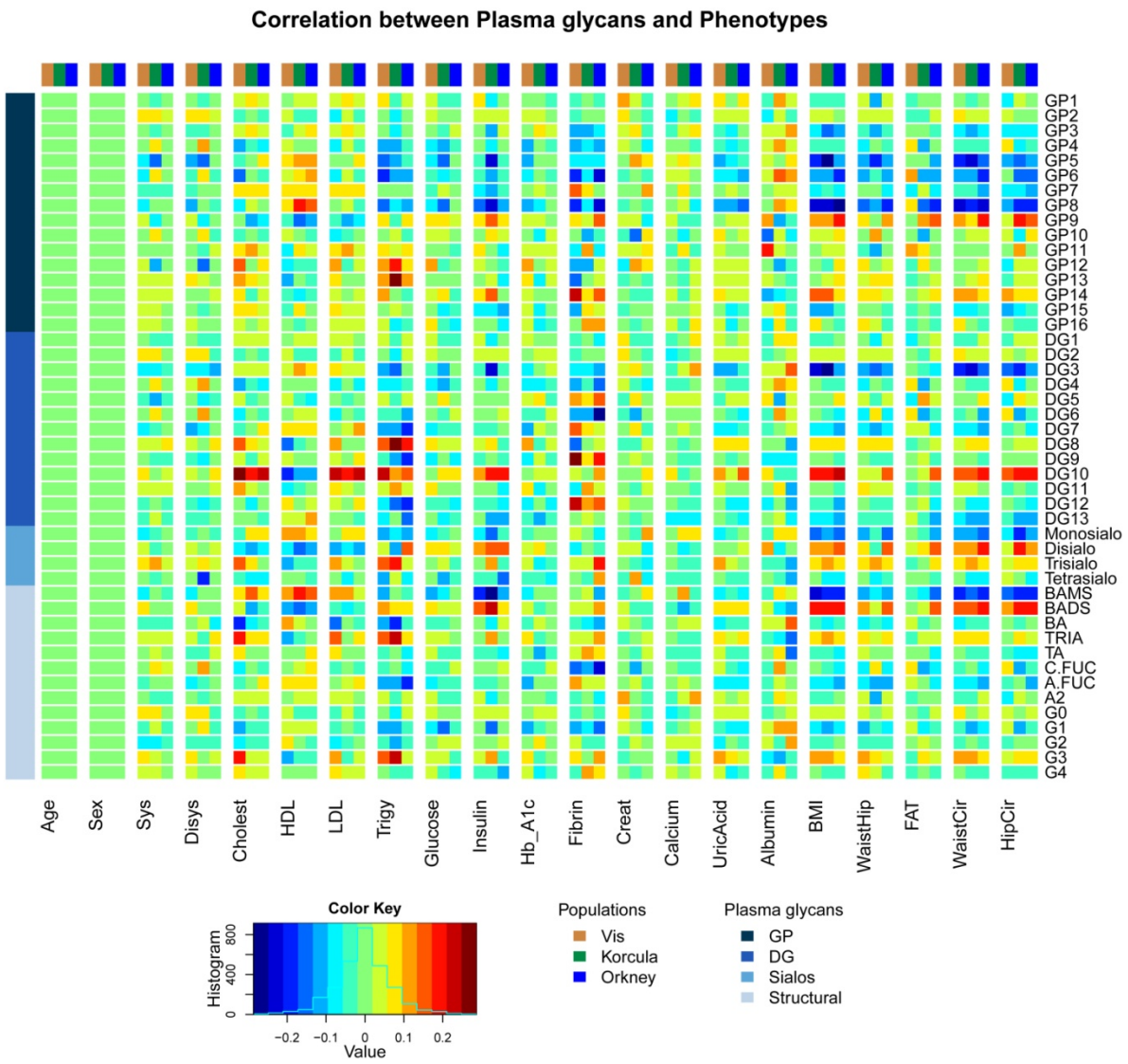
yellow corresponding to low and high values, respectively. The bars above the heatmap depict the cluster division in different shades of grey and the population division coloured as gold for Vis, green for Korčula and blue for Orkney. The bar on the left side of the heatmap indicates the four groups of IgG glycans: Initial (dark blue), Charged (blue), Neutral (medium blue) and Neutral derived (light blue).



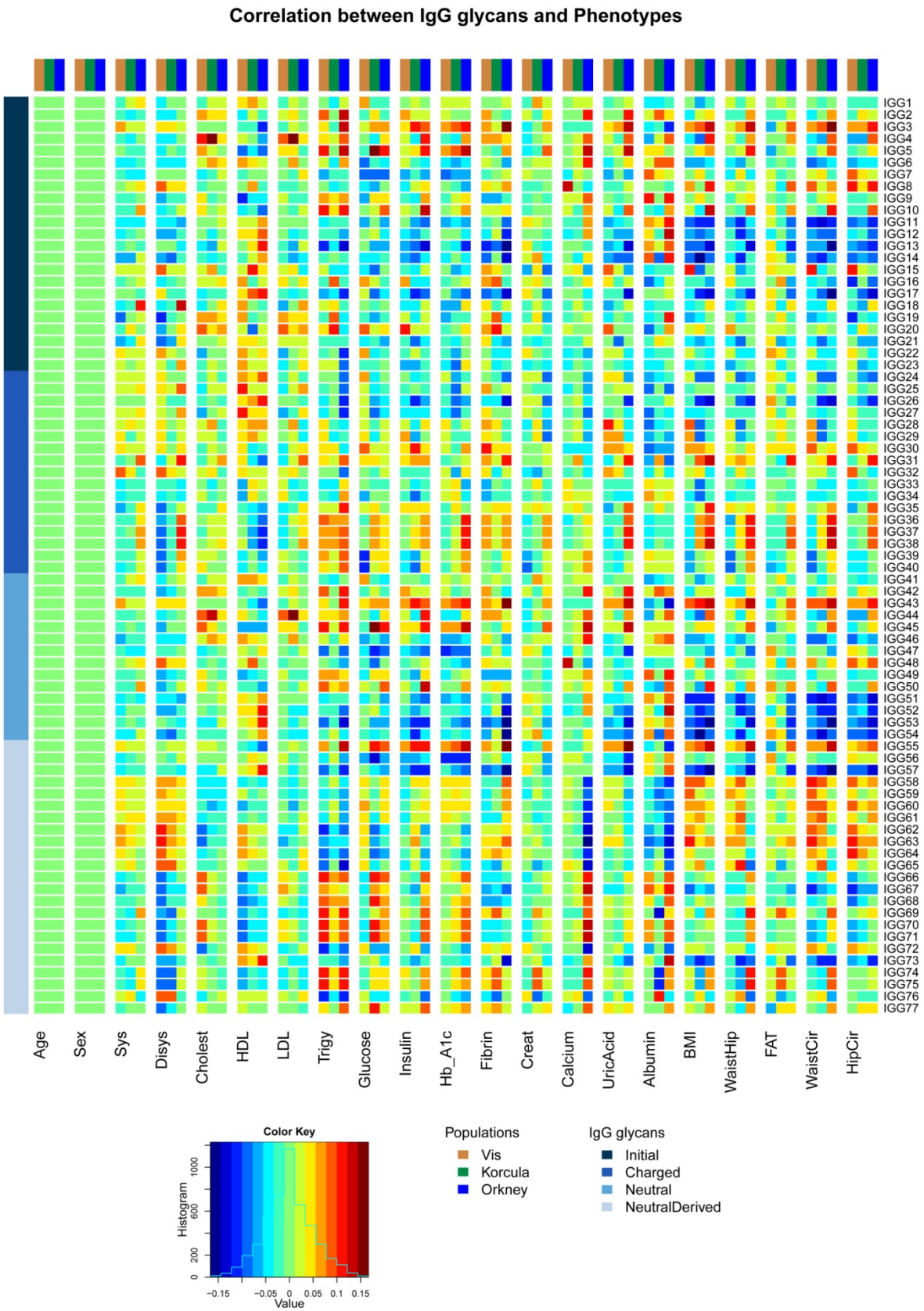
Supplementary figure 4. Correlation of plasma glycans with age and gender. The heatmap depicts the level of correlation between each plasma glycan feature (rows) and age and gender for each population (columns); correlation coefficients range from -0.65 (dark blue) to 0.65 (dark red). The bar above the heatmap indicates the population to which the three columns of each phenotype correspond to: gold for Vis, green for Korčula and blue for Orkney. The bar on the left side of the heatmap indicates the four groups of plasma glycans: GP (dark blue), DG (blue), Sialos (medium blue) and Structural (light blue).


Correlation of IgG glycans with Age and Sex
IGG1
IGG2
IGG3
IGG4
IGG5
IGG6
IGG7
IGG8
IGG9
IGG10
IGG11
IGG12
IGG13
IGG14
IGG15
IGG16
IGG17
IGG18
IGG19
IGG20
IGG21
IGG22
IGG23
IGG24
IGG25
IGG26
IGG27
IGG28
IGG29
IGG30
IGG31
IGG32
IGG33
IGG34
IGG35
IGG36
IGG37
IGG38
IGG39
IGG40
IGG41
IGG42
IGG43
IGG44
IGG45
IGG46
IGG47
IGG48
IGG49
IGG50
IGG51
IGG52
IGG53
IGG54
IGG55
IGG56
IGG57
IGG58
IGG59
IGG60
IGG61
IGG62
IGG63
IGG64
IGG65
IGG66
IGG67
IGG68
IGG69
IGG70
IGG71
IGG72
IGG73
IGG74
IGG75
IGG76
IGG77
Age
Sex
Color Key
Histogram
0
20
40
60
80
120
−0.5
0
0.5
Value
Populations
Vis
Korcula
Orkney
IgG glycans
Initial
Charged
Neutral
NeutralDerived

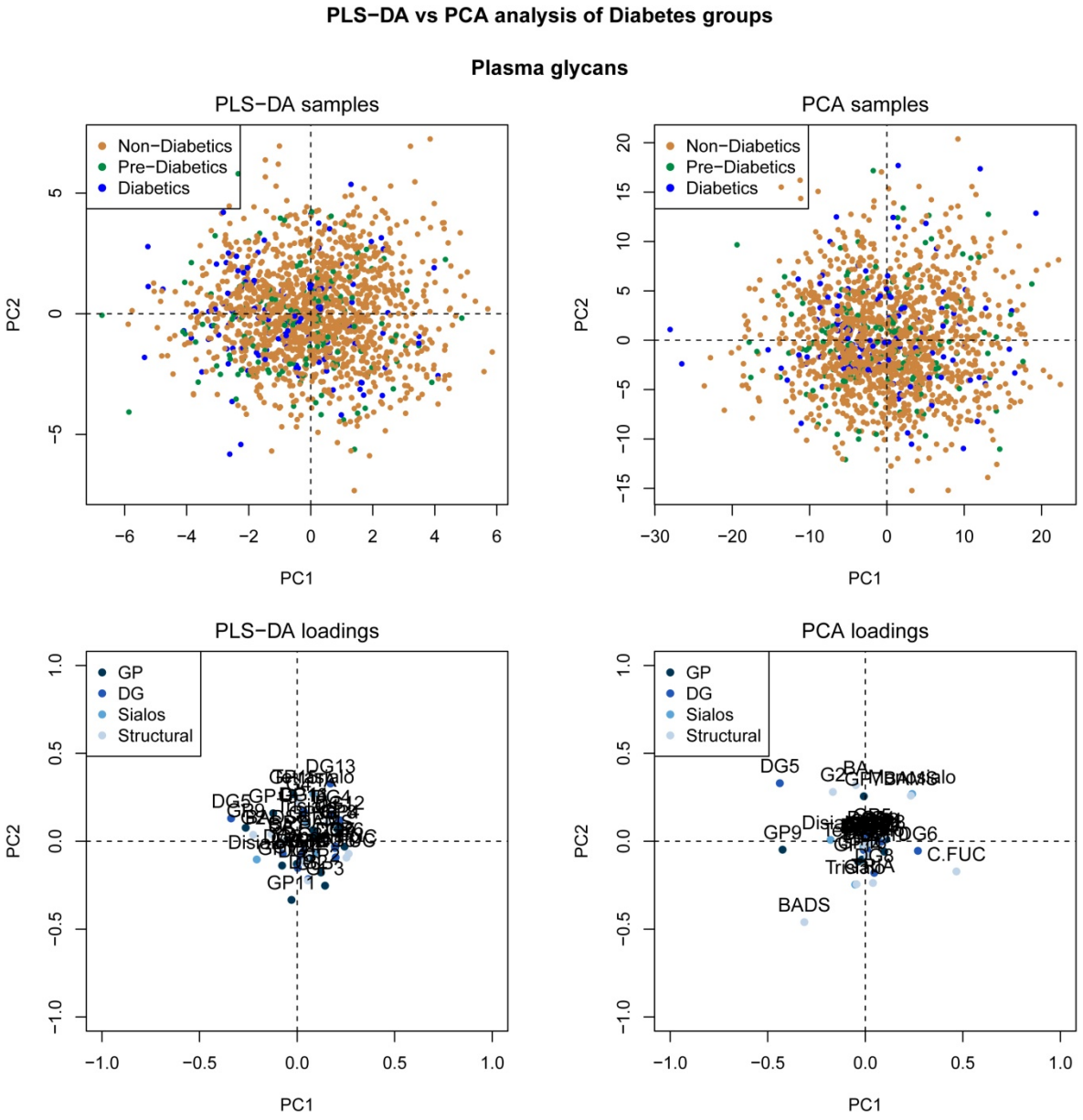
Supplementary figure 5. Correlation of IgG glycans with age and gender. The heatmap depicts the level of correlation between each IgG glycan feature (rows) and age and gender for each population (columns); correlation coefficients range from -0.65 (dark blue) to 0.65 (dark red). The bar above the heatmap indicates the population to which the three columns of each phenotype correspond to: gold for Vis, green for Korčula and blue for Orkney. The bar on the left side of the heatmap indicates the four groups of IgG glycans: Initial (dark blue), Charged (blue), Neutral (medium blue) and Neutral Derived (light blue).



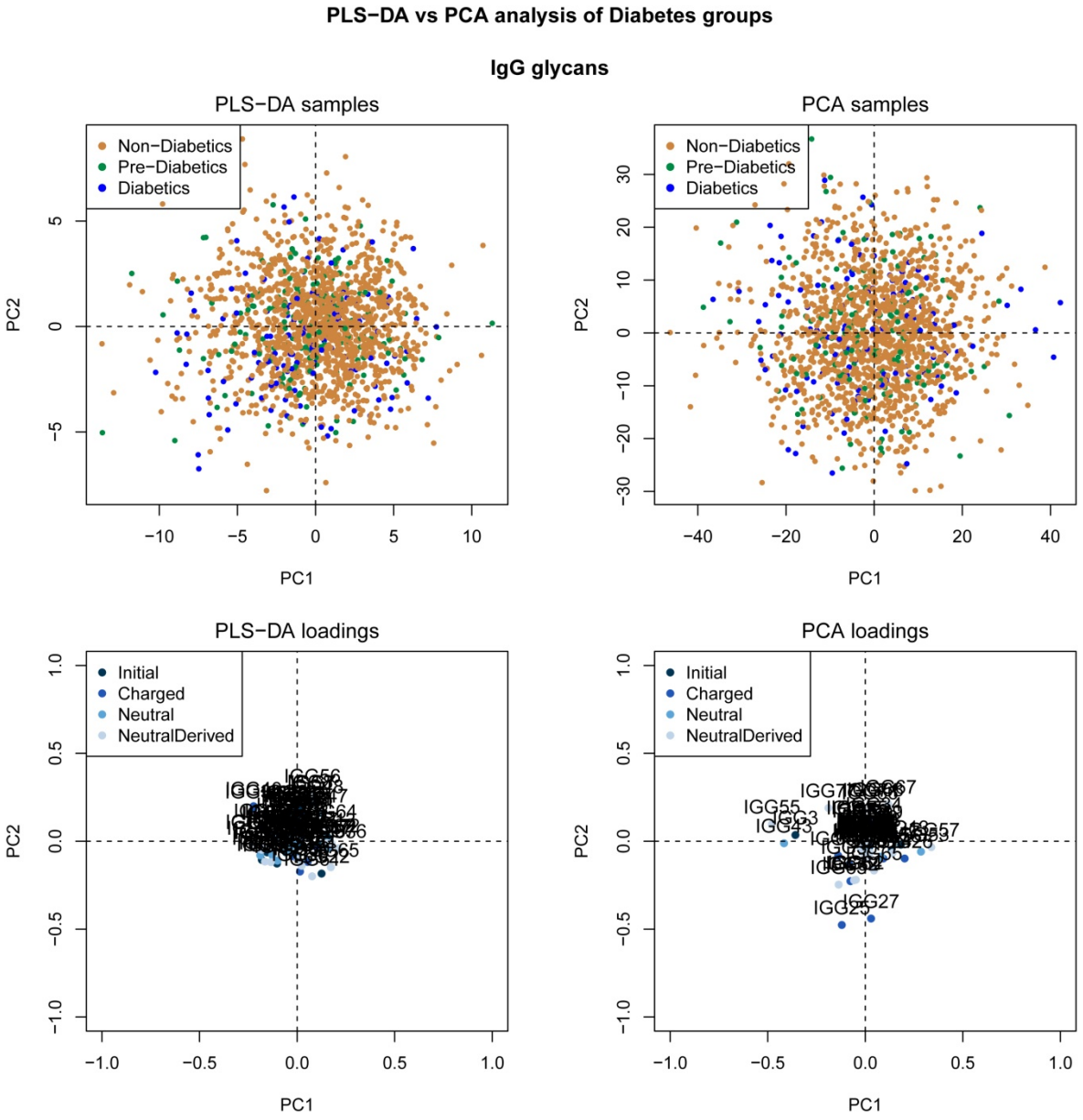
Supplementary figure 6. Correlation between plasma glycans and phenotypes for all populations. The heatmap depicts the level of correlation between each plasma glycan feature (rows) and the phenotypes for each population (columns); correlation coefficients range from -0.3 (dark blue) to 0.3 (dark red). The bar above the heatmap indicates the population to which the three columns of each phenotype correspond to: gold for Vis, green for Korčula and blue for Orkney. The bar on the left side of the heatmap indicates the four groups of plasma glycans: GP (dark blue), DG (blue), Sialos (medium blue) and Structural (light blue).


Correlation between IgG glycans and Phenotypes
IGG1
IGG2
IGG3
IGG4
IGG5
IGG6
IGG7
IGG8
IGG9
IGG10
IGG11
IGG12
IGG13
IGG14
IGG15
IGG16
IGG17
IGG18
IGG19
IGG20
IGG21
IGG22
IGG23
IGG24
IGG25
IGG26
IGG27
IGG28
IGG29
IGG30
IGG31
IGG32
IGG33
IGG34
IGG35
IGG36
IGG37
IGG38
IGG39
IGG40
IGG41
IGG42
IGG43
IGG44
IGG45
IGG46
IGG47
IGG48
IGG49
IGG50
IGG51
IGG52
IGG53
IGG54
IGG55
IGG56
IGG57
IGG58
IGG59
IGG60
IGG61
IGG62
IGG63
IGG64
IGG65
IGG66
IGG67
IGG68
IGG69
IGG70
IGG71
IGG72
IGG73
IGG74
IGG75
IGG76
IGG77
Age
Sex
Sys
Disys
Cholest
HDL
LDL
Trigy
Glucose
Insulin
Hb_A1c
Fibrin
Creat
Calcium
UricAcid
Albumin
BMI
WaistHip
FAT
WaistCir
HipCir
Color Key
Histogram
0
200
600
1000
−0.15
−0.05
0
0.05
0.1
0.15
Value
Populations
Vis
Korcula
Orkney
IgG glycans
Initial
Charged
Neutral
NeutralDerived

Supplementary figure 7. Correlation between IgG glycans and phenotypes for all populations. The heatmap depicts the level of correlation between each IgG glycan feature (rows) and the phenotypes for each population (columns); correlation coefficients range from -0.15 (dark blue) to 0.15 (dark red). The bar above the heatmap indicates the population to which the three columns of each phenotype correspond to: gold for Vis, green for Korčula and blue for Orkney. The bar on the left side of the heatmap indicates the four groups of IgG glycans: Initial (dark blue), Charged (blue), Neutral (medium blue) and Neutral derived (light blue).



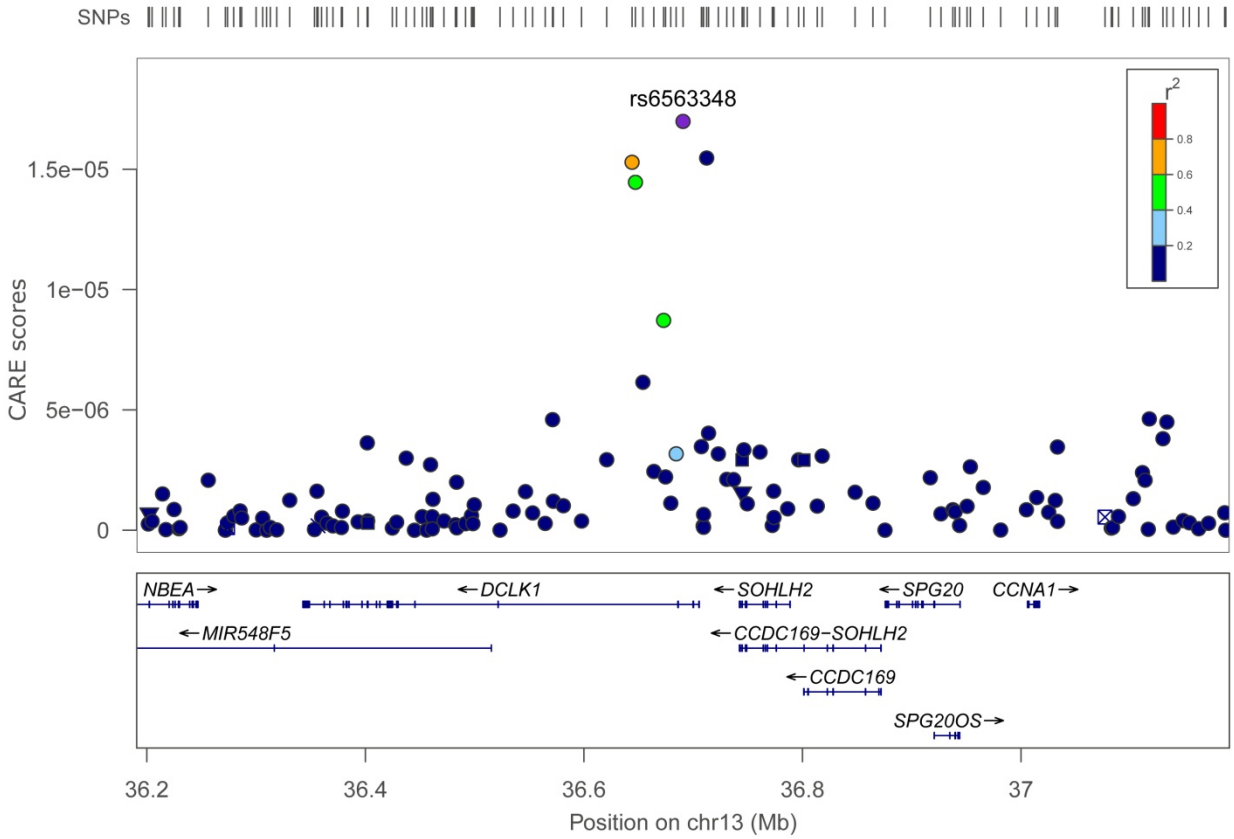
Supplementary figure 8. PLS-DA and PCA analysis of the diabetes groups using plasma glycans data. The score plots representing the data samples by the two first principal components (PC1 on the x-axis and PC2 on the y-axis) are shown on the upper panels; groups are coloured as gold for non-diabetic, green for pre-diabetic and blue for diabetic. The corresponding loading plots establishing the relative contributions of each plasma glycan feature to the overall variation in the groups are shown on the lower panels; glycans are coloured according to their group: GP (dark blue), DG (blue), Sialos (medium blue) and Structural (light blue).



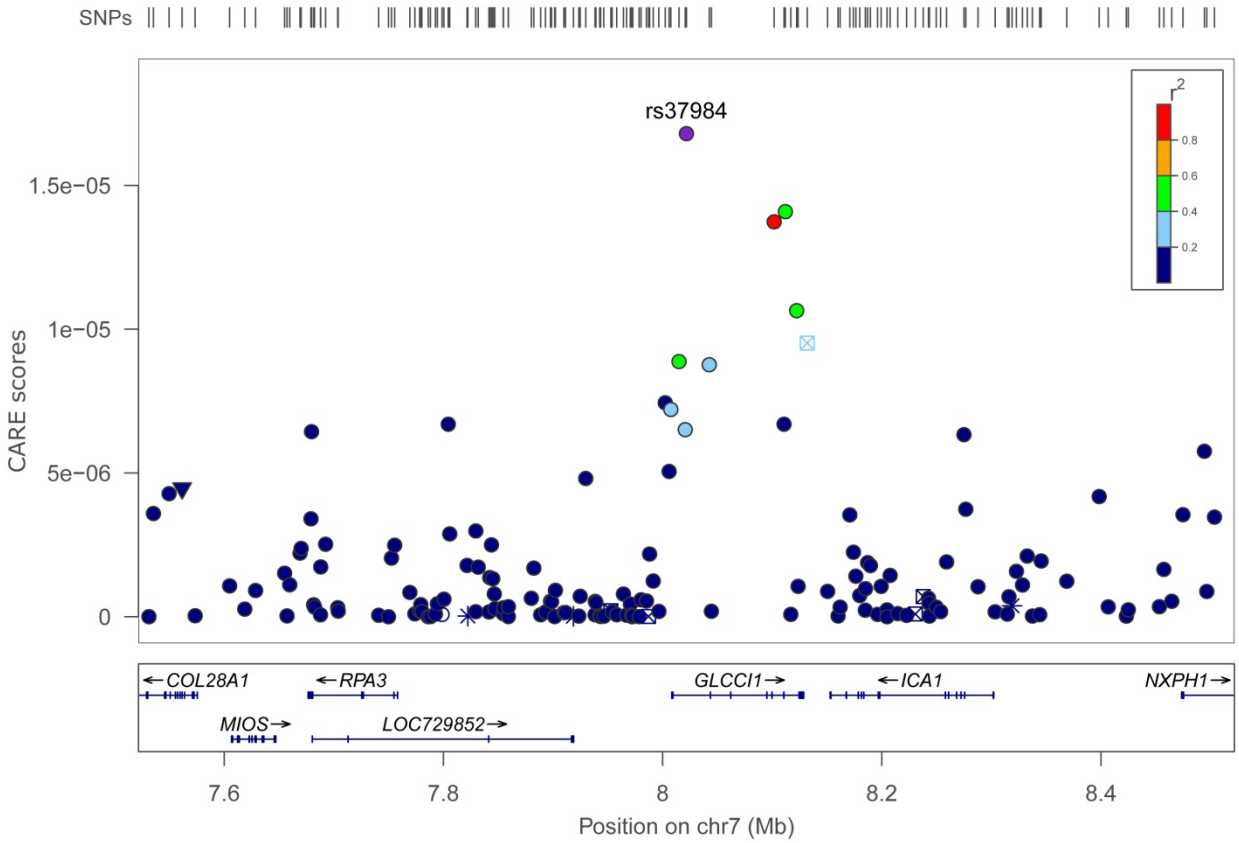
Supplementary figure 9. PLS-DA and PCA analysis of the diabetes groups using IgG glycans data. The score plots representing the data samples by the two first principal components (PC1 on the x-axis and PC2 on the y-axis) are shown on the upper panels; groups are coloured as gold for non-diabetic, green for pre-diabetic and blue for diabetic. The corresponding loading plots establishing the relative contributions of each IgG glycan feature to the overall variation in the groups are shown on the lower panels; glycans are coloured according to their group: Initial (dark blue), Charged (blue), Neutral (medium blue) and Neutral derived (light blue).

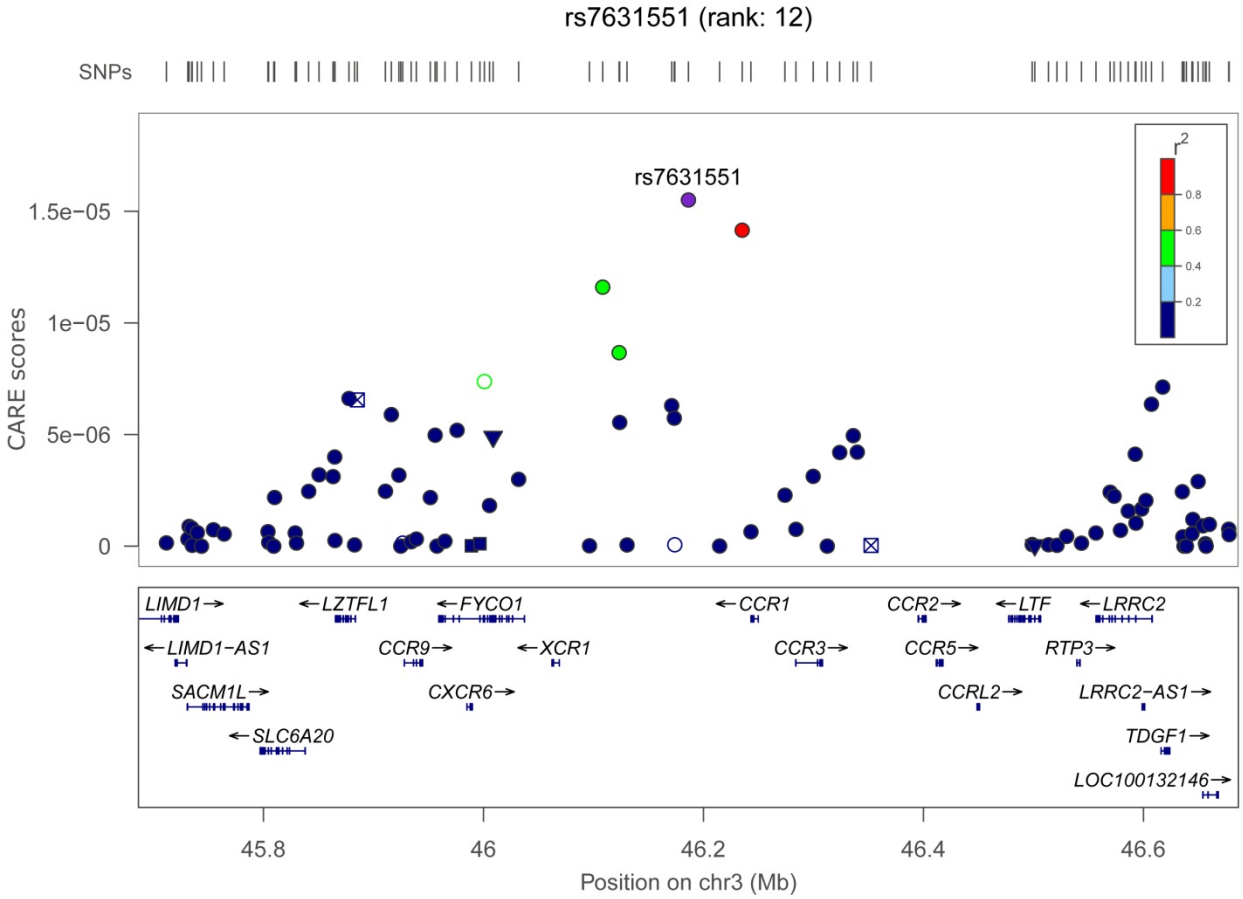
Genetic context of polymorphisms possibly associated with Diabetes

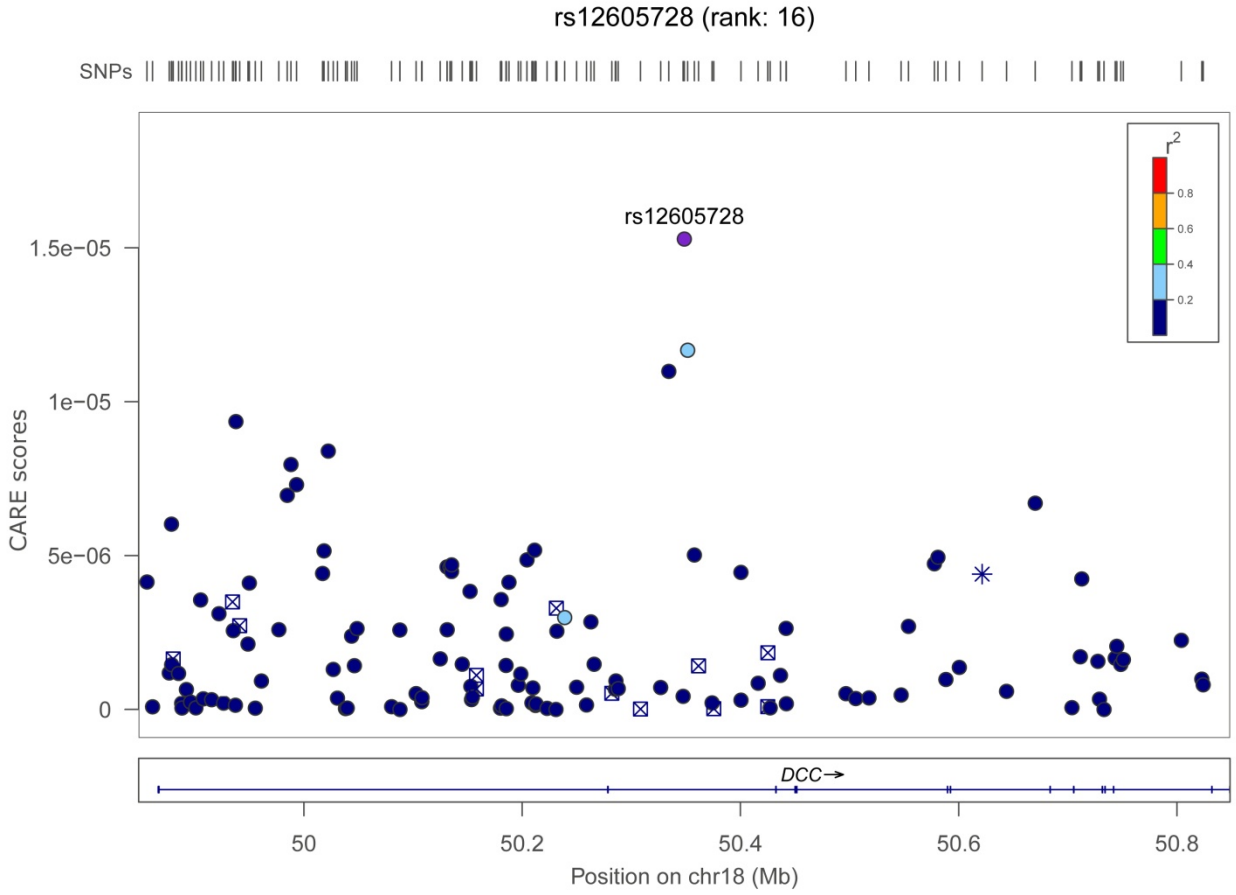
rs6563348 (rank: 4)

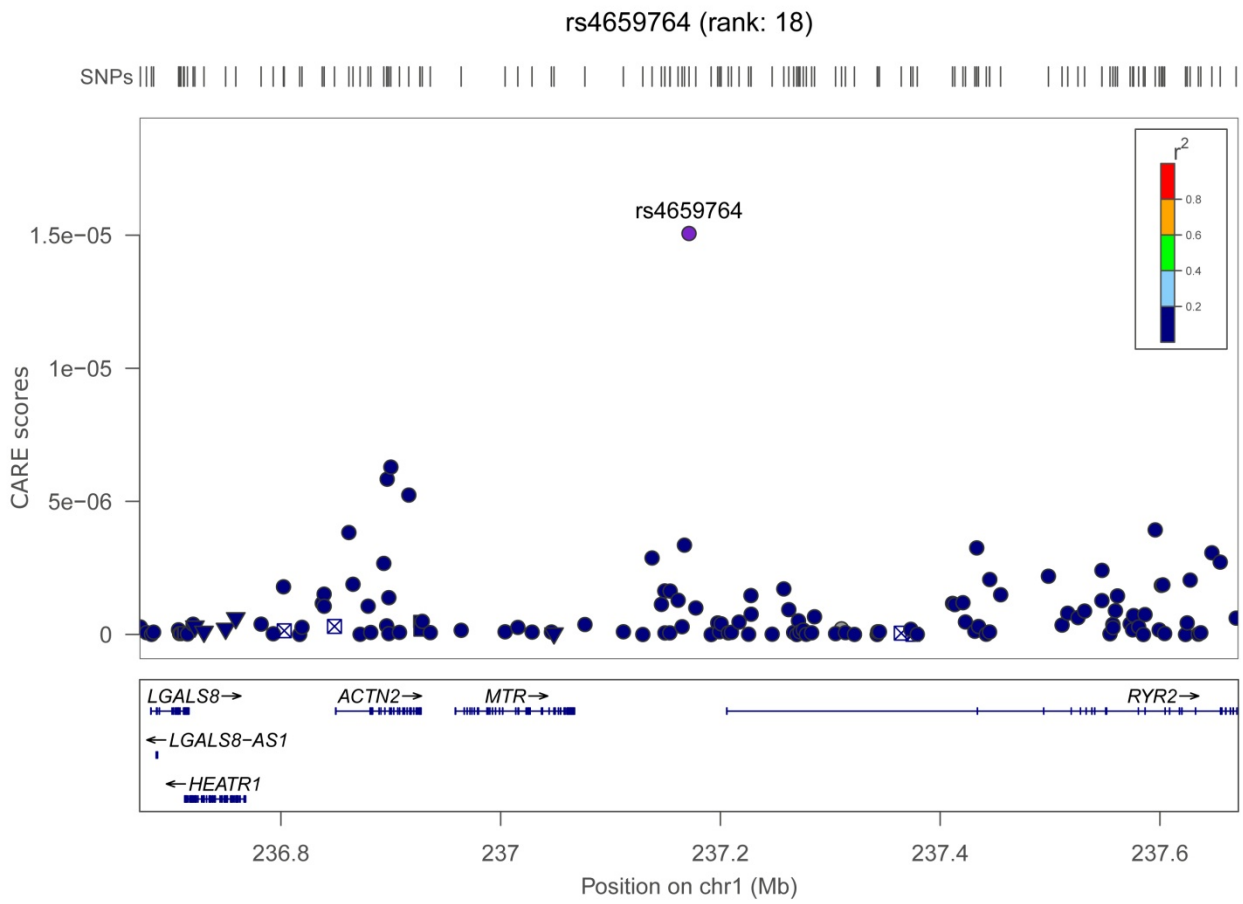


rs37984 (rank: 6)

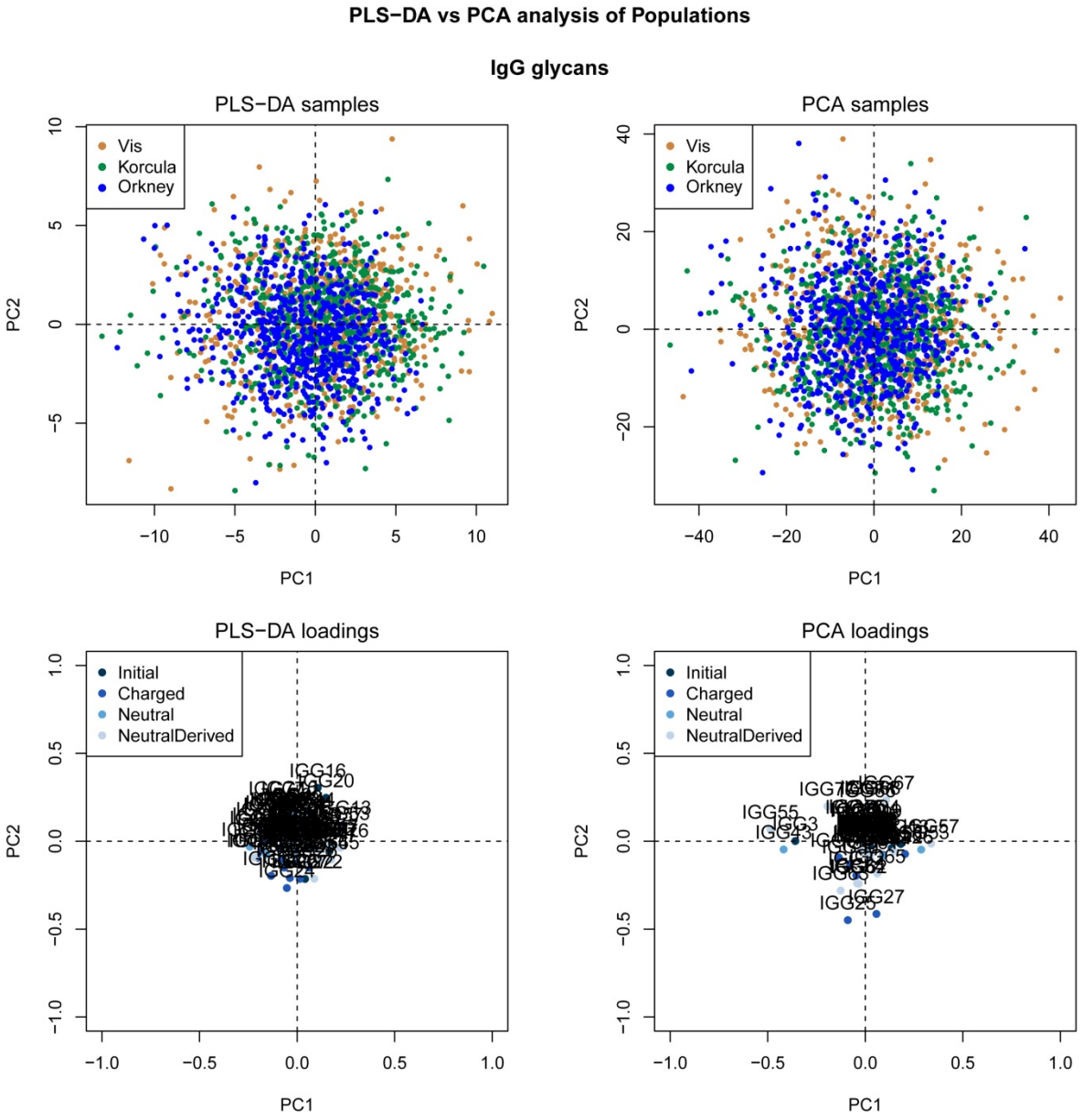



rs7631551 (rank: 12)
SNPs
rs7631551
CARE scores
1.5e−05
1e−05
5e−06
0
r2
0.8
0.6
0.4
0.2
LIMD1
LZTFL1
FYCO1
CCR1
CCR2
LTF
LRRC2
LIMD1−AS1
CCR9
XCR1
CCR3
CCR5
RTP3
SACM1L
CXCR6
CCRL2
LRRC2−AS1
SLC6A20
TDGF1
LOC100132146
45.8
46
46.2
46.4
46.6
Position on chr3 (Mb)

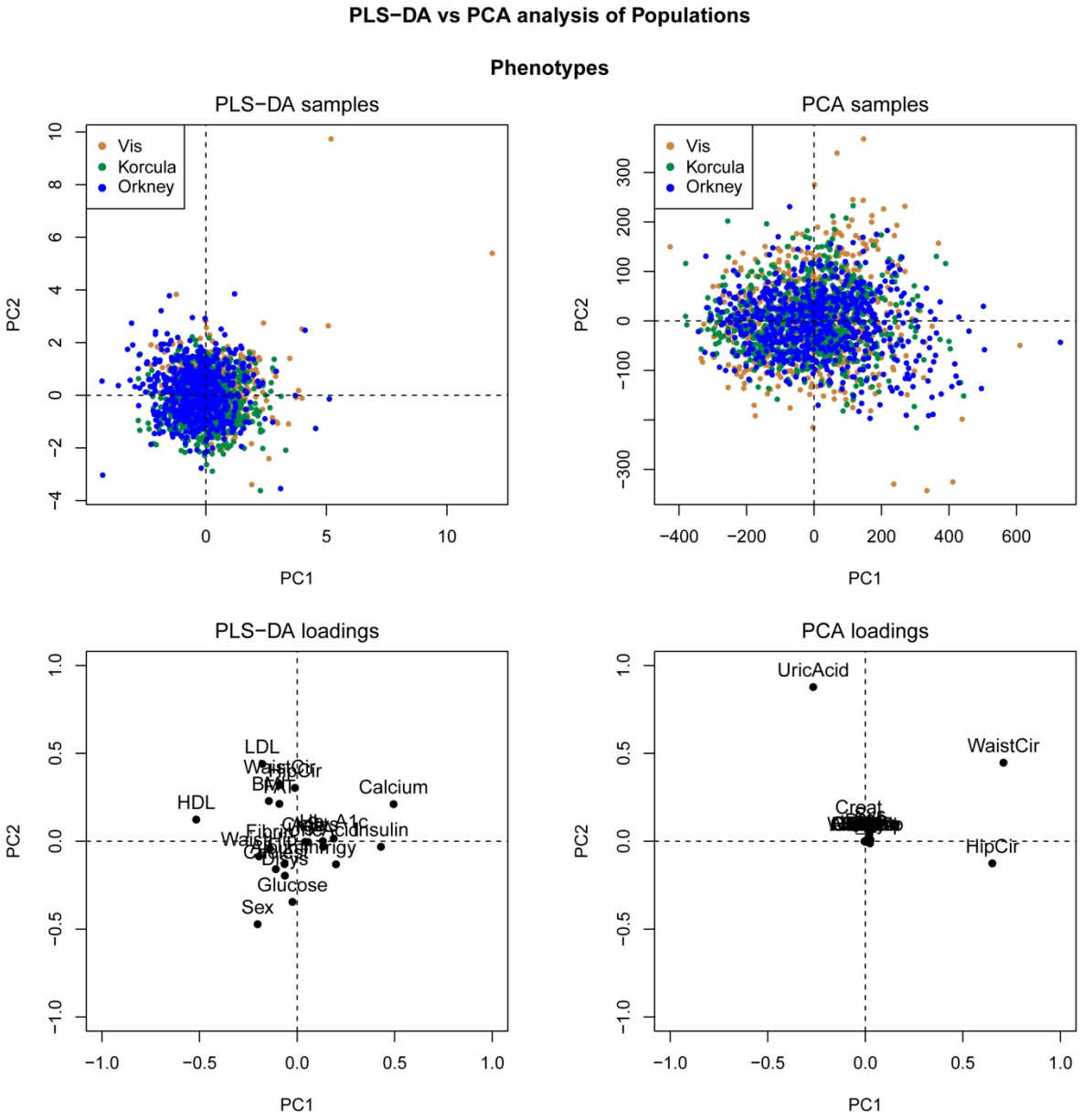

rs12605728 (rank: 16)
SNPs
rs12605728
CARE scores
1.5e−05
1e−05
5e−06
0
r2
0.8
0.6
0.4
0.2
DCC
50
50.2
50.4
50.6
50.8
Position on chr18 (Mb)



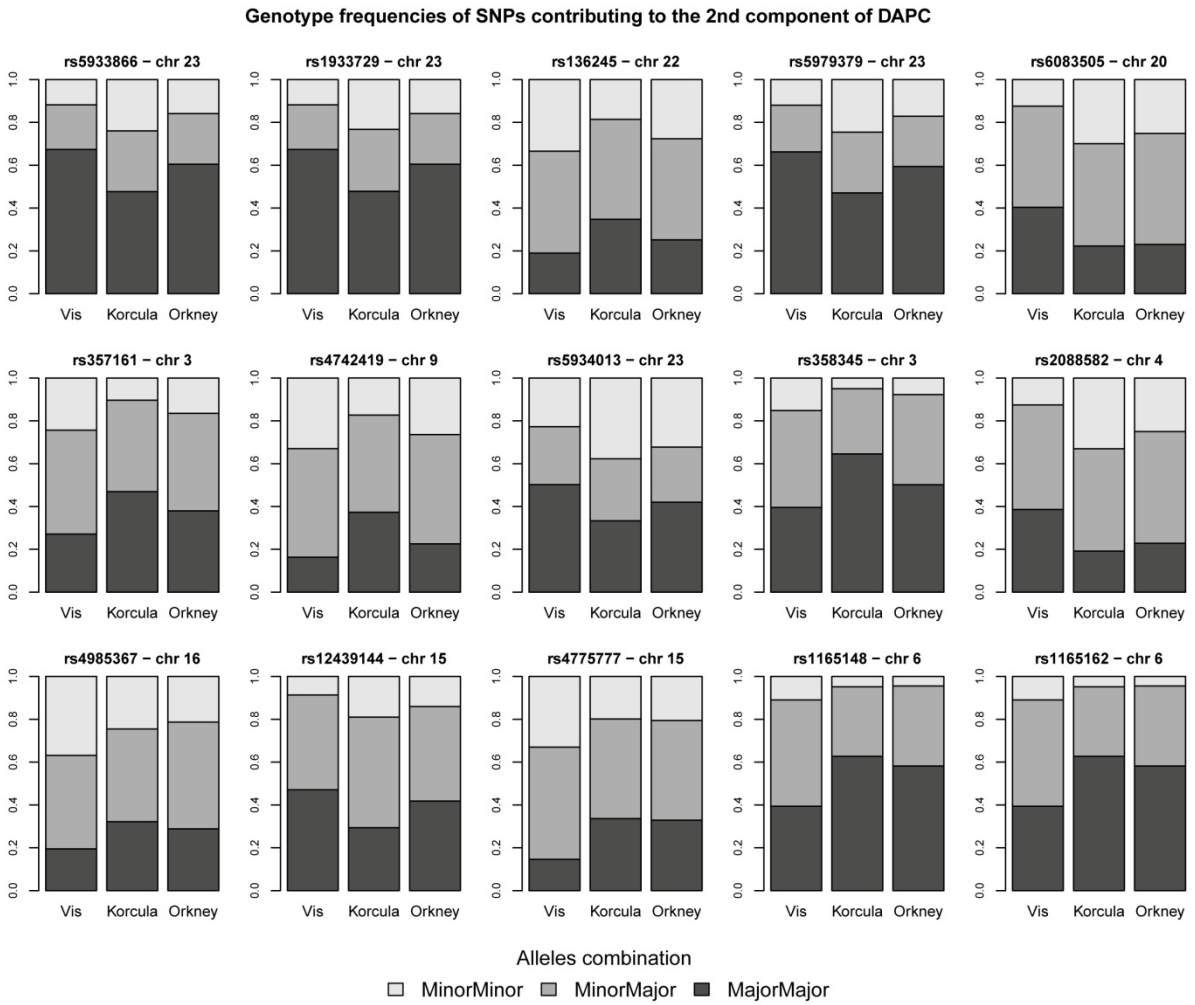
Supplementary figure 10. Genetic context of several polymorphisms possibly associated with the diabetes condition. The regional association plots show the correlation adjusted scores (CARE scores) for SNPs distributed in five genomic regions centred on variants rs6563348 (chromosome 13), rs37984 (chromosome 7), rs7631551 (chromosome 3), rs12605728 (chromosome 18) and rs4659764 (chromosome 1). The flanking region extends 0.5Mb both upstream and downstream of the reference SNP which is labelled and shown in purple. The colour intensity of the other SNPs within the region represents the extent of their linkage disequilibrium (r^2) with the reference SNP: red ($r^2 \geq 0.8$), orange ($0.6 \leq 0.8$), green ($0.4 \leq 0.6$), light blue ($0.2 \leq 0.4$) and dark blue ($r^2 \leq 0.2$). The locations of known genes in the region are depicted below the association plot.



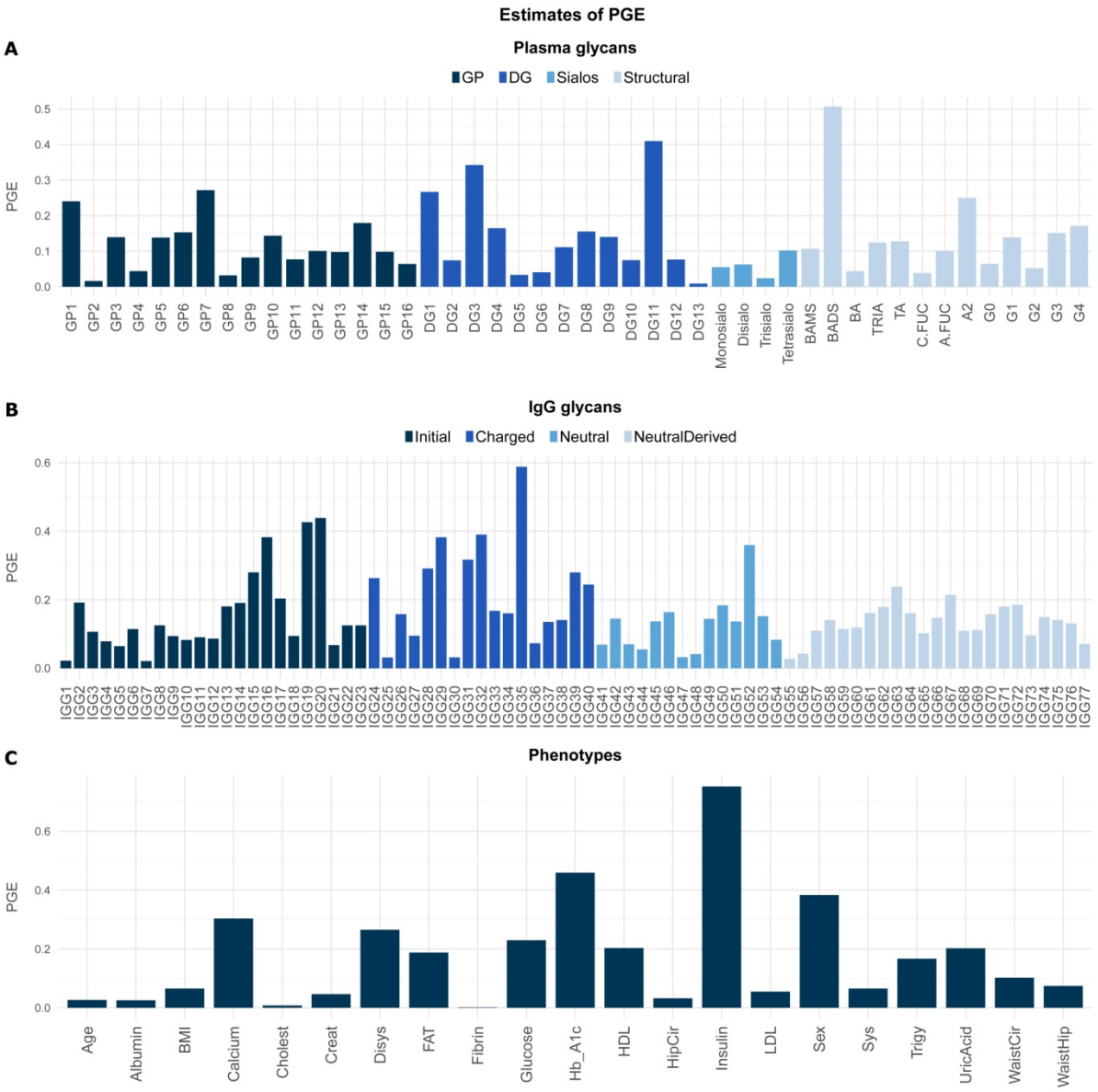
Supplementary figure 11. PLS-DA and PCA analysis of the population cohorts using IgG glycans data. The score plots representing the data samples by the two first principal components (PC1 on the x-axis and PC2 on the y-axis) are shown on the upper panels; the populations are coloured as gold for Vis, green for Korčula and blue for Orkney. The corresponding loading plots establishing the relative contributions of each IgG glycan feature to the overall variation in the populations are shown on the lower panels; glycans are coloured according to their group: Initial (dark blue), Charged (blue), Neutral (medium blue) and Neutral derived (light blue).



Supplementary figure 12. PLS-DA and PCA analysis of the population cohorts using phenotypes data. The score plots representing the data samples by the two first principal components (PC1 on the x-axis and PC2 on the y-axis) are shown on the upper panels; the populations are coloured as gold for Vis, green for Korčula and blue for Orkney. The corresponding loading plots establishing the relative contributions of each phenotype feature to the overall variation in the populations are shown on the lower panels.



Supplementary figure 13. Genotype frequencies of the 15 SNPs most contributing to the second discriminant component of the DAPC analysis of the population cohorts. SNPs are located along various chromosomes and either Vis or Korčula show slightly differences in genotype profiles. Genotypes are coded in grey shades with light grey corresponding to minor-minor allele combination, medium grey corresponding to minor-major allele and dark grey corresponding to major-major allele.



Supplementary figure 14. PGE estimates for all traits of the three feature data sets. The estimates of PGE by the Bayesian sparse linear mixed model are represented for each trait of plasma glycans (A), IgG glycans (B) and phenotypes (C). For plasma and IgG glycans the colours of the bars represent the glycan groups as indicated in the corresponding legend above the barplot.

APPENDIX B. Supplementary Tables

Supplementary table 1. Glycan structures present in different HPLC peaks. Chromatographic peaks obtained with HILIC analysis (GP1-GP16, left) and with HILIC after sialidase treatment (DG1-DG13, right) and the individual glycan structures present in each peak.

PEAK	STRUCTURE
GP1	A2
GP2	A2B A1G1 FA2
GP3	M5 FA2B A2[6]G1 A2[6]BG1
GP4	A2[3]G1 A2[3]BG1 M4A1G1 FA2[6]G1 FA2[6]BG1 A1[6]G1S(3)1 A1[6]G1S(6)1 FA2[3]G1 FA2[3]BG1 M6D1, D2 A1[3]G1S(3)1 A1[3]G1S(6)1
GP5	M6D3 A2[6]G1S(3)1 A2[6]G1S(6)1 A2G2 A2[3]G1S(3)1 A2[3]G1S(6)1 A2BG2
GP6	FA2[6]G1S(3)1 FA2[6]G1S(6)1 FA2[6]BG1S(3)1 FA2[6]BG1S(6)1 M4A1G1S1 FA2G2 FA2[3]G1S(3)1 FA2[3]G1S(6)1 A2BG1S1 FA2[3]BG1S(3)1 FA2[3]BG1S(6)1
GP7	FA2BG2 M7D3 A2G2S(3)1 A2G2S(6)1 M7D1
GP8	A2BG2S(3)1 A2BG2S(6)1 M5A1G1S1 FA2G2S(3)1 FA2G2S(6)1 A3G3 FA2BG2S(3)1 FA2BG2S(6)1
GP9	A2F1G2S(3)1 A2F1G2S(6)1 M8D2, D3 A2G2S(3,3)2 A2G2S(3,6)2 A2G2S(6,6)2 M8D1,D3
GP10	A2BG2S(3,3)2 A2BG2S(3,6)2 A2BG2S(6,6)2 A3BG3S(3)1 A3BG3S(6)1 FA2G2S(3,3)2 FA2G2S(3,6)2 FA2G2S(6,6)2
GP11	FA2BG2S(3,3)2 FA2BG2S(3,6)2 FA2BG2S(6,6)2 M9
GP12	A2F1G2S2 A3G3S(3,3)2 A3G3S(3,6)2 A3G3S(6,6)2 A3BG3S(3,3)2 A3BG3S(3,6)2 A3BG3S(6,6)2
GP13	A3G3F1S2 FA3G3S(3,3)2 FA3G3S(3,6)2 FA3G3S(6,6)2 FA3BG3S(3,3)2 FA3BG3S(3,6)2 FA3BG3S(6,6)2 A3G3S(3,3,6)3 A3G3S(3,6,6)3 A3G3S(6,6,6)3
GP14	A3F1G3S(3,3,6)3 FA3F1G3S(6,6,6)3 A4G4S(6,6)2 A3F1G3S(3,6,6)3 A3F1G3S(6,6,6)3 A4G4S(6,6,6)3 A4F1G4S2 A4G4S3
GP15	A4G4S4 A4F1G4S3
GP16	A4G4S(6,6,6,6)4 A4G4S(3,6,6,6)4 A4BG4S4 FA4G4S4 A4F1G4S4 A4G4LacS4 A4F2G4S4 FA4F1G4S4

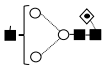
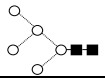
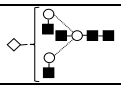
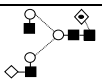
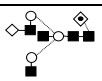
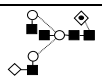
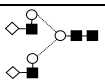
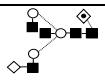
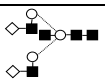
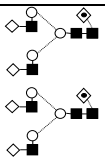
PEAK	STRUCTURE
DG1	A2
DG2	A2B A1G1 FA2
DG3	M5 FA2B A2[6]G1 A2[6]BG1
DG4	M4A1G1 A2[3]G1 A2[3]BG1 FA2[6]BG1 FA2[3]G1 FA2[3]BG1
DG5	M6D1, D2 M6D3 A2G2 A2BG2
DG6	FA2G2 M5A1G1 FA2BG2
DG7	M7D3 A2F1G2 M7D1
DG8	A3G3 A2F2G2 FA3G3 M8D2, D3 M8D1,D3
DG9	FA3BG3 A3F1G3
DG10	M9 FA3F1G3
DG11	A4G4 A4BG4 A3F2G3 FA4G4
DG12	A4F1G4
DG13	A4G4Lac A4F2G4 FA4F1G4

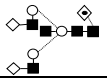
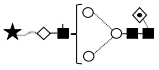
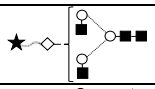
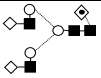
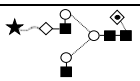
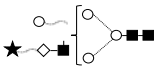
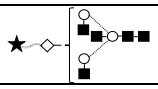
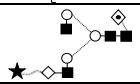
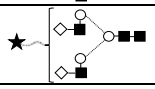
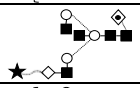
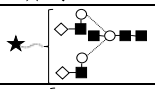
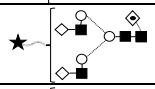
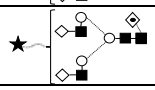
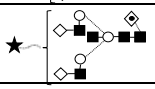
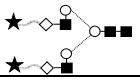
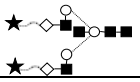
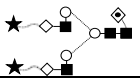
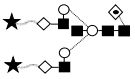
Structure abbreviations: all N-glycans have two core GlcNAcs; F at the start of the abbreviation indicates a core fucose α1-6 linked to the inner GlcNAc; Mx, number (x) of mannose on core GlcNAcs; D1 indicates that the α1-2 mannose is on the Manα1-6Manα1-6 arm, D2 on the Manα1-3Manα1-6 arm, D3 on the Manα1-3 arm of M6 and on the Manα1-2Manα1-3 arm of M7 and M8; Ax, number of antenna (GlcNAc) on trimannosyl core; A2, biantennary with both GlcNAcs as β1-2 linked; A3, triantennary with a GlcNAc linked β1-2 to both mannose and the third GlcNAc linked β1-4 to the α1-3 linked mannose; A4, GlcNAcs linked as A3 with additional GlcNAc β1-6 linked to α1-6 mannose; B, bisecting GlcNAc linked β1-4 to β1-3 mannose; Gx, number (x) of β1-4 linked galactose on antenna; [3]G1 and [6]G1 indicates that the galactose is on the antenna of the α1-3 or α1-6 mannose; F(x), number (x) of fucose linked α1-3 to antenna GlcNAc; Lac(x), number (x) of lactosamine (Galβ1-4GlcNAc) extensions; Sx, number (x) of sialic acids linked to galactose; the numbers 3 or 6 or in parentheses after S indicate whether the sialic acid is in an α2-3 or α2-6 linkage. If there is no linkage number, the exact link is unknown.

Supplementary table 2. Glycan structural features derived from the plasma glycome peaks. Derived plasma glycosylation traits were approximated by adding the chromatographic peaks from either HILIC or HILIC after sialidase treatment sharing the same structural characteristics.

GLYCAN STRUCTURAL FEATURE	TRAIT CODE	DESCRIPTION	FORMULA
Fucosylation (position of fucose)	FUC-C	Core fucosylated	DG6/(DG5+DG6)*100
	FUC-A	Antennary fucosylated	DG7/(DG5+DG7)*100
Degree of branching	BA	Biantennary	DG1+DG2+DG3+DG4+DG5+DG6+DG7
	TRIA	Triantennary	DG8+DG9+DG10
	TA	Tetraantennary	DG11+DG12+DG13
Sialylation of biantennary structures	BAMS	Monosialylated biantennary	(GP7+GP8)/(DG5+DG6+DG7)*100
	BADS	Disialylated biantennary	(GP9+GP10+GP11)/(DG5+DG6+DG7)*100
Galactosylation	G0	Nongalactosylated	DG1+DG2
	G1	Monogalactosylated	DG3+DG4
	G2	Digalactosylated	DG5+DG6+DG7
	G3	Trigalactosylated	GP12+GP13+GP14
	G4	Tetragalactosylated	GP15+GP16
	A2	Biantennary nongalactosylated	(GP1+DG1)/2

Supplementary table 3. Composition of the IgG glycome. The IgG glycome was separated into 24 chromatographic peaks by HILIC-UPLC and the individual glycan structures each peak were determined by mass spectrometry. The peaks are named IGG1-IGG23 along the thesis, with the peak GP3 excluded from the analyis as explained in the main text.

Glycan peak	Peak composition	Structure	%
GP1	F(6)A1		100
GP2	A2		100
GP3	A2B		100
GP4	F(6)A2		100
GP5	M5		63
	F(6)A2		37
GP6	F(6)A2B		97
	A2[6]G1		3
GP7	A2[3]G1		75
	F(6)A2B		25
GP8a	A2BG1		93
	F(6)A2[6]G1		7
GP8b	F(6)A2[6]G1		100
GP9	F(6)A2[3]G1		100
GP10	F(6)A2[6]BG1		100
GP11	F(6)A2[3]BG1		100
GP12	A2G2		91
	F(6)A2[3]BG1		9
GP13	A2BG2		87
			

Glycan peak	Peak composition	Structure	%
GP15	F(6)A2BG2		83
	F(6)A1G1S1		8
	A2G1S1		5
	F(6)A2G2		4
GP16a	F(6)A2[6]G1S1		63
	M4A1G1S1		25
	A2BG1S1		13
GP16b	F(6)A2[3]G1S1		91
	F(6)A2[6]BG1S1		9
GP17	A2G2S1		89
	F(6)A2[3]BG1S1		11
GP18a	A2BG2S1		91
	F(6)A2G2S1		9
GP18b	F(6)A2G2S1		100
GP19	F(6)A2BG2S1		100
GP20	n.d.		/
GP21	A2G2S2		100
GP22	A2BG2S2		100
GP23	F(6)A2G2S2		100
			

	F(6)A2G2	13
GP14	F(6)A2G2	100

GP24	F(6)A2BG2S2	100

Structure abbreviations: all N-glycans have core sugar sequence consisting of two N-acetylglucosamines (GlcNAc) and three mannose residues; F indicates a core fucose α1–6 linked to the inner GlcNAc; Mx, number (×) of mannose on core GlcNAcs; Ax, number of antenna (GlcNAc) on trimannosyl core; A2, biantennary glycan with both GlcNAcs as β1–2 linked; B, bisecting GlcNAc linked β1–4 to β1–3 mannose; Gx, number of β1–4 linked galactose (G) on antenna; [3]G1 and [6]G1 indicates that the galactose is on the antenna of the α1–3 or α1–6 mannose; Sx, number (×) of sialic acids linked to galactose. Structural schemes are given in terms of N-acetylglucosamine (square), mannose (circle), fucose (rhomb with a dot), galactose (rhomb) and sialic acid (star).

Supplementary table 4. Glycan structural features derived from the IgG glycome peaks. Derived IgG glycosylation traits were approximated from the ratios of original IgG glycan peaks (GP1-GP24, excluding GP3) sharing the same structural characteristics as indicate by the formulas.

GLYCOSYLATION FEATURE GROUPS	LABEL	STRUCTURAL FEATURE CODE	STRUCTURAL FEATURE DESCRIPTION	FORMULA
IgG CHARGED glycans (derived parameters)	IGG24	FGS/(FG+FGS)	*The percentage of sialylation of fucosylated galactosylated structures without bisecting GlcNAc in total IgG glycans*	*SUM(GP16 + GP18 + GP23) / SUM(GP16 + GP18 + GP23 + GP8 + GP9 + GP14) * 100*
	IGG25	FBGS/(FBG+FBGS)	*The percentage of sialylation of fucosylated galactosylated structures with bisecting GlcNAc in total IgG glycans*	*SUM(GP19 + GP24) / SUM(GP19 + GP24 + GP10 + GP11 + GP15) * 100*
	IGG26	FGS/(F+FG+FGS)	*The percentage of sialylation of all fucosylated structures without bisecting GlcNAc in total IgG glycans*	*SUM(GP16 + GP18 + GP23) / SUM(GP16 + GP18 + GP23 + GP4 + GP8 + GP9 + GP14) * 100*
	IGG27	FBGS/(FB+FBG+FBGS)	*The percentage of sialylation of all fucosylated structures with bisecting GlcNAc in total IgG glycans*	*SUM(GP19 + GP24) / SUM(GP19 + GP24 + GP6 + GP10 + GP11 + GP15) * 100*
	IGG28	FG1S1/(FG1+FG1S1)	*The percentage of monosialylation of fucosylated monogalactosylated structures in total IgG glycans*	*GP16 / SUM(GP16 + GP8 + GP9) * 100*
	IGG29	FG2S1/(FG2+FG2S1+FG2S2)	*The percentage of monosialylation of fucosylated digalactosylated structures in total IgG glycans*	*GP18 / SUM(GP18 + GP14 + GP23) * 100*
	IGG30	FG2S2/(FG2+FG2S1+FG2S2)	*The percentage of disialylation of fucosylated digalactosylated structures in total IgG glycans*	*GP23 / SUM(GP23 + GP14 + GP18) * 100*
	IGG31	FBG2S1/(FBG2+FBG2S1+FBG2S2)	*The percentage of monosialylation of fucosylated digalactosylated structures with bisecting GlcNAc in total IgG glycans*	*GP19 / SUM(GP19 + GP15 + GP24) * 100*
	IGG32	FBG2S2/(FBG2+FBG2S1+FBG2S2)	*The percentage of disialylation of fucosylated digalactosylated structures with bisecting GlcNAc in total IgG glycans*	*GP24 / SUM(GP24 + GP15 + GP19) * 100*
	IGG33	$F^{total}S1/F^{total}S2$	*Ratio of all fucosylated (+/- bisecting GlyNAc) monosyalilated and disialylated structures in total IgG glycans*	*SUM(GP16 + GP18 + GP19) / SUM(GP23 + GP24)*
	IGG34	FS1/FS2	*Ratio of fucosylated (without bisecting GlcNAc) monosialylated and disialylated structures in total IgG glycans*	*SUM(GP16 + GP18) / GP23*
	IGG35	FBS1/FBS2	*Ratio of fucosylated (with bisecting GlcNAc) monosialylated and disialylated structures in total IgG glycans*	*GP19 / GP24*
	IGG36	FBS^{total}/FS^{total}	*Ratio of all fucosylated sialylated structures with and without bisecting GlcNAc*	*SUM(GP19 + GP24) / SUM(GP16 + GP18 + GP23)*
	IGG37	FBS1/FS1	*Ratio of fucosylated monosialylated structures with and without bisecting GlcNAc*	*GP19 / SUM(GP16 + GP18)*
	IGG38	FBS1/(FS1+FBS1)	*The incidence of bisecting GlcNAc in all fucosylated monosialylated structures in total IgG glycans*	*GP19 / SUM(GP16 + GP18 + GP19)*
	IGG39	FBS2/FS2	*Ratio of fucosylated disialylated structures with and without bisecting GlcNAc*	*GP24 / GP23*
	IGG40	FBS2/(FS2+FBS2)	*The incidence of bisecting GlcNAc in all fucosylated disialylated structures in total IgG glycans*	*GP24 / SUM(GP23 + GP24)*

GLYCOSYLATION FEATURE GROUPS	LABEL	STRUCTURAL FEATURE CODE	STRUCTURAL FEATURE DESCRIPTION	FORMULA
IgG NEUTRAL glycans	IGG41	$GP1^n$	*The percentage of GP1 glycan in total neutral IgG glycans (GP^n)*	*$GP1 / GP^n * 100$*
	IGG42	$GP2^n$	*The percentage of GP2 glycan in total neutral IgG glycans (GP^n)*	*$GP2 / GP^n * 100$*
	IGG43	$GP4^n$	*The percentage of GP4 glycan in total neutral IgG glycans (GP^n)*	*$GP4 / GP^n * 100$*
	IGG44	$GP5^n$	*The percentage of GP5 glycan in total neutral IgG glycans (GP^n)*	*$GP5 / GP^n * 100$*
	IGG45	$GP6^n$	*The percentage of GP6 glycan in total neutral IgG glycans (GP^n)*	*$GP6 / GP^n * 100$*
	IGG46	$GP7^n$	*The percentage of GP7 glycan in total neutral IgG glycans (GP^n)*	*$GP7 / GP^n * 100$*
	IGG47	$GP8^n$	*The percentage of GP8 glycan in total neutral IgG glycans (GP^n)*	*$GP8 / GP^n * 100$*
	IGG48	$GP9^n$	*The percentage of GP9 glycan in total neutral IgG glycans (GP^n)*	*$GP9 / GP^n * 100$*
	IGG49	$GP10^n$	*The percentage of GP10 glycan in total neutral IgG glycans (GP^n)*	*$GP10 / GP^n * 100$*
	IGG50	$GP11^n$	*The percentage of GP11 glycan in total neutral IgG glycans (GP^n)*	*$GP11 / GP^n * 100$*
	IGG51	$GP12^n$	*The percentage of GP12 glycan in total neutral IgG glycans (GP^n)*	*$GP12 / GP^n * 100$*
	IGG52	$GP13^n$	*The percentage of GP13 glycan in total neutral IgG glycans (GP^n)*	*$GP13 / GP^n * 100$*
	IGG53	$GP14^n$	*The percentage of GP14 glycan in total neutral IgG glycans (GP^n)*	*$GP14 / GP^n * 100$*
	IGG54	$GP15^n$	*The percentage of GP15 glycan in total neutral IgG glycans (GP^n)*	*$GP15 / GP^n * 100$*
IgG NEUTRAL glycans (derived parameters)	IGG55	$G0^n$	*The percentage of agalactosylated structures in total neutral IgG glycans*	*$SUM(GP1^n: GP6^n)$*
	IGG56	$G1^n$	*The percentage of monogalactosylated structures in total neutral IgG glycans*	*$SUM(GP7^n: GP11^n)$*
	IGG57	$G2^n$	*The percentage of digalactosylated structures in total neutral IgG glycans*	*$SUM(GP12^n: GP15^n)$*
	IGG58	$F^{n\ total}$	*The percentage of all fucosylated (+/- bisecting GlcNAc) structures in total neutral IgG glycans*	*$SUM(GP1^n + GP4^n + GP5^n + GP6^n + GP8^n + GP9^n + GP10^n + GP11^n + GP14^n + GP15^n)$*
	IGG59	$FG0^{n\ total}/G0^n$	*The percentage of fucosylation of agalactosylated structures*	*$SUM(GP1^n + GP4^n + GP5^n + GP6^n) / G0^n * 100$*
	IGG60	$FG1^{n\ total}/G1^n$	*The percentage of fucosylation of monogalactosylated structures*	*$SUM(GP8^n + GP9^n + GP10^n + GP11^n) / G1^n * 100$*
	IGG61	$FG2^{n\ total}/G2^n$	*The percentage of fucosylation of digalactosylated structures*	*$SUM(GP14^n + GP15) / G2^n * 100$*
	IGG62	F^n	*The percentage of fucosylated (without bisecting GlcNAc) structures in total neutral IgG glycans*	*$SUM(GP1^n + GP4^n + GP5^n + GP8^n + GP9^n + GP14^n)$*
	IGG63	$FG0^n/G0^n$	*The percentage of fucosylation (without bisecting GlcNAc) of agalactosylated structures*	*$SUM(GP1^n + GP4^n + GP5^n) / G0^n * 100$*
	IGG64	$FG1^n/G1^n$	*The percentage of fucosylation (without bisecting GlcNAc) of monogalactosylated structures*	*$SUM(GP8^n + GP9^n) / G1^n * 100$*

GLYCOSYLATION FEATURE GROUPS	LABEL	STRUCTURAL FEATURE CODE	STRUCTURAL FEATURE DESCRIPTION	FORMULA
	IGG65	$FG2^n/G2^n$	*The percentage of fucosylation (without bisecting GlcNAc) of digalactosylated structures*	$GP14^n / G2^n * 100$
	IGG66	FB^n	*The percentage of fucosylated (with bisecting GlcNAc) structures in total neutral IgG glycans*	$SUM(GP6^n + GP10^n + GP11^n + GP15^n)$
	IGG67	$FBG0^n/G0^n$	*The percentage of fucosylation (with bisecting GlcNAc) of agalactosylated structures*	$GP6^n / G0^n * 100$
	IGG68	$FBG1^n/G1^n$	*The percentage of fucosylation (with bisecting GlcNAc) of monogalactosylated structures*	$SUM(GP10^n + GP11^n) / G1^n * 100$
	IGG69	$FBG2^n/G2^n$	*The percentage of fucosylation (with bisecting GlcNAc) of digalactosylated structures*	$GP15) / G2^n * 100$
	IGG70	FB^n/F^n	*Ratio of fucosylated structures with and without bisecting GlcNAc*	$FB^n / F^n * 100$
	IGG71	$FB^n/F^{n\ total}$	*The incidence of bisecting GlcNAc in all fucosylated structures in total neutral IgG glycans*	$FB^n / F^{n\ total} * 100$
	IGG72	$F^n/(B^n + FB^n)$	*Ratio of fucosylated non-bisecting GlcNAc structures and all structures with bisecting GlcNAc*	$F^n/(GP13^n + FB^n)$
	IGG73	$B^n/(F^n + FB^n)$	*Ratio of structures with bisecting GlcNAc and all fucosylated structures (+/- bisecting GlcNAc)*	$GP13^n / (F^n + FB^n) * 1000$
	IGG74	$FBG2^n/FG2^n$	*Ratio of fucosylated digalactosylated structures with and without bisecting GlcNAc*	$GP15^n/GP14^n$
	IGG75	$FBG2^n/(FG2^n + FBG2^n)$	*The incidence of bisecting GlcNAc in all fucosylated digalactosylated structures in total neutral IgG glycans*	$GP15^n/(GP14^n + GP15^n) * 100$
	IGG76	$FG2^n/(BG2^n + FBG2^n)$	*Ratio of fucosylated digalactosylated non-bisecting GlcNAc structures and all digalactosylated structures with bisecting GlcNAc*	$GP14^n/(GP13^n + GP15^n)$
	IGG77	$BG2^n/(FG2^n + FBG2^n)$	*Ratio of digalactosylated structures with bisecting GlcNAc and all fucosylated digalactosylated structures (+/- bisecting GlcNAc)*	$GP15^n/(GP14^n + GP15^n) * 1000$

Supplementary table 5. Correspondence between IgG and plasma glycan peaks. The IgG glycan peaks (GP1-GP24) are combined into 11 plasma glycan peaks (GP1-GP11). Notes about IgG glycan peaks: GP1 has no correspondence in plasma peaks, GP3 was excluded from all analyses as explained in the section 2.2.2, GP20 glycan structures were not determined (n.d.) and minor peaks designated with letters a and b sum up to a major peak (for instance, GP8a+GP8b=GP8).

IgG glycan peaks	Plasma glycans peaks
GP1	-
GP2	GP1
GP3	GP2
GP4	
GP5	GP3
GP6	
GP7	GP4
GP8a	
GP8b	
GP9	
GP10	
GP11	
GP12	GP5
GP13	
GP14	GP6
GP15	
GP16a	
GP16b	
GP17	GP7
GP18a	GP8
GP18b	
GP19	
GP20	GP9
GP21	
GP22	GP10
GP23	

GP24	GP11

Supplementary table 6. Genetic variants potentially associated with the diabetes condition. List of the top 30 SNPs identified by the correlation adjusted scores method as the most important for the diabetes data group division. Genes overlapping with the SNP are annotated without asterisk and neighbour genes of the SNP are annotated with asterisk.

SNP	Chr	Genes
rs7865906	9	NCS1*
rs1292123	6	CDK19*
rs9578030	13	LINC00398*
rs6563348	13	DCLK1
rs7865279	9	IDNK*
rs37984	7	AC006042.8; AC006465.3; GLCCI1
rs7232159	18	RP11-25O3.1*
rs2203586	2	AC092684.1
rs2150228	13	RNY4P29*
rs1373762	18	RP11-25O3.1*
rs12492596	3	AC104637.1*
rs7631551	3	FLT1P1
rs1328650	13	DCLK1*
rs7163551	15	RGMA
rs1926317	13	DCLK1
rs12605728	18	DCC
rs5961574	X	AC074035.1*
rs4659764	1	MT1HL1
rs802684	6	CDK19
rs1910780	12	RP11-955H22.2*
rs3924384	2	AC116609.1
rs3795366	1	SIPA1L2
rs1358725	6	RP1-60O19.1*
rs10897193	11	AP003733.1*
rs2691185	6	CDK19
rs1217770	5	MAP1B*
rs7202468	16	AC009158.1*
rs9954050	18	RP11-25O3.1*
rs7334245	13	DCLK1
rs11653470	17	AC005863.1

Supplementary table 7. Genetic variants most contributing to the genetic structure of populations. List of 35 SNPs consistently identified by the three SNP selection methods to be the most important for population discrimination within the top 100 SNPs. Genes overlapping with the SNP are annotated without asterisk and neighbour genes of the SNP are annotated with asterisk. The number in parenthesis following the name of the method indicates the rank position achieved by the SNP with that particular method. RJ: Random Jungle; CARE: correlation adjusted scores; DAPC: discriminat analysis of principal components.

SNP	Chr	Genes	Methods (rank)
rs1446585	2	R3HDM1	RJ(1); CARE(2); DAPC(1)
rs6730157	2	RAB3GAP1; ZRANB3	RJ(2); CARE(1); DAPC(2)
rs309160	2	DARS	RJ(3); CARE(8); DAPC(6)
rs313519	2	R3HDM1	RJ(4); CARE(14); DAPC(9)
rs313528	2	R3HDM1	RJ(5); CARE(13); DAPC(8)
rs932206	2	AC068492.1*	RJ(6); CARE(4); DAPC(4)
rs1561277	2	ZRANB3	RJ(7); CARE(3); DAPC(3)
rs621341	2	TMEM163	RJ(8); CARE(11); DAPC(15)
rs2011946	2	AC068492.1*	RJ(9); CARE(6); DAPC(7)
rs6739713	2	R3HDM1*	RJ(10); CARE(9); DAPC(11)
rs1469996	2	LCT; UBXN4	RJ(11); CARE(39); DAPC(17)
rs2071556	6	AL645941.1; AL662845.1; AL935042.1; BX088556.1; BX927138.1; CR752645.1; CR759798.1; CR936913.1; HLA-DMB; XXbac-BPG181M17.5	RJ(12); CARE(22); DAPC(21)
rs309137	2	AC093391.2	RJ(13); CARE(7); DAPC(5)
rs2322659	2	LCT	RJ(14); CARE(57); DAPC(12)
rs1869829	2	RAB3GAP1	RJ(15); CARE(20); DAPC(10)
rs3213943	2	R3HDM1	RJ(18); CARE(66); DAPC(20)
rs1042337	6	AL645941.1; AL662845.1; AL935042.1; BX088556.1; BX927138.1; CR752645.1; CR759798.1; CR936913.1; HLA-DMB; XXbac-BPG181M17.5	RJ(20); CARE(26); DAPC(33)
rs7950019	11	ST13P5	RJ(21); CARE(5); DAPC(28)
rs1035798	6	AGER; PBX2; RNF5	RJ(24); CARE(18); DAPC(26)
rs6430585	2	UBXN4	RJ(28); CARE(45); DAPC(18)
rs659445	6	C2; CYP21A2; EHMT2; ZBTB12	RJ(29); CARE(65); DAPC(45)
rs10008492	4	RNA5SP158*	RJ(30); CARE(15); DAPC(14)

SNP	Chr	Genes	Methods (rank)
rs494620	6	CYP21A2; SLC44A4	RJ(31); CARE(58); DAPC(30)
rs382259	6	XXbac-BPG154L12.4*	RJ(32); CARE(12); DAPC(13)
rs4331786	4	TLR10	RJ(37); CARE(33); DAPC(27)
rs9267833	6	NOTCH4	RJ(41); CARE(24); DAPC(63)
rs1123848	2	HNRNPKP2*	RJ(43); CARE(80); DAPC(19)
rs10024216	4	RNA5SP158	RJ(46); CARE(34); DAPC(22)
rs535586	6	CYP21A2; EHMT2	RJ(51); CARE(67); DAPC(46)
rs13296013	9	RPS10P3	RJ(59); CARE(32); DAPC(58)
rs1319281	13	RN7SKP2*	RJ(61); CARE(64); DAPC(84)
rs10496746	2	RN7SKP141*	RJ(62); CARE(99); DAPC(24)
rs2045272	11	ST13P5*	RJ(75); CARE(17); DAPC(49)
rs185819	6	5S_rRNA; RNA5SP206; TNXB	RJ(82); CARE(60); DAPC(44)
rs13149231	4	KLF3*	RJ(84); CARE(10); DAPC(35)

Supplementary table . Genetic variants associated with plasma N-glycan traits. List of SNPs consistently identified by the three SNP selection methods to be associated with each glycan trait within the top 100 SNPs. SNPs which were first-ranked by two of the methods are highlighted in bold. Genes overlapping with the SNP are annotated without asterisk and neighbour genes of the SNP are annotated with asterisk; n.d. means that the annotation for that SNP could not be fetched from Ensembl database. The number in parenthesis following the name of the method indicates the rank position achieved by the SNP with that particular method. RJ: Random Jungle; CARE: correlation adjusted scores; GEMMA: bayesian sparse linear mixed model.

Trait	SNP	Chr	Genes	Methods (rank)
GP1	**rs6573604**	**14**	**CTD-2509G16.5**	**RJ(46); CARE(1); GEMMA(1)**
GP3	rs946808	9	RP11-375O18.2	RJ(6); CARE(23); GEMMA(2)
GP5	rs1530057	3	RBMS3	RJ(78); CARE(7); GEMMA(75)
GP6	**rs9901675**	**17**	**SNORD10; AC113189.5; SNORA67; MPDU1;CD68; EIF4A1; SENP3-EIF4A1**	**RJ(17); CARE(1); GEMMA(1)**
GP7	rs12362065	11	OR10W1*	RJ(74); CARE(87); GEMMA(5)
GP8	rs6725841	2	LINC00299;AC007464.1	RJ(4); CARE(1); GEMMA(5)
GP8	rs10484427	6	RP11-254A17.1*	RJ(10); CARE(2); GEMMA(2)
GP9	rs4487196	3	RPL21P41*	RJ(65); CARE(2); GEMMA(97)
GP9	rs3734087	5	NUDT12	RJ(74); CARE(44); GEMMA(84)
GP10	rs10483776	14	FUT8	RJ(3); CARE(1); GEMMA(3)
GP10	rs4756899	11	USH1C;OTOG	RJ(9); CARE(3); GEMMA(1)
GP10	rs174627	11	FADS3;FADS2	RJ(71); CARE(2); GEMMA(2)
GP11	rs7137203	12	AC139931.1*	RJ(2); CARE(1); GEMMA(1)
GP11	rs4414724	2	LDHAP3*	RJ(85); CARE(29); GEMMA(15)
GP12	rs1281121	4	SH3TC1	RJ(61); CARE(24); GEMMA(4)
GP13	rs13107325	4	SLC39A8	RJ(5); CARE(2); GEMMA(6)
GP14	**rs3760776**	**19**	**FUT6;FUT3**	**RJ(1); CARE(1); GEMMA(2)**
GP14	rs1974491	17	BRIP1*	RJ(17); CARE(2); GEMMA(1)
GP15	rs10812830	9	LINGO2	RJ(27); CARE(8); GEMMA(27)
GP15	rs10743152	11	TH;MIR4686	RJ(79); CARE(1); GEMMA(3)
GP16	rs1569785	22	RP1-293L6.1*	RJ(24); CARE(3); GEMMA(59)
DG1	rs11621121	14	MIR4708*	RJ(2); CARE(5); GEMMA(1)
DG1	rs10132229	14	CTD-2509G16.5	RJ(7); CARE(4); GEMMA(2)
DG2	rs1412990	9	PIP5K1B	RJ(56); CARE(61); GEMMA(13)
DG3	rs4567889	2	ALK	RJ(57); CARE(5); GEMMA(2)

Trait	SNP	Chr	Genes	Methods (rank)
DG4	rs2980542	8	RGS22	RJ(31); CARE(2); GEMMA(3)
DG4	rs1995536	8	CSMD1	RJ(93); CARE(77); GEMMA(80)
DG5	rs1506869	8	DOCK5	RJ(89); CARE(25); GEMMA(69)
DG7	rs315081	1	ST6GALNAC3	RJ(17); CARE(2); GEMMA(3)
DG7	rs3760776	19	FUT6;FUT3	RJ(52); CARE(1); GEMMA(2)
DG7	rs4569731	4	GALNTL6	RJ(72); CARE(86); GEMMA(66)
DG8	rs2446440	8	LINC00967*	RJ(4); CARE(3); GEMMA(1)
DG8	rs1328514	9	AL353707.1*	RJ(29); CARE(6); GEMMA(37)
DG8	rs2472867	6	FARS2	RJ(51); CARE(2); GEMMA(2)
DG8	rs12926250	16	PMFBP1	RJ(54); CARE(13); GEMMA(16)
DG9	rs3760776	19	FUT6;FUT3	RJ(1); CARE(1); GEMMA(2)
DG9	rs1150975	12	RP11-428G5.1	RJ(38); CARE(7); GEMMA(3)
DG9	rs2650000	12	HNF1A-AS1*	RJ(45); CARE(6); GEMMA(1)
DG10	rs3135363	6	BTNL2*	RJ(3); CARE(3); GEMMA(1)
DG11	**rs13203024**	**6**	**NUS1***	**RJ(1); CARE(1); GEMMA(3)**
DG11	rs729724	10	WARS2P1*	RJ(11); CARE(7); GEMMA(5)
DG12	rs3760776	19	FUT6;FUT3	RJ(16); CARE(1); GEMMA(1)
Monosialo	rs10514990	17	CA10	RJ(37); CARE(73); GEMMA(4)
Disialo	rs9847446	3	RP11-231E6.1*	RJ(3); CARE(7); GEMMA(3)
Disialo	rs759602	3	ST6GAL1	RJ(38); CARE(46); GEMMA(92)
Trisialo	rs248230	5	RNF130	RJ(19); CARE(49); GEMMA(34)
Trisialo	rs10211505	2	AC012671.2*	RJ(33); CARE(51); GEMMA(24)
BAMS	rs718858	3	AGTR1	RJ(46); CARE(3); GEMMA(53)
BADS	rs9808120	2	RP11-111J6.2*	RJ(6); CARE(49); GEMMA(71)
BADS	rs11701048	21	CBS	RJ(66); CARE(26); GEMMA(3)
BA	rs1486536	11	RP11-179A10.1*	RJ(83); CARE(95); GEMMA(64)
TRIA	rs2235959	14	FLRT2	RJ(52); CARE(3); GEMMA(1)
C.FUC	rs12702696	7	ICA1;AC006042.6	RJ(41); CARE(52); GEMMA(33)
A.FUC	**rs3760776**	**19**	**FUT6;FUT3**	**RJ(1); CARE(1); GEMMA(2)**
A.FUC	rs17078797	13	RP11-531P20.1*	RJ(2); CARE(91); GEMMA(71)
A.FUC	rs4899579	14	IFT43*	RJ(13); CARE(31); GEMMA(98)
A.FUC	rs4807826	19	RANBP3	RJ(39); CARE(40); GEMMA(51)
A.FUC	rs10795250	10	AKR1C5P	RJ(85); CARE(27); GEMMA(41)

Trait	SNP	Chr	Genes	Methods (rank)
A2	rs10132229	14	CTD-2509G16.5	RJ(7); CARE(3); GEMMA(66)
A2	rs7159888	14	CTD-2509G16.5	RJ(8); CARE(2); GEMMA(1)
A2	rs2305480	17	GSDMB	RJ(96); CARE(15); GEMMA(2)
G0	rs6782811	3	ITPR1	RJ(52); CARE(30); GEMMA(79)
G1	rs1045873	10	PRTFDC1	RJ(11); CARE(4); GEMMA(1)
G1	rs3133679	8	RGS22	RJ(12); CARE(1); GEMMA(2)
G1	rs7789699	7	PRKAG2	RJ(21); CARE(6); GEMMA(24)
G2	rs17735715	3	RP11-23D24.2; RNU6-901P	RJ(67); CARE(61); GEMMA(10)
G3	**rs13107325**	**4**	**SLC39A8**	**RJ(6); CARE(1); GEMMA(1)**
G3	rs469523	5	DCP2*	RJ(32); CARE(6); GEMMA(65)
G3	rs4827341	X	SRPX; RP13-43E11.1; TM4SF2	RJ(39); CARE(30); GEMMA(63)
G4	**rs10743152**	**11**	**TH;MIR4686**	**RJ(10); CARE(1); GEMMA(1)**
G4	rs228376	X	DMD	RJ(77); CARE(30); GEMMA(4)

Supplementary table 9. Genetic variants associated with IgG N-glycan traits. List of SNPs consistently identified by the three SNP selection methods to be associated with each glycan trait within the top 100 SNPs. SNPs which were first-ranked by all methods are highlighted in bold. Genes overlapping with the SNP are annotated without asterisk and neighbour genes of the SNP are annotated with an asterisk. The number in parenthesis following the name of the method indicates the rank position achieved by the SNP with that particular method. RJ: Random Jungle; CARE: correlation adjusted scores; GEMMA: bayesian sparse linear mixed model.

Trait	SNP	Chr	Genes	Methods (rank)
IGG2	rs1269068	14	CTD-2509G16.5	RJ(3); CARE(4); GEMMA(1)
IGG3	rs3818593	9	B4GALT1	RJ(5); CARE(1); GEMMA(1)
IGG3	rs13121519	4	GRID2	RJ(42); CARE(9); GEMMA(85)
IGG4	rs6100044	20	VAPB	RJ(55); CARE(37); GEMMA(79)
IGG5	rs909674	22	MGAT3	RJ(2); CARE(1); GEMMA(2)
IGG6	rs1556463	9	PTPRD	RJ(69); CARE(45); GEMMA(68)
IGG7	rs4908037	1	AGL*	RJ(61); CARE(1); GEMMA(33)
IGG8	rs7570009	2	TMEM131	RJ(42); CARE(65); GEMMA(16)
IGG8	rs1218577	1	KCNN3	RJ(43); CARE(32); GEMMA(52)
IGG9	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(2)
IGG9	rs2292298	4	RELL1	RJ(6); CARE(3); GEMMA(1)
IGG10	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(5)
IGG10	rs4731214	7	POT1*	RJ(57); CARE(15); GEMMA(2)
IGG10	rs3136706	1	CD2	RJ(61); CARE(37); GEMMA(85)
IGG10	rs7232036	18	LINC00908*	RJ(93); CARE(21); GEMMA(54)
IGG11	rs7159888	14	CTD-2509G16.5	RJ(69); CARE(2); GEMMA(1)
IGG12	rs7146952	14	RP11-326E7.1*	RJ(38); CARE(16); GEMMA(46)
IGG13	rs3818593	9	B4GALT1	RJ(1); CARE(1); GEMMA(2)
IGG13	rs6764279	3	ST6GAL1	RJ(11); CARE(2); GEMMA(1)
IGG13	rs7897452	10	CACNB2*	RJ(18); CARE(30); GEMMA(10)
IGG13	rs2142661	22	RIBC2	RJ(22); CARE(23); GEMMA(74)
IGG14	rs9620326	22	SMARCB1	RJ(1); CARE(4); GEMMA(3)
IGG14	rs1539604	6	RP11-278J20.2*	RJ(33); CARE(31); GEMMA(5)
IGG14	rs6444193	3	ST6GAL1	RJ(52); CARE(2); GEMMA(15)
IGG15	**rs6764279**	**3**	**ST6GAL1**	**RJ(1); CARE(1); GEMMA(1)**
IGG15	rs6444193	3	ST6GAL1	RJ(2); CARE(2); GEMMA(3)

Trait	SNP	Chr	Genes	Methods (rank)
IGG15	rs154452	5	AC008592.5	RJ(47); CARE(31); GEMMA(14)
IGG17	rs2363447	4	SLC4A4*	RJ(3); CARE(4); GEMMA(2)
IGG17	rs6764279	3	ST6GAL1	RJ(5); CARE(2); GEMMA(1)
IGG17	rs1368304	5	HMGN1P16	RJ(13); CARE(5); GEMMA(18)
IGG17	rs16939284	8	RP11-706J10.1; ZFHX4-AS1	RJ(25); CARE(10); GEMMA(5)
IGG17	rs3818593	9	B4GALT1	RJ(30); CARE(1); GEMMA(3)
IGG20	rs7201219	16	GSG1L	RJ(94); CARE(5); GEMMA(5)
IGG22	rs6764279	3	ST6GAL1	RJ(47); CARE(1); GEMMA(1)
IGG22	rs4830793	X	FRMPD4	RJ(65); CARE(29); GEMMA(48)
IGG23	**rs6764279**	**3**	**ST6GAL1**	**RJ(1); CARE(1); GEMMA(1)**
IGG23	rs1174864	7	POM121L12*	RJ(11); CARE(91); GEMMA(59)
IGG24	**rs6764279**	**3**	**ST6GAL1**	**RJ(1); CARE(1); GEMMA(1)**
IGG24	rs1358295	2	RNU6-187P	RJ(36); CARE(81); GEMMA(99)
IGG25	rs4677611	3	FOXP1	RJ(2); CARE(38); GEMMA(32)
IGG25	rs6734537	2	KLF7*	RJ(77); CARE(25); GEMMA(1)
IGG26	**rs6764279**	**3**	**ST6GAL1**	**RJ(1); CARE(1); GEMMA(1)**
IGG26	rs3818593	9	B4GALT1	RJ(2); CARE(2); GEMMA(2)
IGG26	rs9405681	6	EXOC2*	RJ(12); CARE(32); GEMMA(73)
IGG27	rs2154637	8	KB-1615E4.2	RJ(22); CARE(2); GEMMA(1)
IGG28	**rs6764279**	**3**	**ST6GAL1**	**RJ(1); CARE(1); GEMMA(1)**
IGG28	rs6444193	3	ST6GAL1	RJ(2); CARE(2); GEMMA(2)
IGG28	rs935653	2	PRKCE	RJ(30); CARE(12); GEMMA(7)
IGG28	rs4940206	18	DCC	RJ(73); CARE(11); GEMMA(14)
IGG29	**rs6764279**	**3**	**ST6GAL1**	**RJ(1); CARE(1); GEMMA(1)**
IGG29	rs6444193	3	ST6GAL1	RJ(2); CARE(2); GEMMA(2)
IGG29	rs2725391	17	AZI1	RJ(5); CARE(6); GEMMA(3)
IGG29	rs8104096	19	CTC-265F19.2;GNG7	RJ(54); CARE(55); GEMMA(71)
IGG31	rs6764279	3	ST6GAL1	RJ(1); CARE(1); GEMMA(2)
IGG31	rs6687262	1	PSAT1P3*	RJ(2); CARE(2); GEMMA(1)
IGG31	rs4887970	16	WWOX	RJ(4); CARE(5); GEMMA(27)
IGG31	rs378268	5	RP11-158J3.2*	RJ(9); CARE(24); GEMMA(65)
IGG31	rs2279913	17	RP11-455O6.2; AZI1	RJ(22); CARE(3); GEMMA(3)
IGG32	**rs6764279**	**3**	**ST6GAL1**	**RJ(1); CARE(1); GEMMA(1)**

Trait	SNP	Chr	Genes	Methods (rank)
IGG32	rs6444193	3	ST6GAL1	RJ(2); CARE(2); GEMMA(2)
IGG33	rs1036585	3	BCHE*	RJ(22); CARE(3); GEMMA(1)
IGG33	rs1816658	8	LINC00966*	RJ(56); CARE(5); GEMMA(7)
IGG33	rs13266168	8	RP11-705O24.1*	RJ(91); CARE(4); GEMMA(2)
IGG34	rs1036585	3	BCHE*	RJ(11); CARE(7); GEMMA(38)
IGG34	rs13083341	3	BCHE*	RJ(51); CARE(8); GEMMA(93)
IGG35	**rs6764279**	**3**	**ST6GAL1**	**RJ(1); CARE(1); GEMMA(1)**
IGG35	rs3777179	5	ELL2	RJ(28); CARE(11); GEMMA(3)
IGG35	rs2149436	13	HTR2A*	RJ(85); CARE(10); GEMMA(8)
IGG36	rs10758192	9	B4GALT1	RJ(4); CARE(2); GEMMA(1)
IGG37	rs6764279	3	ST6GAL1	RJ(1); CARE(1); GEMMA(1)
IGG37	rs10758192	9	B4GALT1	RJ(10); CARE(3); GEMMA(2)
IGG38	rs3818593	9	B4GALT1	RJ(2); CARE(2); GEMMA(2)
IGG38	rs6764279	3	ST6GAL1	RJ(3); CARE(1); GEMMA(1)
IGG38	rs302740	1	RP5-896L10.1	RJ(92); CARE(7); GEMMA(4)
IGG39	rs909674	22	MGAT3	RJ(1); CARE(1); GEMMA(4)
IGG39	rs9620326	22	SMARCB1	RJ(7); CARE(8); GEMMA(3)
IGG39	rs3818593	9	B4GALT1	RJ(34); CARE(9); GEMMA(2)
IGG40	rs9620326	22	SMARCB1	RJ(2); CARE(7); GEMMA(2)
IGG40	rs5757659	22	TAB1	RJ(5); CARE(2); GEMMA(3)
IGG40	rs3818593	9	B4GALT1	RJ(10); CARE(9); GEMMA(1)
IGG42	rs10132229	14	CTD-2509G16.5	RJ(2); CARE(2); GEMMA(2)
IGG42	rs7159888	14	CTD-2509G16.5	RJ(7); CARE(3); GEMMA(1)
IGG43	rs1445779	5	FTH1P9	RJ(69); CARE(3); GEMMA(2)
IGG45	rs9620326	22	SMARCB1	RJ(5); CARE(6); GEMMA(3)
IGG45	rs5757659	22	TAB1	RJ(6); CARE(3); GEMMA(1)
IGG45	rs7573966	2	STRN	RJ(53); CARE(13); GEMMA(2)
IGG45	rs5757721	22	RPS19BP1*	RJ(83); CARE(11); GEMMA(37)
IGG46	rs3798174	6	SLC22A1	RJ(1); CARE(7); GEMMA(66)
IGG46	rs6573604	14	CTD-2509G16.5	RJ(9); CARE(2); GEMMA(1)
IGG46	rs11650354	17	TBX21	RJ(14); CARE(5); GEMMA(2)
IGG46	rs7789913	7	IKZF1	RJ(35); CARE(6); GEMMA(6)
IGG46	rs9285339	13	SLITRK6*	RJ(52); CARE(24); GEMMA(3)

Trait	SNP	Chr	Genes	Methods (rank)
IGG48	rs10151805	14	C14orf80*	RJ(66); CARE(1); GEMMA(49)
IGG49	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(3)
IGG49	rs909674	22	MGAT3	RJ(16); CARE(2); GEMMA(4)
IGG49	rs2185781	1	ADIPOR1	RJ(50); CARE(38); GEMMA(46)
IGG50	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(3)
IGG50	rs7232036	18	LINC00908*	RJ(86); CARE(8); GEMMA(7)
IGG50	rs7563350	2	PROM2*	RJ(90); CARE(69); GEMMA(45)
IGG51	rs11650354	17	TBX21	RJ(10); CARE(1); GEMMA(2)
IGG51	rs1341138	13	HSPD1P8	RJ(54); CARE(57); GEMMA(7)
IGG52	rs12256995	10	PPIAP31*	RJ(20); CARE(27); GEMMA(64)
IGG52	rs1028531	14	RP11-816J8.1*	RJ(79); CARE(1); GEMMA(1)
IGG53	rs10758192	9	B4GALT1	RJ(21); CARE(4); GEMMA(1)
IGG53	rs1998930	6	RP11-230C9.1*	RJ(58); CARE(5); GEMMA(8)
IGG53	rs10057083	5	CSNK1A1; CTB-89H12.4	RJ(97); CARE(28); GEMMA(73)
IGG54	rs10517927	4	SPOCK3	RJ(6); CARE(1); GEMMA(5)
IGG54	rs7857028	9	RNU6-996P*	RJ(62); CARE(17); GEMMA(61)
IGG56	rs441233	9	LINC00094*	RJ(18); CARE(1); GEMMA(13)
IGG56	rs5905956	X	RP11-342D14.1	RJ(93); CARE(3); GEMMA(6)
IGG57	**rs3818593**	**9**	**B4GALT1**	**RJ(1); CARE(1); GEMMA(1)**
IGG57	rs2861806	5	CTB-63M22.1*	RJ(84); CARE(11); GEMMA(30)
IGG58	rs7789913	7	IKZF1	RJ(13); CARE(20); GEMMA(11)
IGG58	rs7159888	14	CTD-2509G16.5	RJ(18); CARE(2); GEMMA(1)
IGG58	rs7079570	10	VSTM4	RJ(40); CARE(35); GEMMA(28)
IGG59	**rs6573604**	**14**	**CTD-2509G16.5**	**RJ(1); CARE(1); GEMMA(1)**
IGG59	rs8074094	17	ITGB3; ITGB3	RJ(9); CARE(17); GEMMA(2)
IGG59	rs7453920	6	HLA-DQB2	RJ(79); CARE(73); GEMMA(9)
IGG61	rs6573604	14	CTD-2509G16.5	RJ(11); CARE(2); GEMMA(1)
IGG61	rs11643717	16	LINC00311; CTC-786C10.2	RJ(44); CARE(53); GEMMA(14)
IGG62	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(2)
IGG62	rs909674	22	MGAT3	RJ(3); CARE(2); GEMMA(3)
IGG62	rs7789913	7	IKZF1	RJ(11); CARE(3); GEMMA(1)
IGG62	rs8102799	19	ZNF160	RJ(50); CARE(20); GEMMA(5)
IGG62	rs1859425	7	ZNF804B	RJ(78); CARE(47); GEMMA(55)

Trait	SNP	Chr	Genes	Methods (rank)
IGG63	rs9620326	22	SMARCB1	RJ(1); CARE(4); GEMMA(3)
IGG63	rs909674	22	MGAT3	RJ(3); CARE(1); GEMMA(4)
IGG63	rs7781977	7	IKZF1	RJ(7); CARE(6); GEMMA(1)
IGG63	rs1041350	9	SUMO2P2*	RJ(15); CARE(9); GEMMA(2)
IGG63	rs10483766	14	RHOJ	RJ(54); CARE(14); GEMMA(38)
IGG64	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(3)
IGG64	rs7781977	7	IKZF1	RJ(46); CARE(3); GEMMA(2)
IGG64	rs6570330	6	TXLNB*	RJ(56); CARE(11); GEMMA(1)
IGG65	rs2427032	20	CDH4	RJ(1); CARE(4); GEMMA(2)
IGG65	rs11643717	16	LINC00311; CTC-786C10.2	RJ(28); CARE(19); GEMMA(29)
IGG65	rs7159888	14	CTD-2509G16.5	RJ(93); CARE(6); GEMMA(1)
IGG66	**rs9620326**	**22**	**SMARCB1**	**RJ(1); CARE(1); GEMMA(1)**
IGG66	rs909674	22	MGAT3	RJ(2); CARE(2); GEMMA(2)
IGG66	rs5750811	22	TAB1	RJ(6); CARE(6); GEMMA(6)
IGG67	rs909674	22	MGAT3	RJ(2); CARE(1); GEMMA(4)
IGG67	rs9620326	22	SMARCB1	RJ(3); CARE(5); GEMMA(3)
IGG67	rs1041350	9	SUMO2P2*	RJ(9); CARE(8); GEMMA(2)
IGG68	**rs9620326**	**22**	**SMARCB1**	**RJ(1); CARE(1); GEMMA(1)**
IGG68	rs10506022	12	RP11-709A23.1; PPFIBP1	RJ(3); CARE(7); GEMMA(3)
IGG68	rs909674	22	MGAT3	RJ(6); CARE(2); GEMMA(2)
IGG68	rs7475361	10	SEPHS1	RJ(10); CARE(5); GEMMA(6)
IGG68	rs3802586	10	PHYH	RJ(95); CARE(18); GEMMA(45)
IGG69	rs9620326	22	SMARCB1	RJ(1); CARE(2); GEMMA(3)
IGG69	rs1159709	2	ERBB4	RJ(29); CARE(20); GEMMA(88)
IGG69	rs3818593	9	B4GALT1	RJ(49); CARE(3); GEMMA(2)
IGG70	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(2)
IGG70	rs909674	22	MGAT3	RJ(3); CARE(2); GEMMA(3)
IGG70	rs7789913	7	IKZF1	RJ(50); CARE(9); GEMMA(1)
IGG71	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(2)
IGG71	rs909674	22	MGAT3	RJ(2); CARE(2); GEMMA(3)
IGG71	rs1390156	13	TDRD3*	RJ(12); CARE(64); GEMMA(55)
IGG71	rs10139559	14	RP11-353P15.1*	RJ(17); CARE(33); GEMMA(72)
IGG71	rs31340	5	FSTL4	RJ(20); CARE(23); GEMMA(25)

Trait	SNP	Chr	Genes	Methods (rank)
IGG71	rs7789913	7	IKZF1	RJ(67); CARE(9); GEMMA(1)
IGG72	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(2)
IGG72	rs909674	22	MGAT3	RJ(2); CARE(2); GEMMA(3)
IGG72	rs7781977	7	IKZF1	RJ(14); CARE(7); GEMMA(1)
IGG72	rs1355925		n.d.	RJ(32); CARE(16); GEMMA(4)
IGG73	rs11954386	5	PARP8	RJ(71); CARE(58); GEMMA(20)
IGG74	rs9620326	22	SMARCB1	RJ(3); CARE(1); GEMMA(4)
IGG74	rs10758192	9	B4GALT1	RJ(8); CARE(2); GEMMA(3)
IGG75	rs9620326	22	SMARCB1	RJ(1); CARE(1); GEMMA(2)
IGG75	rs10758192	9	B4GALT1	RJ(2); CARE(2); GEMMA(1)
IGG75	rs31340	5	FSTL4	RJ(36); CARE(4); GEMMA(5)
IGG75	rs2092168	22	RPS19BP1*	RJ(70); CARE(5); GEMMA(3)
IGG76	rs9620326	22	SMARCB1	RJ(3); CARE(1); GEMMA(3)
IGG76	rs31340	5	FSTL4	RJ(11); CARE(3); GEMMA(2)
IGG76	rs7789913	7	IKZF1	RJ(54); CARE(21); GEMMA(8)
IGG76	rs3818593	9	B4GALT1	RJ(69); CARE(6); GEMMA(5)

Supplementary table 10. Genetic variants associated with phenotypes. List of SNPs consistently identified by the three SNP selection methods to be associated with each phenotype within the top 100 SNPs. Genes overlapping with the SNP are annotated without asterisk and neighbour genes of the SNP are annotated with asterisk. The number in parenthesis following the name of the method indicates the rank position achieved by the SNP with that particular method. RJ: Random Jungle; CARE: correlation adjusted scores; GEMMA: bayesian sparse linear mixed model.

Trait	**SNP**	**Chr**	**Genes**	**Methods (rank)**
Sys	rs10485097	6	PPIL4	RJ(9); CARE(51); GEMMA(22)
Sys	rs7001273	8	RP11-628E19.4*	RJ(56); CARE(24); GEMMA(46)
Sys	rs10507382	13	PAN3;FLT1	RJ(65); CARE(3); GEMMA(7)
HDL	rs995538	3	CPNE4	RJ(7); CARE(42); GEMMA(21)
Trigy	rs2131905	1	AKNAD1	RJ(2); CARE(46); GEMMA(59)
Trigy	rs159382	5	CTD-2176I21.2*	RJ(45); CARE(1); GEMMA(3)
Insulin	rs10026220	4	PI4K2B	RJ(2); CARE(4); GEMMA(64)
Insulin	rs965972	1	RP11-452J13.1*	RJ(16); CARE(2); GEMMA(3)
Insulin	rs6679047	1	AL450244.1*	RJ(49); CARE(12); GEMMA(5)
Calcium	rs7914270	10	WAPAL;RP11-77P6.2	RJ(16); CARE(3); GEMMA(4)
UricAcid	rs1014290	4	SLC2A9	RJ(35); CARE(1); GEMMA(1)