Analiza N-glikoma imunoglobulina G u blizanaca s križoboljom

Vučenović, Dunja

University of Zagreb
Faculty of Science
Department of Biology

Dunja Vučenović

Glycomic phenotyping of twins with lumbago

Graduation Thesis

Zagreb 2015

Ovaj rad, izrađen u Zavodu za istraživanje blizanaca pri King's College London, UK,
pod vodstvom prof.dr.sc. Frances MK Williams,
predan je na ocjenu Biološkom odsjeku Prirodoslovno-matematičkog fakulteta
Sveučilišta u Zagrebu radi stjecanja zvanja
***Magistar molekularne biologije**.*

This work, completed at the Department of Twin Research, King's College London
under the supervision of Frances MK Williams PhD FRCP(E)
was submitted for assessment to Department of Biology, Faculty of Science
at University of Zagreb in order to acquire
***Master's degree in molecular biology**.*

Acknowledgements

This Thesis would not have been possible without the guidance and the help of people who in one way or another contributed to this work, for which I am truly grateful.

First and foremost, special thanks to my supervisor Frances Williams whose help and encouragement I will never forget. I would like to thank her for giving me the opportunity to join Department of Twin Research during my thesis work. Special thanks to Stella and Max for helping me with all the obstacles I encountered during the research.

I thank Professor Gordan Lauc for arranging the internship at DTR and giving me the opportunity to learn a lot about exciting new scientific disciplines.

Special thanks to all my friends for being by my side for the last five years and for making them especially enjoyable.

Last but not least, I would like to thank my family– without them and their selfless support, understanding, encouragement and love during my education path, none of this would be possible.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Diplomski rad

Analiza *N*-glikoma imunoglobulina G u blizanaca s križoboljom

Dunja Vučenović

Horvatovac 102a,
10000 Zagreb

Križobolja je sveprisutan mišićno-skeletni sindrom koji se javlja u svim dobnim skupinama. Detaljne informacije o dijagnozi ovog sindroma potrebne su zbog njegove kliničke i socijalne važnosti. Trenutno postoje ograničeni biomarkeri (uglavnom su to snimke dobivene magnetskom rezonancijom) zbog čega je potrebno razviti nove, jednostavnije metode rane detekcije križobolje. Jedan od glavnih uzroka kronične križobolje jest lokalizirana upala u epiduralnom prostoru. Kako je svaki upalni proces povezan s brojnim promjenama na mjestu upale, tako su uočene i promjene u glikanima vezanim za Fc regiju imunoglobulina G (IgG).
U ovom radu nastojala sam odrediti postoje li različiti profili količine glikana kod blizanca iz TwinsUK baze podataka s križoboljom i zdravih pojedinaca. Da bih to ostvarila, analizirala sam *N*-glikom IgG-a 4416 blizanaca koji su odgovorili na upitnike o povijesti boli, čiji je glikanski profil određen i koji su podvrgnuti snimanju kralježnice magnetskom rezonancijom. S dobivenim podatcima sam, koristeći statističke metode za analizu blizanaca, tražila asocijaciju *N*-glikana s fenotipovima boli. Dodatno sam koristila i metode koje se koriste u genomici za određivanje razlike u ekspresiji gena kod pacijenata i kontrola te tako dobila module različitih količina *N*-glikana. Dobivenim modulima kreirala sam prediktivne algoritme koji govore o vjerojatnosti razvoja križobolje kod novih ispitanika.

(42 stranice, 15 slika, 7 tablica, 60 literaturnih navoda, jezik izvornika: Engleski)
Rad je pohranjen u Središnjoj biološkoj knjižnici.
Ključne riječi: Križobolja, glikani, regresija, mješoviti modeli, WGCNA, LASSO
Voditelj: dr. sc. Frances Williams, prof.
Suvoditelj: dr. sc. Vlatka Zoldoš, izv. prof.
Ocjenitelji: dr. sc. Vlatka Zoldoš, izv. prof
dr.sc. Vesna Benković, izv.prof
dr. sc. Silvija Černi, doc
Zamjena: dr. sc. Kristian Vlahoviček, prof.

Rad prihvaćen: 3. lipnja 2015.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Department of Biology

Graduation Thesis

Glycomic phenotyping of twins with lumbago

Dunja Vučenović

Horvatovac 102a,
10000 Zagreb

Lumbago is a common musculoskeletal condition in all ages. Owing to its clinical and social impact, a clear diagnosis of this syndrome is needed in order to define besides pain intensity also the interference of pain with daily and work activity. Currently, there are limited biomarkers (mostly imaging), but there is a need for novel and simpler detection methods. One of the major determinants of pain in persistent lumbago is localised inflammation in epidural space. As every inflammatory process is linked with changes on the inflamed region, so are altered *N*-glycan IgG compositions. In this study, using classical twin design conducted between twins from TwinsUK registry, I assessed whether twins reporting episodes of lumbago had detectable levels of altered IgG glycosylation. I analysed IgG *N*-glycome of 4416 twins having completed pain history questionnaires, whose glycan profile was determined and who underwent magnetic-resonance spine scan. With this data, using statistical methods for twin studies, I looked for association between glycan quantities and pain phenotypes. Additionally, I used methods applied in genomics and obtained modules of glycans that show different patterns of glycan quantities. With these modules I built prediction algorithms that are giving likelihood of person reporting episodes of lumbago.

(42 pages, 15 figures, 7 tables, 60 references, original in: English)

Thesis deposited in the Central Biological Library.
Key words: Lumbago, glycans, regression, mixed models, WGCNA, LASSO
Supervisor: FRCP(E) Frances Williams, Prof
Cosupervisor: Dr. Vlatka Zoldoš, Assoc. Prof.
Reviewers: Dr. Vlatka Zoldoš, Assoc. Prof.
Dr. Vesna Benković, Assoc. Prof.
Dr. Silvija Černi, Asst. Prof.
Substitution: Dr. Kristian Vlahoviček, Prof.

Thesis accepted:.3rd June 2015

Contents

Abbreviations

BMI	Body mass index
CLBP	Chronic low back pain
CSUM	Cervical component extracted from MRI scans
CWP	Chronic widespread pain
DZ	Dizygotic
GU	Glucose unit
HILIC	Hydrophilic interaction liquid chromatography
LBP	Low back pain
LSUM	Lumbar component extracted from MRI scans
MRI	Magnetic resonance imaging
MZ	Monozygotic
WGCNA	Weighted gene correlation network analysis

1. Introduction

1.1. Biology and importance of lumbago

Low back pain (LBP) is a common musculoskeletal condition in all ages (Brooks, 2006). The lifetime prevalence of non-specific LBP may reach 80%, with the annual prevalence ranging between 25% and 60% in different ethnic groups (Louw et al. 2007; Andersson, 2015). Moreover, lumbago is the most common cause of disability in people who are 45 years old or less, causes 4% of people to change employment, and is a problem most severe in industrialized nations (Garofalo and Polatin, 1999). The impact of lumbago associated disability on work is significant, as an estimated 22% of chronic LBP patients are on some form of medical leave from work and another 11% work in a reduced capacity (Wynne-Jones, Dunn, and Main, 2008).

Pain is traditionally categorized as acute or chronic. Most individuals initially suffer from acute pain, indicating the pain was the result of an injury or damage (Geisser, et al. 2006). Chronic pain represents pain that has lasted at least 3 months (von Korff, 1999; Thorn, 2004). On occasion, chronic pain does not result from injury, but rather has an insidious gradual onset over time (Thorn, 2004).

It is common for chronic lumbago patients to have endured numerous types of treatment without success and have significantly altered their lives (Vasudevan, 1992). When the lifestyle changes become significant, some individuals become disabled. Disability can be understood as a significant inability to engage in meaningful and necessary activities in one's daily life (Battié and May, 2001). Such disability is not limited to back pain patients, as individuals may become disabled from other medical conditions or cognitive disabilities.

Lumbago is a diverse group of mixed pain syndromes with different molecular pathologies at different structural levels displaying similar clinical manifestations. In principle, LBP could be divided according to the cause of the pain into: discogenic pain, facet joint pain, sacroiliac joint pain, widespread pain and spinal stenosis. One of the major determinants of pain in persistent chronic LBP (CLBP) syndrome is localised inflammation in epidural space (both following surgery and without previous surgery at this level) (Broos and Aebi, 2008). There is some evidence that there is a correlation of inflammatory cell type expression in the epidural space and severity of CLBP, but actually the correlation between different expression of pro-inflammatory cytokines and severity of CBLP and why there is such huge inter-personal variability in this expression is yet to be investigated (Kraychete et al., 2010).

Owing to its clinical and social impact, a clear diagnosis of this syndrome is needed in order to define besides pain intensity also the interference of pain with daily and work activity. Currently, there are limited biomarkers (mostly imaging) or clinical findings that can be used objectively to help the physician in precise anatomic diagnosis leading to the safest and most cost-effective treatment for the patient.

1.1.1. Measuring pain

The most difficult aspect of studying pain is that it is a "private experience" (Geisser et al., 2003; Jensen and Karoly, 2001). As such, it is only possible to determine how much pain an individual perceives that he or she is experiencing, and not the true pain intensity. This reveals a primary problem with pain measurement: Individual pain ratings do not yield any useful information about the source or severity of the patient's pain problem. Investigators and clinicians are unable to determine that a patient with extreme pain is suffering from a problem that is any worse than a patient who is suffering from minimal pain. The differences may be rooted in how individuals interpret the painful sensation or what individuals use as the standard to which they relate current pain to past painful experiences, such as kidney stones, childbirth, or post-surgical pain. Thus, the assessment of pain may simply be a description by the patient of how this pain compares to other pain that he or she has experienced. This is further evidenced in studies that have demonstrated that the interpretation of pain is a combination of one's expectations and the actual sensory experience (Brown et al., 2008). However, it is unclear whether expectations about pain influence the actual experience of pain or simply how one rates his or her pain (Wager, 2005).

The assessment of pain is commonly conducted through asking the patient to categorize his or her pain or rate the pain on visual analogue scales, verbal rating scales, or numerical rating scales (Chapman et al., 1985; Jensen and Karoly, 2001). The meaning of the resulting ratings are impossible to determine, as there is no way to assess the reliability or validity of individuals' estimations of their pain levels (Turk and Melzack, 2001). Moreover, pain ratings have questionable utility, as they are hindered by variations between individuals on a variety of variables, such as experiences, situations, personality variables, psychosocial variables,

behavioural contingencies, and variations in sensitivity to pain (Turk and Melzack, 2001). As a result, studies rarely simply assess pain ratings in the absence of other variables.

1.2. Glycosylation is complex post-translational modification

Glycans, sugar chains attached to macromolecules, constitute the most abundant and diverse form of the post-translational modifications. All cell surface and secreted glycoproteins that contain appropriate sequences (Asn-X-Ser/Thr where X is any amino acid except proline) can potentially acquire N-linked oligosaccharides (*N*-glycans) while they travel through the endoplasmic reticulum and the Golgi compartments (Marino et al., 2012). Glycans can influence disease development in many syndromes such as cancer, disorders of glycosylation, rheumatoid arthritis and AIDS (Ohtsubo and Marth, 2006). Glycans are crucial for the proper functioning of immune system, as some of the most important interactions between the immune system and viruses and bacteria are mediated by protein-glycan interactions. Moreover, glycans are key in the recognition of non-self events and an altered glycome may lead to autoimmune disorders (Arnold et al., 2007). The biological functions of glycans go from basic structural roles to development, protein folding and immune response. Glycosylation is known to be affected by factors such as sugar nucleotide concentration, type of glyco-enzymes and their expression levels (Marino et al., 2012).

In comparison with total cell proteome, the glycome is estimated to be several orders of magnitude larger, depending on the species (Freeze, 2006). Progress towards describing and explaining the molecular basis of glycan function has been rather slow, partly due to technology restraints and partly due to the fact that the biosynthesis of glycans is not genetic-template-driven. Glycans are generally constructed from nine monosaccharide building blocks which connect to one another through glycosidic bond (Dube and Bertozzi, 2005) (Figure 1).

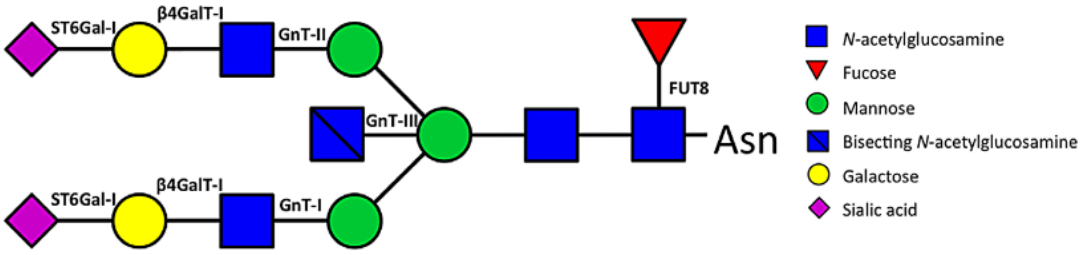

Figure 1 Example of glycan structure: Numerous glycan structures can be formed from nine monosaccharide building blocks which connect through glycosidic bond (Adapted from Lauc et al. 2013)

In eukaryotes there are 11 biosynthetic pathways that add glycans to proteins and lipids. There are two types of glycoproteins: *O*- and *N*- linked. *N*-linked protein glycosylation is the essential process for multicellular life, and its complete absence is embryonically lethal. *N*-glycan core has a fundamental role in glycoprotein function and it is therefore homogeneous and not subject to extensive variability or changes. On the other hand, variability of monosaccharides at the end of glycan antennae is common (e.g., ABO blood groups). This mechanism of glycan diversity enables adapting to changing environment, contributes to glycoproteome heterogeneity and can be advantageous for evading pathogens (Pučić *et al.*, 2010).

Enzymes that take part in the process of glycan synthesis are shared among different glycan structures. When we look at glycan structures, we can see how structurally similar they are. On the Figure 2, hierarchical representation of glycan structures is shown in a way that on the top of the dendogram are placed the simplest structures and then branching is done depending on the enzymes needed to synthesize each of the structures.

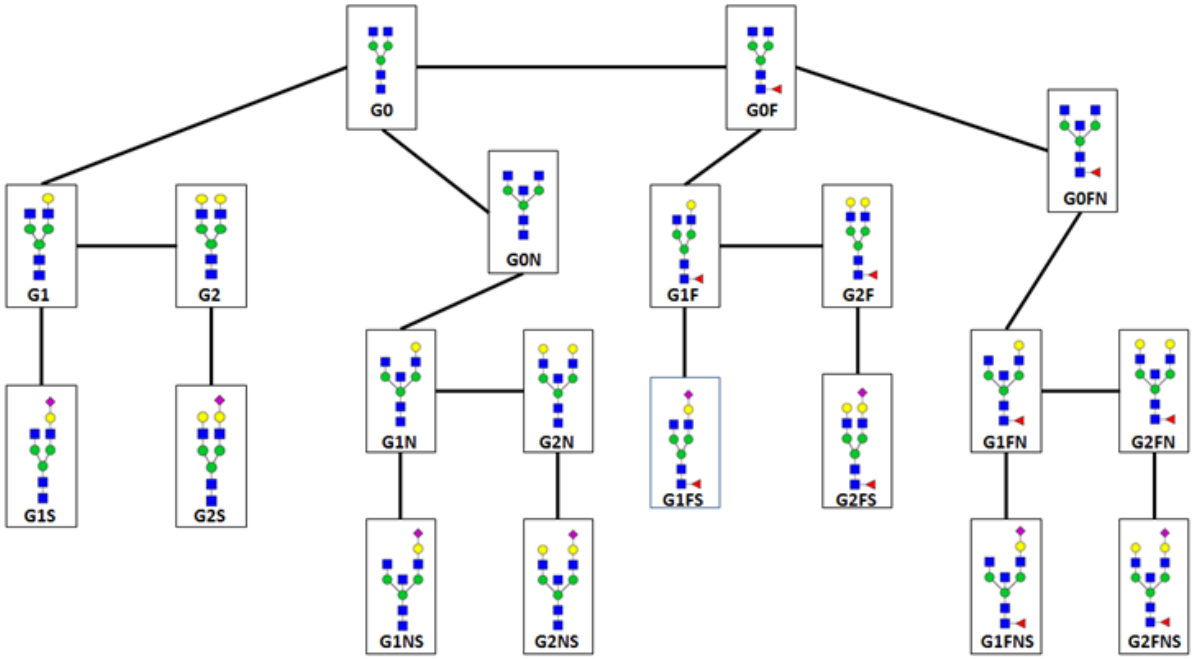

Figure 2 Structural hierarchy of glycans – biological clustering

1.2.1. Analysis of glycan composition of immunoglobulin G (IgG)

At the moment there is no "gold standard" method to analyse protein glycosylation with absolute precision, thus it is not possible to decide which of the methods most accurately reflect the real biological situation. An ideal technique that detects changes of glycan levels in a given environment would have to be fast, sensitive and able to integrate glycogene expression with glycan structural analysis (Alvarez-Manilla et al., 2007).

One of the most widely used method of analysis is high-performance liquid chromatography (HPLC) where glycans are fluorescently tagged and afterwards detected. The fluorescent tags most frequently used in this technique are 2-aminopyridine, 2-aminobenzamide and anthranilic acid (Royle *et al.*, 2008; Huffman et al., 2014). Fluorescent tagging-HPLC method for glycan analysis is sensitive enough and has ability to obtain quantitative data with good precision and reproducibility. However, this technique is time consuming, requires the availability of standards for every glycan to be identified and the resolution of this procedure is dependent on the HPLC or capillary electrophoresis column.

Studies have shown that it is possible to combine separation of oligosaccharides in a chromatography column and the mass spectrometry to increase the resolution, (e. g. HPLC-ESI-MS) (Alvarez-Manilla *et al*., 2007).

Another widely used method used to study glycans is mass spectrometry (MS) and its derivatives like matrix assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF-MS) and electrospray ionization mass spectrometry (ESI-MS). MALDI and ESI-MS can be used to record the spectra of intact glycoproteins, but these instruments are highly dependent on not only mass of the given protein but also on the degree of glycosylation. TOF instruments can generally detect small glycoproteins with a limited number of glycans attached. Glycoproteins with a high level of homogeneity can be identified with usage of ESI instruments designed for that purpose (Mills *et al*., 2003). Advantages of these mass spectrometric techniques are their speed, sensitivity, and high resolution, but they are still very limited for quantifying glycans in glycoproteomic studies.

1.3. Twin studies

Twin studies proved to be an invaluable tool in genetic epidemiology. Twins are two offspring resulting from the same pregnancy born in close succession. They may be either monozygotic (MZ) or dizygotic (DZ). The rate of MZ twinning is relatively stable, occurring in approximately four pregnancies out of 1000 across countries (Blickstein and Keith, 2005). DZ twinning rates, by contrast, vary by geographical region; in Asia about 6 in 1000, in Europe and USA about 10-20 in 1000 and in Africa about 40 in 1000 pregnancies are DZ twin pregnancies (Hall, 2003). The tendency to give birth to DZ twins is inherited and increases with maternal age and use of fertility drugs or in vitro fertilization procedures. About one-third of twins born are MZ, one-third are DZ same-sex (DZSS), and the remaining one-third are DZ opposite-sex (DZOS) (Blickstein and Keith, 2005). DZ twins almost invariably have their own placentas and are dichorionic and diamnionic (DC-DA) in placental membrane structure (Hall, 2003).

1.3.1. Importance of twin studies

Quantitative genetic analyses examine the nature of individual differences as well as similarities between family members and other relatives. In order to address the question regarding genetic and environmental influences on dissimilarities, the variance of a trait is studied. To do so it is necessary to perform studies on subjects with different degrees of genetic and environmental relationships. The twin-design is, therefore, of great use for genetic studies as same sex twin pairs are sampled from the same gene pool and they share same genes, although to a different degree. MZ twins share in principal all of their genes, and DZ twins share, on average, half of their segregating genes such as ordinary full siblings, but unlike ordinary full siblings, twins are matched on age and in most registries, sex. The classical twin study aims to explain the inter-individual variation in a trait. Studying twins offers a unique opportunity to determine the genetic and environmental effects on multifactorial polygenic traits such as body weight and habitual dietary intake. Genetic effects may arise from cumulative effects of multiple genes (additive genetic effects), or because of interaction between the alleles of these genes (dominance genetic effects). Environmental differences may arise from the environment unique to the individual (non-shared environment) making the twins within a pair less alike or from the environment common to co-twins (shared environment) making the twins alike. The greater similarity of MZ than DZ twin pairs is regarded to result from genetic effects as there is good evidence that MZ and DZ share twins share an environment to the same degree (Tan, 2010). This is the basic principle of twin methodology. The classic twin study compares phenotypic resemblances of MZ and DZ twins. Comparing the resemblance of MZ twins for a trait with the resemblance of DZ twins offers the first estimate of the extent to which genetic variation determines phenotypic variation of the trait. If MZ twins resemble each other more than do DZ twins, then the narrow heritability (h^2) of the phenotype can be estimated from twice the difference between MZ and DZ correlations. The proportion of the variance that is due to the environment is the difference between the total twin correlation and the part that is explained by heritability (Boomsma, 2002).

1.3.2. TwinsUK database

The UK Adult Twin Registry is a cohort of volunteer adult twins that has evolved rapidly since its inception in 1992. Originally, several hundred adult female twins were recruited by media campaign to allow investigation of osteoporosis and osteoarthritis, conditions with high prevalence in women. The success of these early studies, and the realisation that many traits hitherto considered environmental in aetiology could be investigated, led to expansion of the collection and the inclusion of males. The registry now incorporates twins from the Aberdeen Twin registry and Institute of Psychiatry Adult Registry. The cohort is one of the most highly and deeply phenotyped in the world, and is being enriched by comprehensive genotyping. Today, the database includes approx. 13,000 twins aged 16–85 (mean age 48), with a ratio of MZ to DZ twins of approximately 50:50. Participants are sent regular questionnaires for completion and are also invited to attend clinical visits.

1.4. Aim of this study

In this study, by using classical twin design conducted between twins from TwinsUK registry, I aim to assess whether twins reporting episodes of LBP had detectable levels of IgG glycosylation that differed from those not reporting episodes of LBP. In the case of different levels of IgG glycosylation between cases and controls, I would be able to suggest glyco-profiles which are more or less likely to be associated with LBP.

It is hard to distinguish weather patients with chronic widespread pain (CWP) are also suffering from LBP (their back pain could be result of CWP and not of LBP) so in order to make the project specifically about back pain, I obtained information about twins having CWP and LBP, in further analyses I used only patients with LBP excluding those having CWP.

In addition to regression analysis of glycan composition in cases and controls, the goal of this study is to perform weighted gene correlation network analysis (WGCNA) on glycan data. WGCNA is widely used in genomics to define modules (clusters) and network nodes among gene expression (high-dimensional) data. This analysis hasn't been applied to a glycan data yet, but because this is highly similar high-dimensional data, it is possible to create correlation networks of glycan data.

Twin studies are invaluable tool in genetic epidemiology that tells about relative and absolute importance of genetic and environmental influences on a phenotype. Especially interesting are cases of discordant twin pairs where one of the siblings is affected by some disease and the other is not. I decided to do separate analysis of only discordant twins so I could, in more detail, find out about effects of gens and environment in development of low back pain.

After learning which glycan structures are significantly associated with disease outcome I made an algorithm that will be able to predict, based on IgG *N*-glycan quantities, disease status. Depending on the strength of association and number of associated glycans prediction accuracy will vary.

2. Materials and methods

2.1. Sample

The participants in the present study were a sample of MZ and DZ twins enlisted in the TwinsUK registry who had undergone height and weight measurements used to calculate BMI (Sambrook, 1999). Collection of socio-demographic, CWP and LBP data was carried out during clinical visit or via a postal self-completion questionnaire. In order not to influence twins' pain perception, they were unaware of the precise research hypothesis addressed in this study. LBP and CWP data were available on 1932 twins, comprising 224 full MZ pairs, 414 full DZ pairs, and 704 participants whose co-twins did not take part in study. For CWP association analysis, IgG glycome was analysed in 4416 twins. Both the CWP and the LBP phenotypes were defined as a binary traits based on questionnaire responses (e.g., 1 = affected and 0 = nonaffected). The modified version of London Fibromyalgia Epidemiology Study (LFES) questionnaire contained four questions about musculoskeletal pain lasting over a week in the upper limbs, the lower limbs and the thorax/neck/back, and two further questions about fatigue and its chronicity and severity (White et al., 1999). A diagnosis of CWP was made if respondents answered positively to all four pain questions and positively to either both a right and left side response or on both sides. Volunteers were considered as case if in any of the questions asked about back pain answered positive. Basic descriptive statistics of the dataset are shown in Table 1 and 2.

Table 1 Sample characteristics of LBP without CWP study participants

	Overall sample (n = 1932)			**Monozygotic twins (n = 801)**			**Dizygotic twins (n = 1131)**		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Age	55,66	11,77	17-83	54,54	13,04	17 - 83	56,45	10,71	18 - 82
BMI	26,52	5,01	15,7-55,18	26,08	4,82	15,7-52,71	26,84	5,1	16,9-55,18
		N	**%**	**N**	**%**		**N**	**%**	
Sex	Male	101	5,3%	45	5,71%		56	5%	
	Female	1831	94,7%	756	94,29%		1075	95%	
LBP without CWP	Cases	440	22,93	175	21,77%		265	23,75	
	Controls	1492	77,07	626	78,23%		866	76,25	

Table 2 Sample characteristics for CWP study participants

	Overall sample (n = 4416)			**Monozygotic twins (n = 1687)**			**Dizygotic twins (n = 2728)**		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Age	51,60	14,06	17.27 – 83,44	51,51	14,86	17,27 – 83,44	51,66	13,58	17,31 – 82,07
BMI	26,26	5,01	13,22 – 55,18	26,09	5,03	15,71 – 52,71	26,37	4,99	13,22 – 55,18
		N	**%**	**N**	**%**		**N**	**%**	
Sex	Male	336	7,6	145	8,6		191	7	
	Female	4080	92,4	1542	91,4		2537	93	
CWP	Cases	1289	29,20	441	26,14		848	31,09	
	Controls	3127	70,80	1246	73,86		1880	68,91	
CWP without MRI	Cases	895	20,27	329	19,95		566	20,75	
	Controls	3521	79,73	1358	80,05		2162	79,25	

The study was approved by the St Thomas' Hospital research Ethics Committee and all twins provided informed consent.

2.2. Assessment of lumbago

Twins participating in the spine study (1997-2000) had attended an assessment that included a nurse-led interview and a number of clinical and laboratory tests (Sambrook, 1999). As part of the study, the twins completed a standardized questionnaire relating to their lifetime history of low back and pain symptoms. The questionnaires were completed by each twin separately. The lumbago questionnaire followed the format of questions used in the Medical Research Council Nurses Study (Smedley et al., 1998). It included both written questions and a pain diagram allowing an assessment of the timing, distribution, radiation, severity, and duration of pain together with information relating to functional disability. Lumbago was defined on a mannequin as being located between the 12th rib and the gluteal folds. This analysis focused

on pain with a total duration of more than a month and associated with disability. Disability was defined as having resulted in any one of the following activities being impossible: walking around the house, standing for 15 minutes, getting up from a low chair, getting out of the bath, getting in and out of a car, going up and down the stairs, putting on socks and tights, and cutting toenails. So in future those that are cases should be referred to as having episodes of severe and disabling LBP.

2.3. Magnetic resonance imaging (MRI)

MRI was performed using a Siemens (Munich, Germany) 1.0T superconducting magnet. Sagittal images were obtained using a fast spin-echo sequence of time to recovery (TR)/time to echo (TE) 5000– 4500/112 msec, with a slice thickness of 4 mm. Grading was performed on T2-weighted images, although T1 images were also obtained for certain measurements. Axial sections were obtained at selected levels to assess structural changes in individuals who had features suggesting prolapse. To avoid problems related to diurnal variation in disc height all MRI scans were performed more than an hour after the subjects arose from sleep in the morning, with no exercise or other rest allowed between arising and the scan, and importantly, each twin pair was scanned at the same appointment and on the same machine (Paajanen et al., 1994). A disease severity score was constructed from the sum of scores for disc bulge, height, signal change, and narrowing in the lumbar and in the cervical spine.

2.4. Overview of experimental glycan analysis

The analysis I performed for this study were mainly computational but in order to understand how the data were obtained and its characteristics, I will briefly explain experimental methods. Experimental data was obtained from Department of Twin research, King's College, London, UK and it was analysed in the research group of Professor Gordan Lauc at Glycobiology group, Genos d.o.o, Zagreb, Croatia.

The IgG was isolated using protein G monolithic plates as described previously (Gornik et al., 2009). The *N*-glycans from IgG samples were released and labelled with 2-aminobenzamide (LudgerTag 2-AB labeling kit Ludger Ltd., Abingdon, U.K.). Labelled glycans were then subjected to hydrophilic interaction high performance liquid chromatography (HILIC).

Glycans were analysed on the basis of their elution positions and measured in glucose units (GU). The chromatograms obtained were all separated into 24 peaks and the amount of glycans in each peak was expressed as % of total integrated area. In addition to 24 directly measured glycan structures, 55 derived traits were calculated.

2.5. Pre-processing and filtering

Directly measured glycan levels were normalized and experimental noise was removed through filtering and batch correction. First, I filtered out most extreme values from the dataset (beyond 0.999% percentile). Then, I applied quotient normalization using median values across the dataset as a reference (Dieterle et al., 2006). Because samples were analysed in four batches and distributions of glycan quantities varied among batches, I did batch correction using ratio-based method with either geometric mean or median (Chen et al., 2011). As the results of the two corrections were almost equivalent, herewith, I report only the results for the dataset corrected with geometric mean.

After these steps, I estimated 55 derived glycan levels from the directly measured glycans (Huffman et al., 2014) using glycanr package for R. These derived traits average particular glycosylation features (galactosylation, fucosylation, sialylation) across different individual glycan structures and consequently they are more closely related to individual enzymatic activities and underlying genetic polymorphisms. Finally, I applied inverse transformation of ranks to normality to obtain standard Normal distribution using rntransform function from GenABEL package for R (Aulchenko et al., 2007).

After pre-processing I assessed the dependency between the glycan traits and confounders such as age, sex, and body-mass index (BMI). I found out that age is a complex cofounder and that it has piece-wise relationship with glycan levels which did not justify its inclusion in linear regression models as a confounder. Therefore, before further analysis I corrected glycan levels for age (through residuals) by segmented regression approach using 40-45 years as initial break points as implemented in segmented package for R (Muggeo, 2003). The choice of the break-down points was done based on the observation of the correlation clouds for age and glycans followed by a bootstrap based search for "true" breakpoints.

2.6. Regression analysis

Regression analysis is a statistical method for estimating the relationships among variables. Regression analysis helps understand how the typical value of the dependent variable (variable of interest) changes when any one of the independent variables (known, measured variables) is varied, while the other independent variables are held fixed. It is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

Many techniques for carrying out regression analysis have been developed. One of the most widely used method is linear regression which studies linear, additive relationships between variables. In linear regression, data is modelled using linear predictor functions, and unknown model parameters are estimated from the data. This method is extensively used because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. One of the important assumptions linear regression uses is that it requires each observation to be independent. That is that the data-points (in this case study participants) should not be from any dependent samples design, e.g., before-after measurements, or matched pairings. Also the model should have little or no multicollinearity (independent variables should be independent from each other).

How in my dataset many volunteers have their siblings, which are genetically either 100% identical or on average 50% identical, in the same database, I couldn't use linear regression to explain relationship between outcome and predictors because linear regression assumes independence between patients. I had to use generalized linear mixed models which are extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects. They are especially convenient when dealing with grouped data. In my case mixed models will distinguish between variability in glycan levels within a twin pair and between twin pairs. Variation in glycan levels within monozygotic twin pairs is taken as a random effect in regression equation because observed variation is caused mostly by environmental effect while variation in glycan levels between twin pairs or within dizygotic twins is caused not only by environmental effects, but also by genetic effects.

2.6.1. Mixed-models analysis of association between glycan levels and disease status

Because of the twin structure of the dataset, association analyses between disease status (LBP without CWP) and glycan traits were performed using a linear mixed model analysis (R programming language, lme4 package) with sex and BMI included as fixed covariate and variation in IgG glycan quantities within twin pairs as random effect. In order to get significance of association for the glycans, I analysed two different types of models: ones in which I tried to explain a twin glycan quantity by using their disease status, sex and BMI and the others in which I didn't use disease status (Figure 3). After comparing goodness of fit of both models I was able to calculate p-values which explained how well a specific glycan associates with disease status. The association was analysed for each glycan separately. False discovery rate was controlled by Sidak's correction for multiple testing with the significance level of 0.0027 (p-value calculation explained in 2.9. chapter). Covariates found to have a significant association with LBP without CWP were entered into a multivariable regression model and assessed for significance.

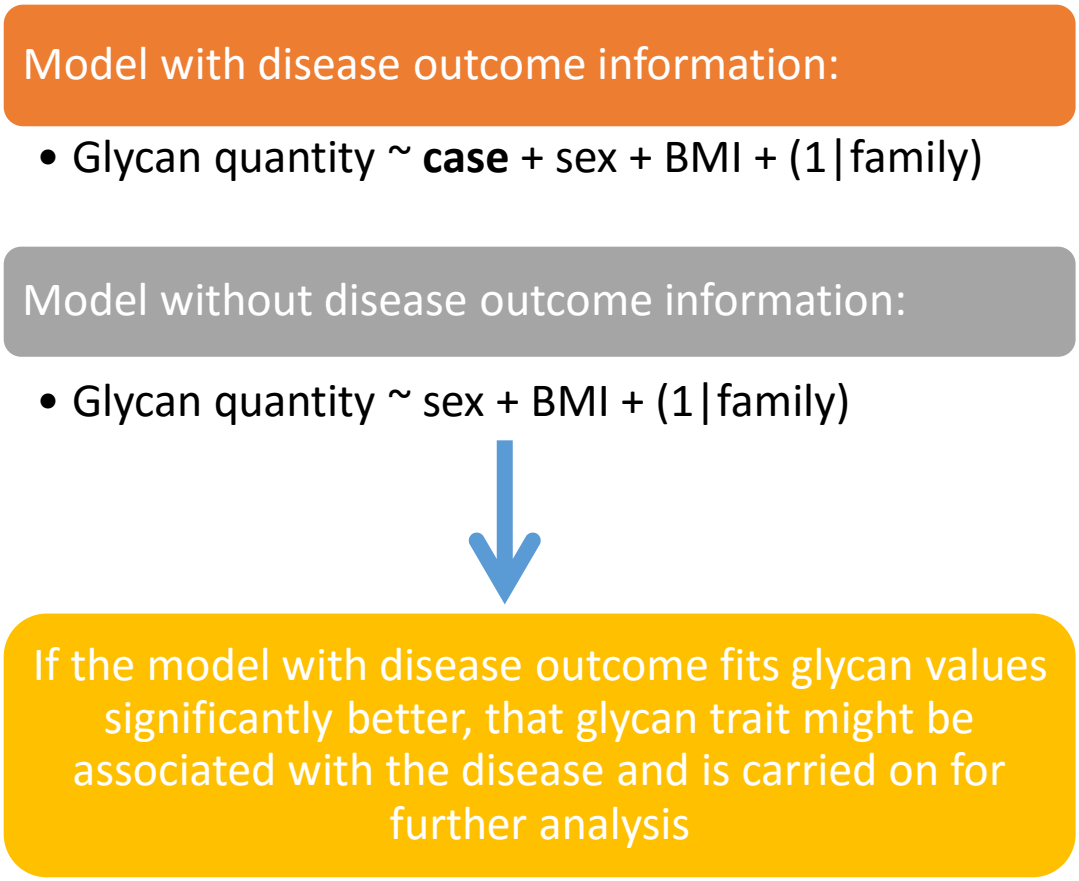

Figure 3 Schematic representation of association analysis. In the given equations (1|family) term signifies random effect term in linear mixed model while other terms after ~ sign are fixed effect terms.

Additionally, because of the bigger dataset and higher statistical power, I examined association of glycan quantities with CWP. The analyses were the same as previously described but outcomes of regressions were CWP – assessed only from questionnaires and CWP with additional MRI information about twin's possible spine defects. By comparing significantly associated glycans in models where LBP was outcome with models where CWP was outcome I concluded about altered glycan quantities in LBP.

In all data analysis in this study was analysed and visualized using R programming language (version 3.0.1).

2.7. Weighted glycan "expression" networks

To investigate if there exist patterns of similar amounts of some glycans among affected and unaffected twins, I performed weighted correlation network analysis (WGCNA). To perform that I used WGCNA package for R (Langfelder, Horvath, 2008; Langfelder, Horvath, 2012) which carries out an exploratory analysis of "network" dependencies between the glycan traits. The algorithm of the analysis is based on the estimation of correlations between the glycan levels across the dataset followed by extraction of relatively independent modules of correlated glycans. Glycan levels were adjusted for age, sex and BMI before the analysis. Signed network algorithms were used which takes into account the direction of the correlation between glycans. The modules (represented by their eigenvalue estimated as first principal component for the glycans in every module) were then correlated with the pain phenotypes, including chronic wide-spread pain (CWP), low back pain (LBP), LBP without CWP (LBP.N.CWP), and MRI traits: CSUM and LSUM. Module memberships is calculated for each glycan, which is a correlation between a glycan and module eigengenes (eigengenes because the approach was developed for gene expression). To estimate correlations between glycan modules and pain phenotypes I used point-biserial correlation coefficients and Pearson's correlation coefficients for qualitative and quantitative traits, respectively.

2.8. Discordant twins analysis

Discordant twin pairs are those where one of the siblings is considered as affected and the other acts as a control. These twin pairs are interesting because although they are genetically identical (MZ) or about 50% identical, and yet they may show very different phenotypes.

I compared the glycan levels in MZ and DZ twins discordant for pain phenotypes (CWP, LBP, LBP.N.CWP) using Wilcoxon's signed-rank test and mixed-models linear regression with twin relationship as random factor using *nlme* package for R (Pinheiro et al., 2015). Before Wilcoxon's test, glycan levels were adjusted for age, sex and BMI, while for mixed-models analysis, glycan levels were adjusted for age only, while sex and BMI were included into the models as fixed effects.

2.9. P-value consideration

There is an essential correlation between the glycan traits, many of which were derived from the original set of directly measured glycans. This complicates straightforward application of correction for multiple testing due to violation of the requirement for the independence of the tests. Taking this into account, I estimated the effective number of independent statistical tests as of 19 (Li and Ji, 2005), which after Sidak's correction for multiple testing provided the significance level of 0.0027.

2.10. Prediction of a disease status

As a final part of this study I made an algorithm for predicting disease status based on twins' glycan quantities and information about their sex and BMI. Before doing any kind of predictive model, I had to solve a problem of correlated predictor values (quantities of some glycans are often correlated) because correlated predictors add unnecessary robustness to a model. Also, if we are using too many predictors we could build a model that perfectly describes our dataset, but performs poorly with new subjects. In order to avoid robustness and prevent overfitting, I performed variable selection where I assessed the predictive potential of each glycan and kept only the most significant ones. In the case of correlated glycan quantities only one glycan from

the group was chosen to represent all others. I used LASSO method which is an innovative variable selection method for regression. It minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Important thing to note is that unlike other shrinkage methods, LASSO completely removes less significant variables from the equation while other methods usually just penalise them. This way LASSO uses fewer variables in model construction and tries to optimise robustness and accuracy ratio. LASSO not only helps to improve the prediction accuracy when dealing with multicolinearity data, but also carries several nice properties such as interpretability and numerical stability.

The whole dataset was split in two parts: 80% of the dataset was randomly assigned to training part and the rest into test part. An algorithm was built only from the training dataset and the validation of its predictive power was estimated on the test part. This process was repeated hundred times with different values for coefficients in LASSO algorithm which define how strictly variables (glycans) will be penalised and removed from equation.

For each iteration of the algorithm I calculated root mean squared error for the prediction and choose which parameters of the LASSO algorithm perform best when predicting on the test dataset. Depending on parameters chosen different numbers of variables and different variables will be used in regression models which will be used for prediction of likelihood of disease.

Additionally I performed the mixed model regression for each of the glycan modules obtained from WGCNA analysis (so that only glycans from that module are used) and built prediction algorithm from those models. Finally, I compared performance of each of the predictive models and concluded whether glycan modules can be useful for variable selection purposes.

3. Results

IgG glycome composition for LBP was analysed in 1932 twins (of those: 224 full MZ pairs, 414 full DZ pairs, and 704 singletons – unpaired twins). For CWP association analysis, IgG glycome was analysed in 4416 twins.

IgG glycosylation analysis was performed using a recently developed high-throughput analysis method (Cassidy, 1998) that reliably separates and individually quantifies nearly all IgG glycans. Distributions of the first 24 directly measured glycans divided by cases and controls for LBP without CWP can be seen on Figure 4.

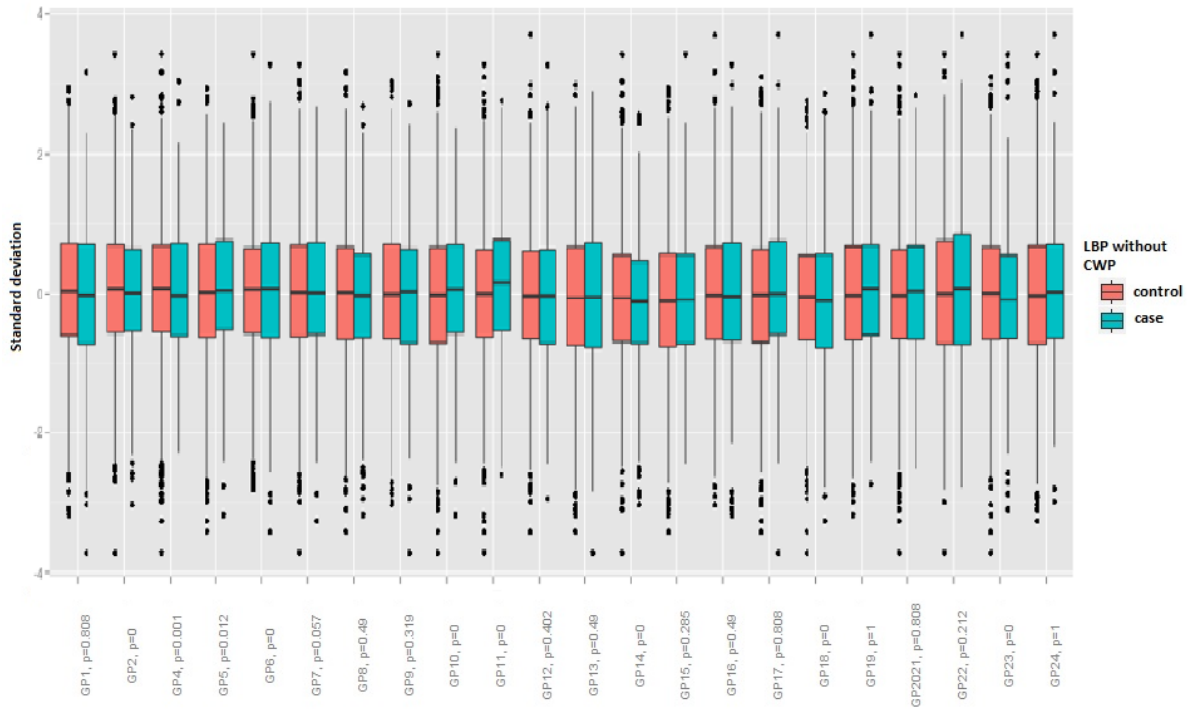

Figure 4 Distribution of glycan quantities between cases and controls for LBP without CWP phenotype

Distributions of first 24 directly measured glycans divided between cases and controls for CWP can be seen on Figure 5.

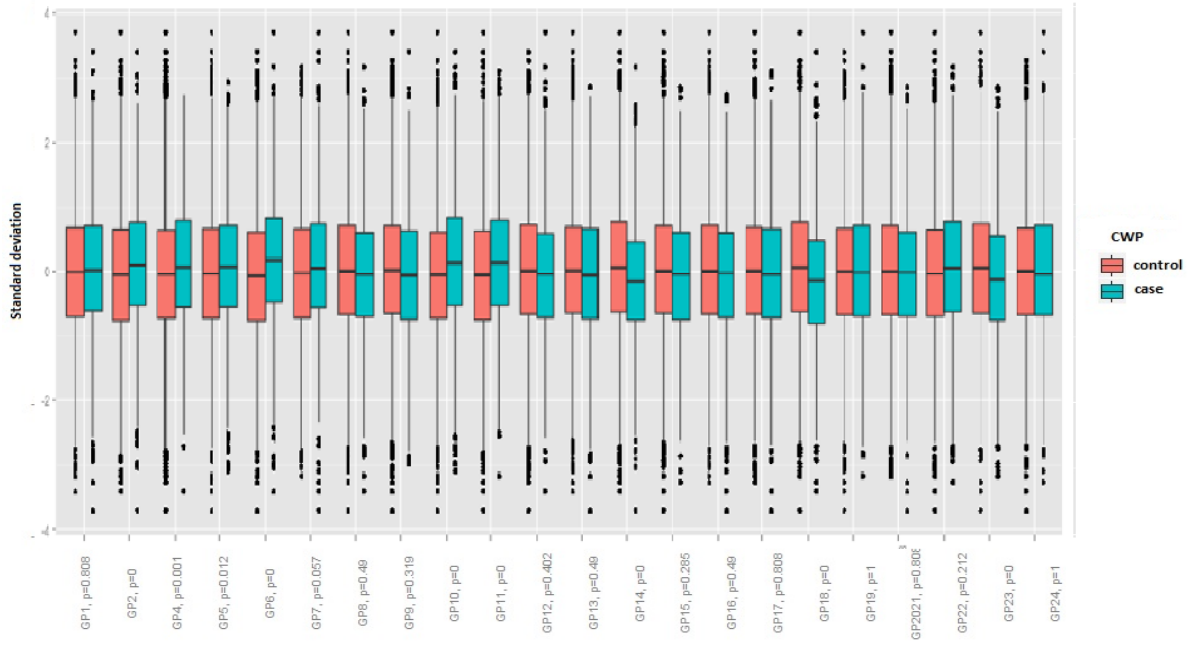

Figure 5 Distribution of glycan quantities between cases and controls for CWP phenotype

3.1. Linear mixed models results

Performing linear mixed model analysis with BMI, sex and disease outcome included as fixed covariate and variation in IgG glycan quantities within twin pairs as random effect for LPB without CWP phenotype show significant difference in glycan trait IGP50 which is derived from GP11. Graphical representation of p-values for these regressions are shown on Figure 6 and Table 3.

Table 3 Lowest p-values for LBP without CWP regression without correction for multiple testing

Glycan	IGP50	GP11	IGP76	IGP74
P-value	9.222886e-05	5.061528e-03	2.238026e-02	2.348379e-02

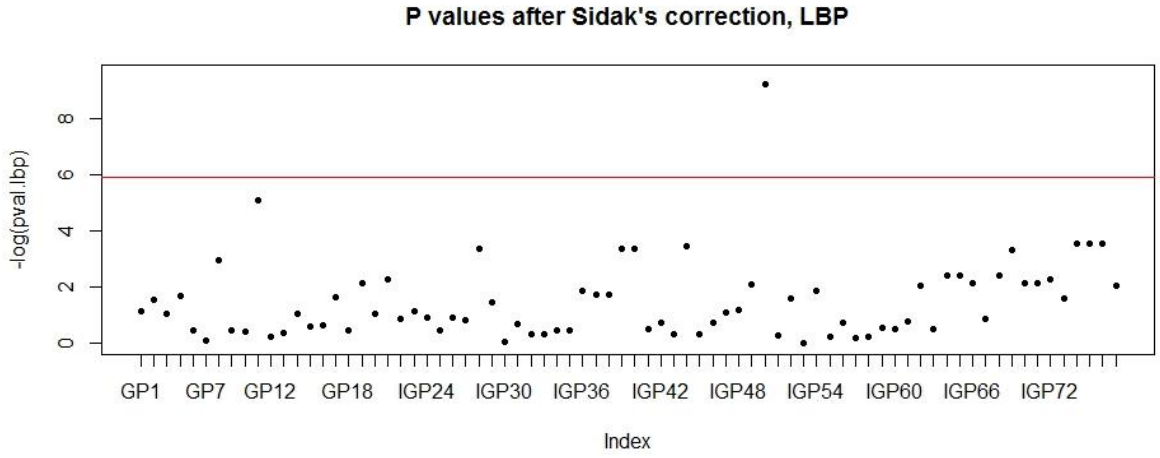

Figure 6 Significance of association of LBP without CWP to glycan quantity, horizontal red line represents significance threshold. P-values (-log2) are reported.

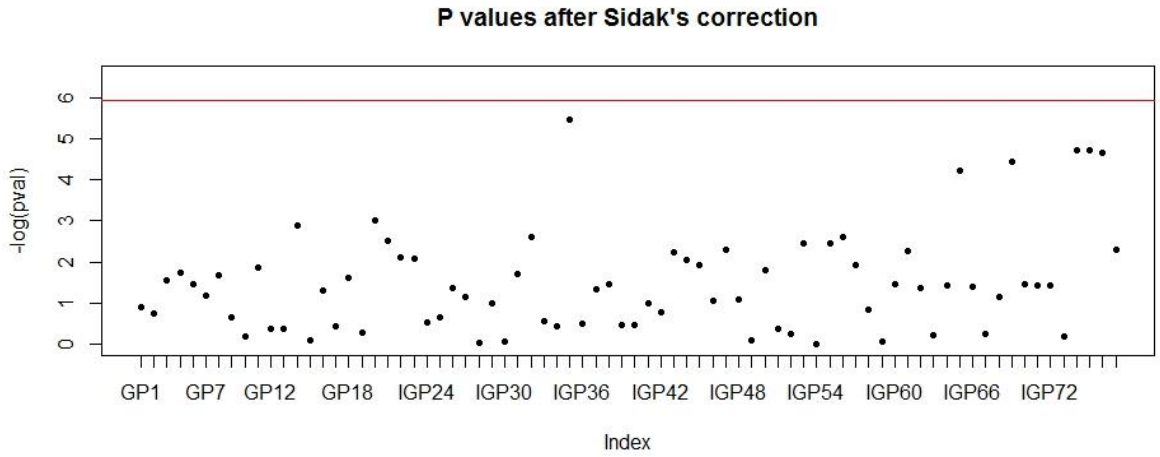

Figure 7 Significance of association of CWP to glycan quantity, horizontal red line represents significance threshold. P-values (-log2) are reported.

Analysis of CWP phenotypes didn't show any difference in glycan quantities as can be seen on Figure 7. Table 4 shows p-values of 4 most significantly associated glycans. None of them passed Sidak's multiple testing corrected threshold of significance.

Table 4 Lowest p-values for CWP regression without correction for multiple testing

Glycan	IGP35	IGP74	IGP75	IGP76
P value	0.003901948	0. 012613641	0. 012620624	0. 012409583

3.2. WGCNA results

To find clusters of similarly expressed glycans that are correlated with disease outcome, I firstly performed hierarchical clustering (Figure 8). By doing this, I grouped glycans that are generally similarly expressed. This dendogram clearly shows how derived glycans traits that were derived from the same or similar glycans are clustered together which proves how data handling done so far was done appropriately. To get glycans clusters that are similarly expressed only in cases but differently in controls I performed weighted (gene) correlation network analysis (WGCNA).

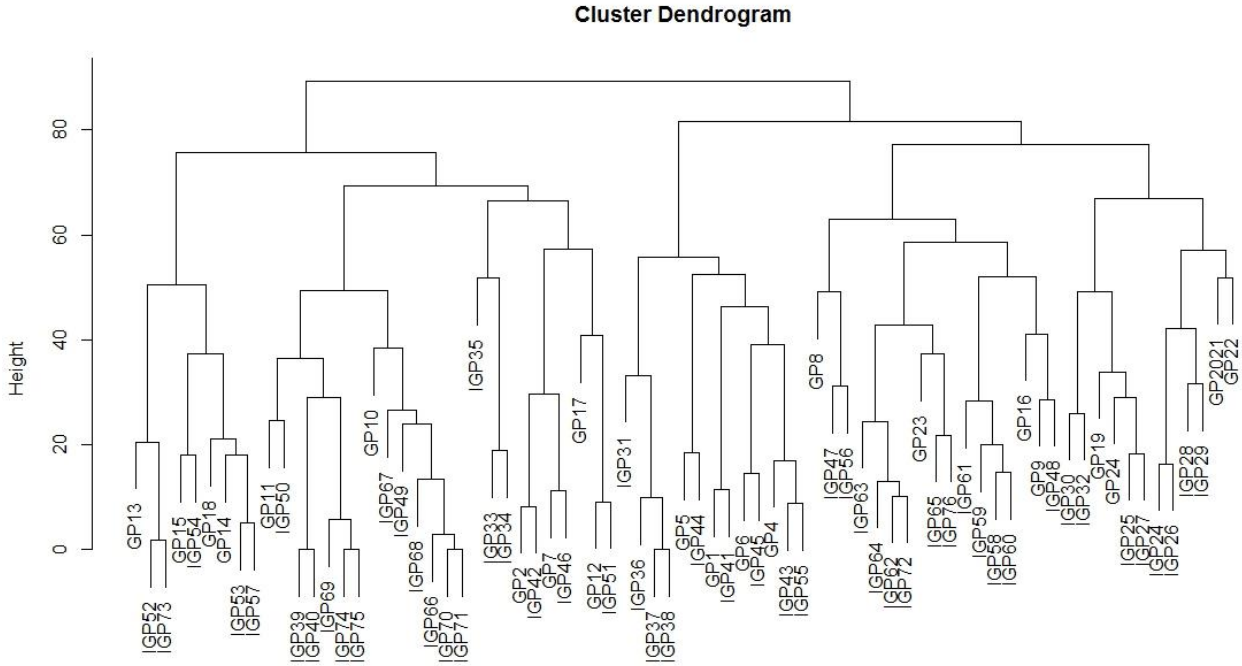

Figure 8 Hierarchical clustering of glycan values

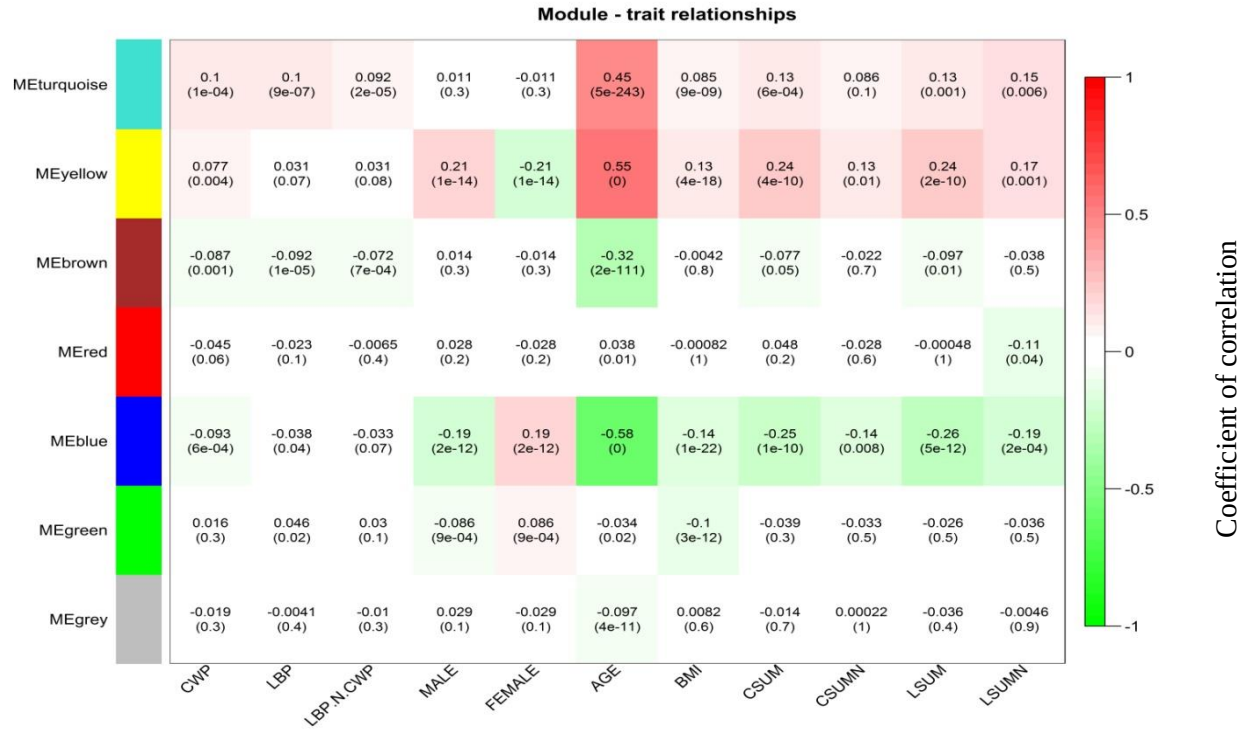

Figure 9 Relationships between modules of correlated glycans and traits. Pearson's or point-biserial correlation coefficients are provided (p-values). Red colour represents positive correlation while green represents negative correlation.

I used WGCNA methodology to analyse glycan levels and pain phenotypes. With signed networks, I identified seven modules of correlated glycans as can be seen on Figure 9. First analysis revealed remarkable correlation between module and phenotypes; however, the true relationships between pain phenotypes and modules were masked by age, sex, and BMI (Figure 9). Therefore, the analysis was repeated with glycans adjusted for these confounders (Figure 10 and 11). Again I got 7 modules which can be grouped into two big branches: comprising yellow, brown and turquoise modules, on one hand, and black, green, blue and red modules, on the other hand (Figure 11).

None of the modules was found to be associated with CWP, CSUM and LSUM; however, blue, brown and turquoise modules were associated with LBP with or without CWP (Figure 12). Blue module was negatively correlated with LBP (trait is associated with the decreased level of glycans), while brown and turquoise positively (trait is associated with the increased level of glycans). The correlations were extremely weak, though. The most promising results were obtained for turquoise module and 'LBP without CWP' (Their correlation was found to be point-biserial $R = 0.061$, $p = 0.004$).

Table 5 shows detail results of WGCNA analysis where all glycans are divided in 7 modules and their p-values for the significance of association with pain traits are listed.

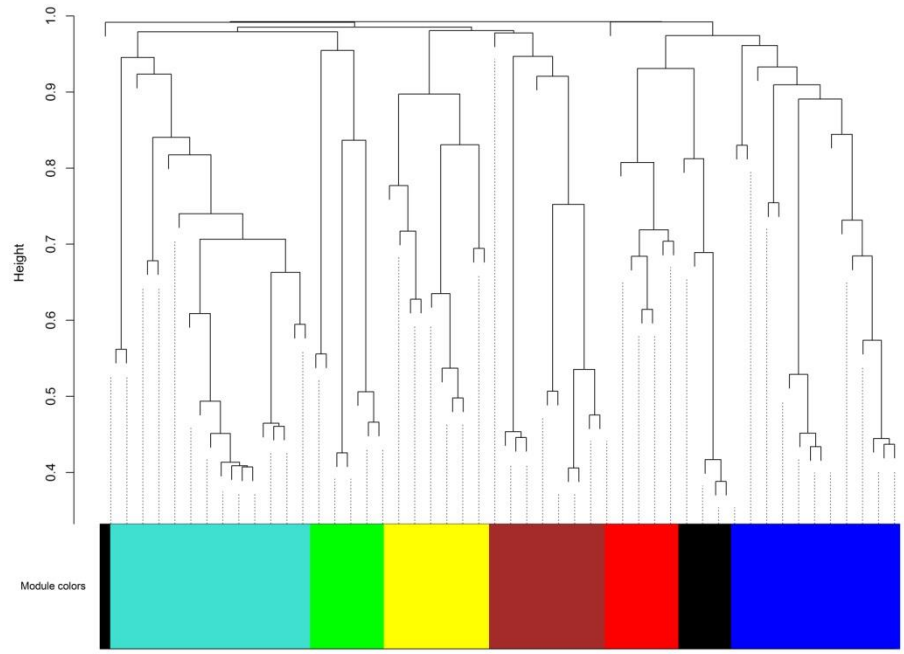

Figure 10 Hierarchical representation of modules of correlated glycans after adjustment for age, sex and BMI

Figure 11 Relationship between modules of correlated glycans after adjustment for age, sex and BMI

Table 5 Glycan-module allocation and their correlation (p-value) with traits

Glycan	module	CWP	LBP	LBP.N.CWP	CSUM	LSUM
GP1	green	0.409	**0.019**	0.070	0.342	0.996
GP2	brown	0.303	0.262	0.357	0.093	0.054
GP4	green	0.297	**0.030**	0.145	0.928	0.966
GP5	green	0.485	0.452	0.467	0.646	0.171
GP6	turquoise	0.339	0.366	0.229	0.948	0.622
GP7	brown	0.187	0.060	0.198	0.390	0.331
GP8	blue	**0.024**	0.323	0.077	0.521	0.908
GP9	blue	0.163	**0.028**	0.102	0.912	0.605
GP10	turquoise	**0.044**	**0.013**	0.066	0.749	0.510
GP11	turquoise	0.124	0.386	0.089	0.222	0.622
GP12	brown	0.065	0.054	0.313	0.991	0.678
GP13	brown	0.341	0.174	0.219	0.239	0.763

GP14	yellow	**0.007**	0.478	0.093	0.204	0.354
GP15	yellow	0.071	**0.037**	0.148	0.461	0.327
GP16	blue	**0.027**	0.136	0.338	0.685	0.808
GP17	brown	0.132	0.131	0.156	0.769	0.314
GP18	yellow	0.229	0.410	0.141	0.279	0.532
GP19	black	0.314	0.463	0.372	0.732	0.372
GP2021	red	0.388	0.183	0.344	0.334	0.306
GP22	brown	0.067	0.105	0.166	0.682	0.180
GP23	red	0.113	0.207	0.059	0.648	0.359
GP24	red	0.271	0.412	0.399	0.859	0.164
IGP24	yellow	0.154	0.491	0.458	0.793	0.805
IGP25	red	0.222	0.103	0.216	0.690	0.806
IGP26	yellow	0.419	0.241	0.438	0.707	0.708
IGP27	red	0.295	0.277	0.344	0.973	0.520
IGP28	yellow	**0.004**	0.436	0.189	0.820	0.780
IGP29	yellow	**0.002**	0.364	0.390	0.884	0.749
IGP30	red	0.388	0.155	0.195	0.500	0.939
IGP31	black	**0.019**	0.090	0.359	0.443	0.843
IGP32	red	0.445	0.248	0.338	0.861	0.473
IGP33	turquoise	0.207	0.386	0.413	0.663	0.692
IGP34	turquoise	0.132	0.204	0.123	0.616	0.910
IGP35	black	0.149	0.495	0.362	0.454	0.551
IGP36	black	0.400	0.320	0.110	0.506	0.696
IGP37	black	0.325	0.365	0.143	0.440	0.795
IGP38	black	0.325	0.365	0.143	0.439	0.796
IGP39	turquoise	0.180	0.084	**0.008**	0.624	0.917
IGP40	turquoise	0.178	0.083	**0.008**	0.614	0.905
IGP41	green	0.271	**0.048**	0.161	0.410	0.932
IGP42	brown	0.402	0.170	0.223	0.080	0.062
IGP43	green	0.078	**0.044**	0.298	0.601	0.722
IGP44	green	0.285	0.242	0.179	0.431	0.116
IGP45	turquoise	0.448	0.134	0.053	0.858	0.581

IGP46	brown	0.340	**0.028**	0.097	0.314	0.281
IGP47	blue	0.061	0.091	0.487	0.907	0.694
IGP48	blue	**0.010**	0.143	0.447	0.483	0.906
IGP49	turquoise	0.162	**0.001**	**0.005**	0.911	0.569
IGP50	turquoise	**0.028**	0.083	**0.006**	0.139	0.596
IGP51	brown	0.149	**0.028**	0.173	0.890	0.648
IGP52	brown	0.398	0.060	0.067	0.463	0.930
IGP53	yellow	0.058	0.319	0.280	0.380	0.535
IGP54	yellow	0.231	**0.013**	0.051	0.777	0.575
IGP55	green	0.114	0.138	0.498	0.600	0.790
IGP56	blue	0.373	**0.007**	0.054	0.652	0.685
IGP57	yellow	0.065	0.199	0.415	0.413	0.571
IGP58	blue	0.218	**0.037**	0.145	0.575	0.440
IGP59	blue	0.227	0.057	0.174	0.268	0.170
IGP60	blue	0.368	0.064	0.144	0.352	0.302
IGP61	blue	0.382	0.086	0.116	0.596	0.321
IGP62	blue	0.287	**0.003**	**0.006**	0.688	0.985
IGP63	blue	0.104	**0.004**	**0.023**	0.876	0.745
IGP64	blue	0.282	**0.005**	**0.009**	0.930	0.721
IGP65	blue	0.136	**0.019**	**0.008**	0.340	0.408
IGP66	turquoise	0.273	**0.004**	**0.005**	0.940	0.440
IGP67	turquoise	0.101	**0.004**	**0.020**	0.595	0.436
IGP68	turquoise	0.305	**0.008**	**0.011**	0.911	0.527
IGP69	turquoise	0.076	0.056	**0.008**	0.308	0.693
IGP70	turquoise	0.304	**0.005**	**0.007**	0.736	0.812
IGP71	turquoise	0.304	**0.005**	**0.007**	0.735	0.812
IGP72	blue	0.308	**0.004**	**0.006**	0.775	0.794
IGP73	brown	0.411	0.054	0.065	0.488	0.961
IGP74	turquoise	0.080	**0.036**	**0.005**	0.297	0.616
IGP75	turquoise	0.080	**0.036**	**0.005**	0.296	0.614
IGP76	blue	0.055	0.053	**0.007**	0.311	0.514
IGP77	turquoise	**0.034**	0.337	0.091	0.591	0.201

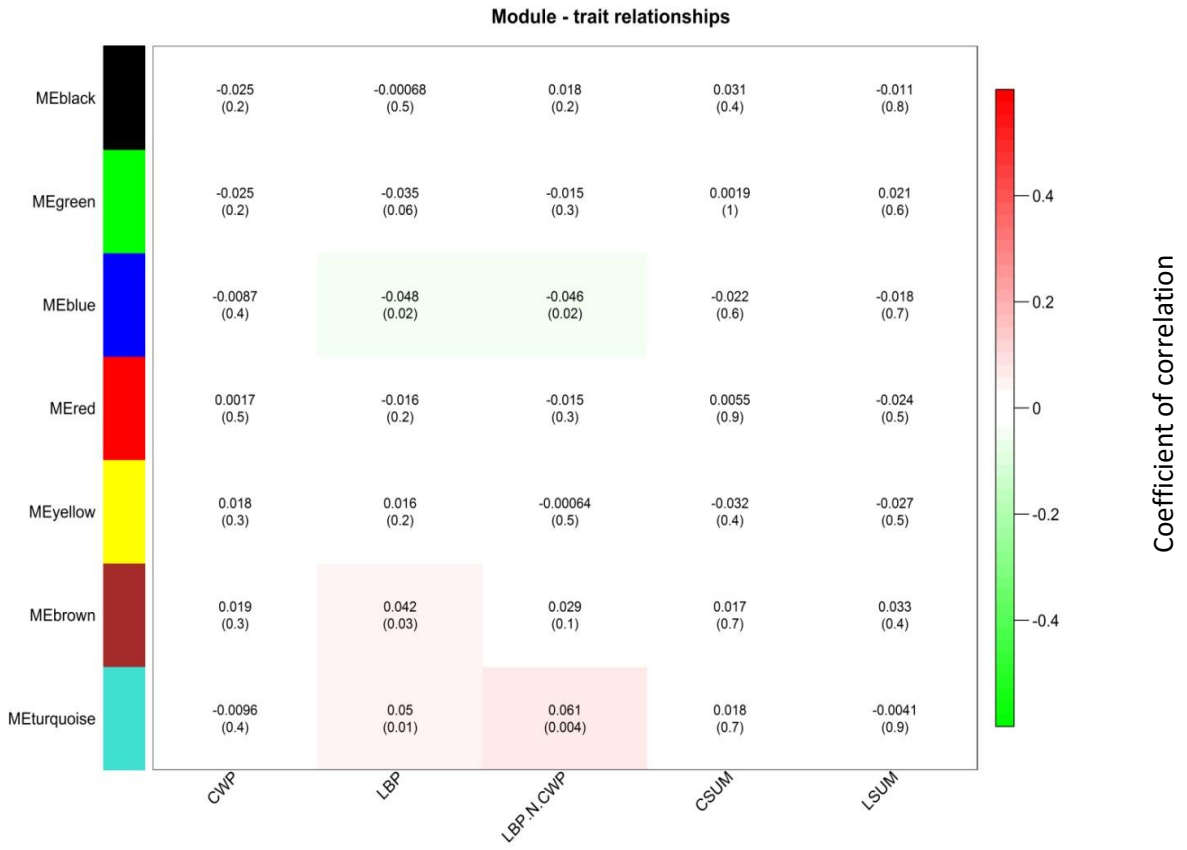

Figure 12 Correlations between module eigenvalues with pain phenotypes and MRI traits after adjustment for age, sex and BMI. Red colour represents positive correlation while green represents negative correlation.

3.3. Discordant twins analysis

To further analyse the relationships between glycome and pain phenotypes, I carried out comparisons of glycan levels in twins discordant for CWP, LBP, and 'LBP without CWP' traits. First, I used all the glycans and compared their levels by Wilcoxon's test and mixed-models regression between the MZ and DZ twins.

For MZ twins, I identified statistically significant differences between the twins with and without LBP for the IGP65, IGP74, IGP75, and IGP76 derived traits (Figure 13; $p<0.0027$ for at least one of the statistical tests used). Notably, these four glycan traits belong to the blue and turquoise modules identified in the WGCNA analysis. Accordingly, IGP65 and IGP76 of the

blue module were found to be elevated in MZ twins without LBP, while IGP74 and IGP75 of the turquoise module were elevated in MZ twins with LBP (Figure 14). The four glycan traits were derived from neutral glycans GP14 and GP15, and also GP13 for IGP76, with GP14 being the numerator for IGP65 and IGP76, while GP15 the numerator for the other two. Intriguingly, neither GP14, nor GP15 showed any trend to association with LBP.

No differences were found for other pain phenotypes. Also, no differences were found for DZ twins or MZ and DZ twins combined.

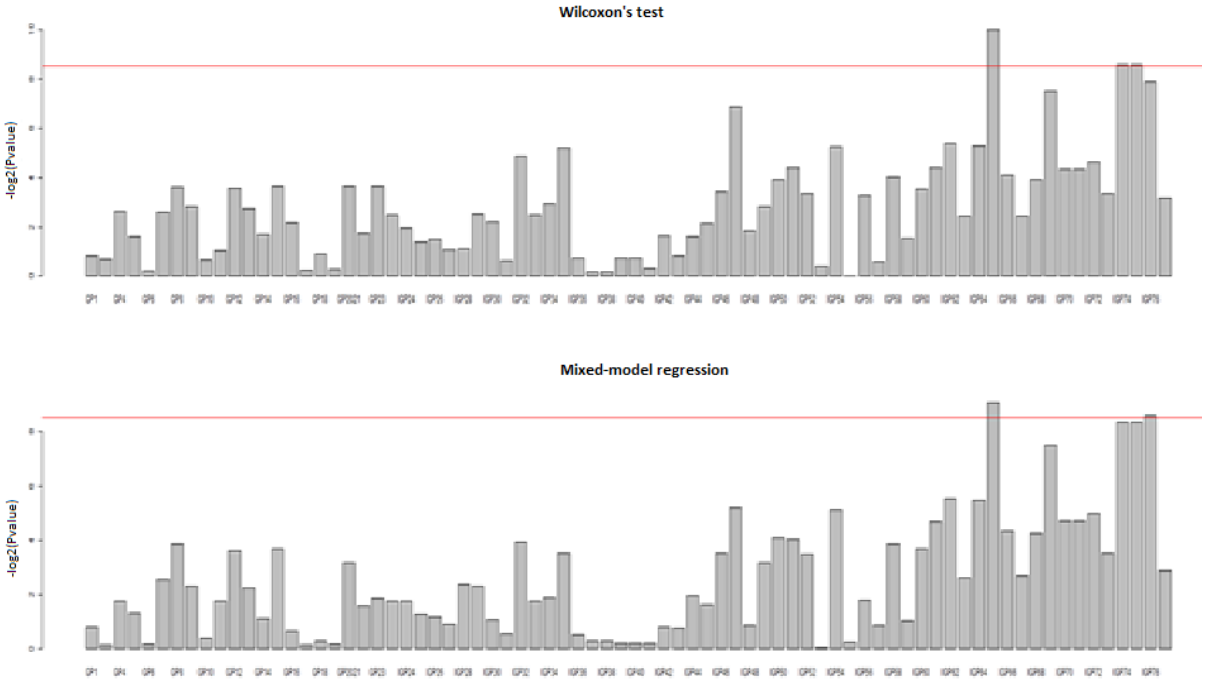

Figure 13 P-values (-log2) for comparisons of mean glycan levels in MZ twins discordant for LBP phenotype. Each column represents one glycan (from GP1 to IGP77) Horizontal red line corresponds to p=0.0027 which was taken as the significance threshold based on the 19 effective independent tests with Sidak's correction for multiple testing

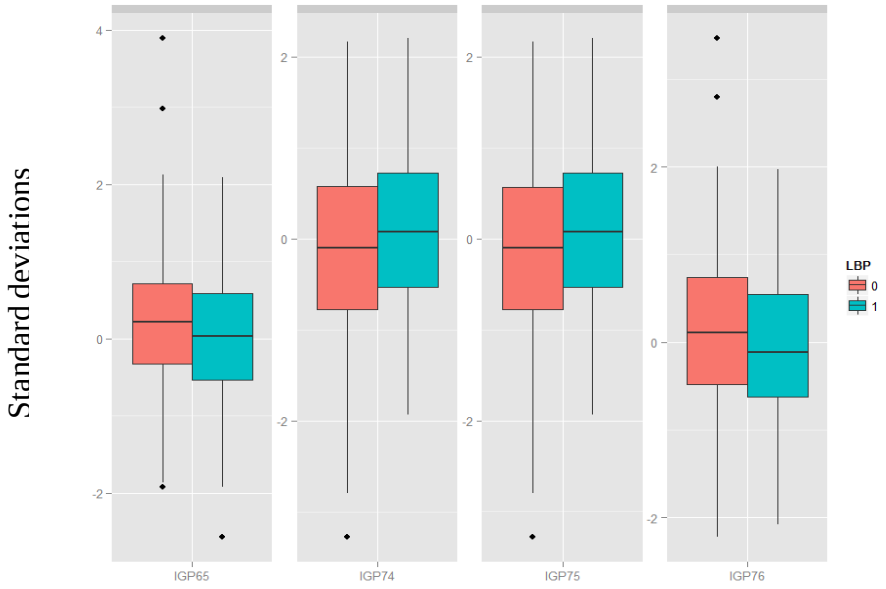

Figure 14 Glycan distributions of significantly different glycans between discordant MZ twins in LBP study

3.4. Prediction of a disease status

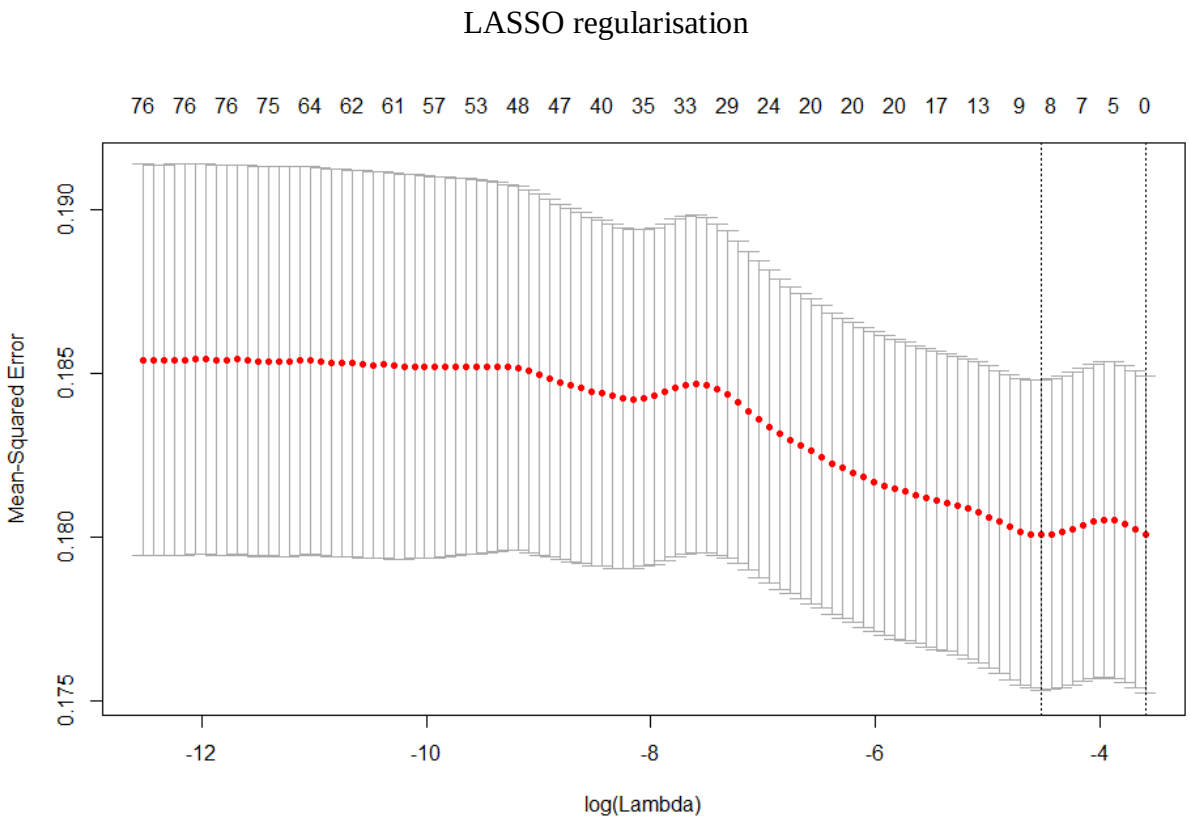

Figure 15 Mean-Squared error of predictions for 100 prediction models with different LASSO parameters

Figure 15 shows how mean-squared error changes when predicting disease status on a test part of the dataset with different parameters for LASSO algorithm and when using different number of variables. If I would use all glycan variables for prediction, the accuracy of prediction would be the lowest. According to this graph, the best lambda parameter is 0.01082604 which corresponds to model with eight glycan variables (vertical line on the plot). Eight glycans used in that model are: GP1, GP17, GP22, IGP37, IGP40, IGP50 and IGP56.

Regression model in this case was, because of the prediction purposes, defined other way around compared to previous models built in Mixed models chapter. All models looked like this: **LBP ~ selected glycans + sex + BMI + (1|familyID).**

Prediction algorithm built with these eight glycans showed 76.85% accuracy in predicting new patients from their glycan profile. Table 6 summarises accuracies of prediction of glycan subset that was derived after LASSO regularisation and different glycan modules.

Table 6 Accuracy of prediction for glycan modules

Glycan subset	Accuracy of prediction on test dataset
LASSO regularisation	76.85%
Turquoise module	77.99%
Blue module	76.94%
Brown module	76.94%
Yellow module	76.76%
Green module	76.76%
Red module	76.26%
Black module	75.94%

4. Discussion

4.1. Regression analysis

Immunoglobulin G is an excellent glycoprotein model as its glycosylation is well defined and many important functional effects of alternative IgG glycosylation have been described (Gornik, 2012). *N*-glycans attached to the conserved asparagine 297 in the Fc part of IgG are important modulators of IgG effector functions (Gornik, 2012) For example, glycosylation acts as a switch between pro- and anti-inflammatory IgG functionality. Malfunction of this system is associated with different inflammatory and autoimmune diseases such as SLE (Lauc et al., 2013), rheumatoid arthritis and inflammatory bowel diseases (Ohtsubo and Marth, 2006). Because one of the major determinants of pain in persistent CLBP syndrome is localised inflammation in epidural space (Broos, Aebi, 2008), and as glycans are known to be associated with inflammatory diseases, in this study I have evaluated, using a classical twin study design, association between quantities of plasma IgG *N*-glycans and LBP. This ended as a very difficult task because distributions of glycan quantities between cases with LBP and controls were very similar (Figure 6). Although my sample size was big enough to detect significant variations between cases and controls, only one glycan trait that came out significant was IGP50 which is derived from GP11. Interestingly GP11 had second lowest P-value but didn't pass significance threshold. For the CWP analysis I used even bigger dataset, and here differences in distributions of glycan quantities were even bigger (Figure 7), but after performing mixed model regression, no significantly different glycans showed up.

Generally differences in glycan quantities, when researched from the whole plasma proteins can be attributed to multiple effects like different ratio of plasma proteins, different levels of glycosylation, but in this study I avoided both problems by isolating a single protein from plasma (IgG), which is produced by a single cell type (B lymphocytes), thus effectively excluding differential regulation of gene expression in different tissues, and the "noise" introduced by variation in plasma IgG concentration and by *N*-glycans on other plasma proteins.

4.2. WGCNA

A new emphasis on the thoughtful use and adoption of statistical analyses is required in order for the biological sciences to keep pace with the increasing dominance of complex and highly multivariate systems biology data. Hence, in this study I used methods that were developed for genomic data and applied it to other multivariate omic data. This is the first time that WGCNA is applied to glycan data.

Weighted correlation network analysis is a powerful methodology for revealing clusters (modules) of multiple omic traits, such as genome-wide gene expression or global methylation profiles, and placing them into a biological context through the analysis of associations between the clusters and diseases or traits of interest (Horvath et al., 2006; Presson et al., 2008; Saris et al., 2009; van Eijk et al., 2012).

WGCNA defines a network that continuously links all variables and then clusters the most highly co-expressed variables in flexibly defined modules. WGCNA can be used to create both signed networks, which separate positively and negatively correlated nodes into separate modules and also unsigned networks, which assess correlations by their absolute values.

Here I applied WGCNA on glycan data to assess whether there are different patterns of glycan quantities among twins and is there a biological meaning relating to chronic pain behind these modules. As a result I obtained seven different modules of glycan quantities. At first, due to my lack of experience there was a strong association among the most of the modules with various pain phenotypes, but after I investigated further, it turned out that actually most of the associations are linked to age, BMI or sex (Figure 9). I have fixed that problem and corrected glycan quantities for age, sex and BMI. After running this analysis again with corrected values, I found much smaller associations. Interestingly only 3 modules were associated with LBP while only two modules were associated with LBP without CWP. This means that in the brown module (the module where glycans are only associated with LBP, but not with LBP without CWP) are glycans that are actually associated with CWP and not with LBP but because of manifestations of CWP we get the impression that these glycans could be related to LBP. According to the results of WGCNA for the further studies that are seeking for association between LBP and glycan quantities, it would be the most efficient to investigate only glycans that were clustered in turquoise and blue module because they showed significant differences between cases and controls in both LBP and LBP without CWP while others didn't.

Grouping features into modules has several advantages. First, condensing a very large network into a small number of modules or, alternatively, hub nodes allows external traits to be compared to a limited number of variables, providing a solution to the multiple testing problem. Second, module construction provides a means by which the roles of poorly characterized glycans can be inferred from their better-annotated neighbours. The identification of co-regulated modules helps to annotate the results from systems biology scale experiments, adding valuable biological information. Third, since the influence of minor variables is not masked by the most dramatic differences in terms of absolute scale, as occurs in PCA, WGCNA allows the combining of disparate datasets.

4.3. Discordant twins analysis

The discordant twin design allows for a comparison of probands and controls while "matching" for the underlying genetic or shared environmental factors that may influence general cognitive ability.

I used MZ and DZ twins discordant for LBP in the validation analysis. Glycan levels may be influenced by many factors including genetics, age and environment (Lauc at al., 2013). As identical twins share 100% of their genetic makeup, and are matched perfectly for age, gender, social class, *etc*., I was able to validate the role of IgG on LBP; isolating the nongenetic contribution. These data help us to understand the complex interplay between genetic and nongenetic influences that determine LBP.

In these analysis the only significantly associated glycans are: IGP65, IGP74, IGP75, and IGP76 glycan derived traits, while the glycans from which they were derived are not significantly associated with disease outcome. If I check the meaning behind these derived glycan traits that are shown on Table 7, I notice that all traits are connected with fucosilation of digalactosylated structures, both bisecting and non-bisecting.

According to this finding, from discordant twin analysis, I would suggest that there is a link between low back pain and fucosilation of dygalactosilated IgG *N*-glycan structures.

Table 7 Significantly associated glycan traits with LBP, discordant twin study

Glycan	Description
IGP65	The percentage of fucosylation of digalactosylated structures (without bisecting GlcNAc) in total neutral IgG glycans
IGP74	Ratio of fucosylated digalactosylated structures with and without bisecting GlcNAc in total neutral IgG glycans
IGP75	The incidence of bisecting GlcNAc in all fucosylated digalactosylated structures in total neutral IgG glycans
IGP76	Ratio of fucosylated digalactosylated non-bisecting GlcNAc structures and all digalactosylated structures with bisecting GlcNAc in total neutral IgG glycans

4.4. Predictive power of WGCNA modules

Knowing that associations between various pain phenotypes and glycan values are low, if any, I was sure that my predictive models will not be very accurate or applicable to other datasets. Nevertheless I have still decided to build them in order to see whether WGCNA modules can be a good variable selection tool.

As a standard upon which comparisons of accuracy will be made I have built predictive model after LASSO regularisation. This method is widely used in genomic studies when there is a need for reduction of number of variables due to a lot of genomic markers used as predictors (Tibshirani, 1997). LASSO regularisation method found out that the lowest mean squared error is observed when model uses only eight glycan variables. For each glycan LASSO was checking their contribution to prediction accuracy in model, and not any biological prior knowledge. On the other hand, WGCNA modules were built based on glycan quantities which are directly linked to biological features because of common biological pathways.

Glycans chosen by LASSO are not members of only one WGCNA module, but belong to four different modules although most of them are from blue and turquoise module, modules that showed association with LBP.

It is interesting to see how prediction accuracy of turquoise, blue and brown module was slightly higher than with LASSO, while other modules preformed worse. This proves that

variable selection based on biological prior knowledge could in future drastically help in building prediction models.

The reason why prediction accuracies were very similar in all models is because of uneven quantities of cases and controls and males and females in my dataset. How 75% of the database were controls model learned much better to distinguish controls, so if it was unsure whether some patient is case or control based on their glycan profile, in all models, it would assign it as control.

There are some limitations of this study. First, there is a female predominance in our study sample (95% of the individuals are, for historical reasons, women). Second, this population being volunteers is slightly healthier than average with a lower rate of diabetes and results might not be generalizable to more severe diabetes populations. Third, the cross-sectional nature of our data does not allow us to draw conclusions as to whether the glycans identified are causative of LBP or merely correlated with it. Finally, I cannot provide reliable estimates as to what proportions of the identified glycans were from Fc and from Fab, respectively. However, in a small pilot of Fc-glycopeptides by nano-liquid chromatography tandem mass spectrometry (Huffman at al. 2014) on 96 representative age-matched individuals from the extremes of the eGFR distribution, I find the same direction of effect in this study which suggests that my initial observations mostly come from the Fc glycans.

Additionally, association of pain phenotypes and glycan quantities is almost not evident in this study, which is partly caused by the way pain phenotypes are defined. Patients filling in questionnaires about their pain history is not the best solution to assess someone's pain perception. For future studies I suggest to correlate tests of objective measures of pain` perception with answers on questionnaires. To make sample more random, database should have even number of males and females which should be chosen randomly from the population.

Along all drawbacks of this study analysis of discordant twin pairs suggests how there is an association between pain phenotypes and glycans, but new studies are needed to define which phenotypes are associated with differential glycan quantity and what is the biological meaning behind subset of those glycans.

5. Conclusion

- Pre-processing and filtering of glycan values was successfully done for 1932 twins for LBP without CWP study and 4416 twins for CWP study.
- Cofounders of glycan quantities were examined and all glycan quantities were corrected for age of volunteer.
- Regression mixed models were built for LBP without CWP and CWP pain phenotypes which showed how associations between glycans with these phenotypes are very weak.
- Seven modules of glycans with different quantity pattern between cases and controls were obtained by WGCNA method.
- All significantly associated glycan traits that I found in discordant MZ twins analysis were linked with fucosylation of digalactosylated glycan structures which gives a reason to suspect how there is a link between glycans and pain phenotypes.
- Prediction models were built that would, based on person's glycan quantities, predict whether that person is case or control.
- Three WGCNA modules that showed association with LBP predicted disease outcome the best, even better than LASSO selection model.
- In future replication in a male dataset would be beneficial, also one containing different forms of cases and controls having objective measures of pain sensitivity.

6. References

Abbott, K. L., Nairn, A. V., Hall, E. M., Horton, M. B., McDonald, J. F. et al. (2008): **Focused glycomic analysis of the N-linked glycan biosynthetic pathway in ovarian cancer**. *Proteomics* 8 (16): 3210-3220. doi: 10.1002/pmic.200800157.

Alvarez-Manilla, G., Warren, N. L., Abney, T., Atwood III, J., Azadi, P., York, W. S., Pierce, M., Orlando, R. (2007): **Tools for glycomics: relative quantitation of glycans by isotopic permethylation using 13CH3I**. *Glycobiology* 17: 677–687.

Andersson, G. B. (1999): **Epidemiological features of chronic low-back pain.** *Lancet.* 354: 581–585.

Arnold, J. N., Wormald, M. R., Sim, R. B., Rudd, P. M., Dwek, R. A. (2007): **The impact of glycosylation on the biological function and structure of human immunoglobulins**. *Annu Rev Immunol* 25: 21-50. doi:10.1146/annurev.immunol.25.022106.141702. PubMed: 17029568.

Aulchenko, Y. S., Ripke, S., Isaacs, A., van Duijn, CM. (2007): **GenABEL: an R library for genome-wide association analysis**. *Bioinformatics* 23 (10): 1294-1296.

Battié, M. C., May, L. (2001): **Physical and occupational therapy assessment approaches**. *In D. C. Turk, R. Melzack (Eds.), Handbook of pain assessment* (2nd ed.). New York, NY: Guilford Press.

Blickstein, I., Keith, LG. (2005): **Multiple Pregnancy: Epidemiology, Gestation, and PerinatalOutcome**, Second Edition. *In-forma HealthCare.*

Boomsma, D., Busjahn, A., Peltonen, L. (2002): **Classical twin studies and beyond**. *Nat Rev Genetics* 3(11): 872-882.

Brooks, P. M. (2006): **The burden of musculoskeletal disease - A global perspective**. *Clin Rheumatol*. 25: 778–781.

Broos, N., Aebi, M. (2008): **Spinal Disorders: Fundamentals of Diagnosis and Treatment.** *Springer-Verlag Berlin Heidelberg.*

Brown, C. A., Seymour, B., Boyle, Y., El-Deredy, W., Jones, A. K. P. (2008): **Modulation of pain ratings by expectation and uncertainty: Behavioral characteristics and anticipatory neural correlates.** *Pain* 135(3): 240-250.

Cassidy, J. D., Carroll, L. J., Cote, P. (1998): **The Saskatchewan health and back pain survey: the prevalence of low back pain and related disability in Saskatchewan adults**. *Spine* 23: 1860–1866.

Chapman, C. R., Casey, K. L., Dubner, R., Foley, K. M., Gracely, R. H., Reading, A. E. (1985): **Pain measurement: an overview**. *Pain* 22: 1-31.

Chen, C., Grennan, K., Badner, J., Jin, L. et al. (2011): **Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods**. *PLoS ONE* 6(2): e17238. doi:10.1371/journal.pone.0017238.

Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006): **Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR** *Metabolomics*. *Analytical Chemistry* 78 (13): 4281-4290 doi: 10.1021/ac051632c.

Dube, D. H., Bertozzi, C. R. (2005): **Glycans in cancer and inflammation – potential for therapeutics and diagnostics**. *Nat. Rev. Drug Discovery* 4: 477–488.

Edmond, S. L., Felson, D. T. (2000): **Prevalence of back symptoms in elders**. *J Rheumatol*. 27: 220–225.

Freeze, H. H. (2006): **Genetic defects in the human glycome.** *Nat Rev Genet* 7: 537-551 doi: 10.1038/nrg1894 .

Garofalo, J. P., and Polatin, P. (1999): **Low back pain: an epidemic in industrialized countries.** *In R. J. Gatchel, D. C. Turk (Eds.), Psychosocial factors in pain: critical perspectives. New York, NY: Guilford Press.*

Geisser, M. E., Robinson, M. E., Miller, Q. L., and Bade, S. M. (2003): **Psychosocial factors and functional capacity evaluation among persons with chronic pain**. *Journal of Occupational Rehabilitation* 13: 259-276.

Geisser, M. E., Roth, R. S., and Williams, D. A. (2006): **The allure of a cure**. *Journal of Pain* 7: 804-806.

Gornik, O., Pavić, T., Lauc, G. (2012): **Alternative glycosylation modulates function of IgG and other proteins - implications on evolution and disease**. *Biochim Biophys Acta* 1820: 1318–1326.

Gornik, O., Wagner, J., Pucić, M., Knezević, A., et al. (2009): **Stability of N-glycan profiles in human plasma**. *Glycobiology* 19: 1547-1553 doi: 10.1093/glycob/cwp134.

Hall, J. G. **Twinning**. (2003): *Lancet* 30 362(9385): 735-743.

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., Laurance, M. F., Zhao, W. et al. (2006): **Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target**. *PNAS* 103(46): 17402-17407.

Huffman, J. E., Pucic-Bakovic, M., Klaric, L., et al. (2014): **Comparative performance of four methods for high-throughput glycosylation analysis of immunoglobulin G in genetic and epidemiological research.** *Molecular and cellular proteomics : MCP* 13(6): 1598-1610.

Huffman, J. E., Knezevic, A., Vitart, V., Kattla, J., Adamczyk, B. 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: 5000-5011. doi:10.1093/hmg/ddr414. PubMed: 21908519.

Jensen, M. P., Karoly, P. (2001): **Self-report scales and procedures for assessing pain in adults**. *In D. C. Turk, R. Melzack (Eds.), Handbook of pain assessment (2nd ed.). New York, NY: Guilford Press.*

Kraychete, D. C., Sakata, R. K. et al. (2010): **Serum cytokine levels in patients with chronic low back pain due to herniated disc: analytical cross-sectional study.** Sao Paulo Med J. 128(5): 259-262.

Langfelder, P, Horvath S (2012): **Fast R Functions for Robust Correlations and Hierarchical Clustering**. Journal of Statistical Software: 46(11): 25-28.

Langfelder, P., Horvath, S. (2008): **WGCNA: an R package for weighted correlation network analysis.** *Bioinformatics* 9: 559.

Lauc, G., Huffman, J. E., Pučić, M., Zgaga, L., Adamczyk, B. et al. (2013): **Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers**. *PLoS Genet* 9: e1003225. PubMed: 23382691.

Lauc, G., Essafi, A., Huffman, J. E., Hayward, C., Knezevic, A. et al. (2010): **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. doi: 10.1371/journal.pgen.1001256

Leboeuf-Yde, C., Kyvik, K. O., Bruun, N. H. (1998): **Low back pain and lifestyle. Part I: smoking. Information from a population- based sample of 29,424 twins**. *Spine* 23: 2207–2213.

Li, J. and Ji, L. (2005): **Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix**. *Heredity* 95: 221–227.

Louw, Q. A., Morris, L. D., Grimmer-Somers, K. (2007): **The prevalence of low back pain in Africa: a systematic review.** *BMC Musculoskelet Disord.* 8: 105. Accessed May 6, 2015.

Marino, K., Saldova, R., Adamczyk, B., Rudd, P. M. (2012): **Changes in Serum N-Glycosylation Profiles: Functional Significance and Potential for Diagnostics**. *In: AP RauterT Lindhorst. Carbohydrate Chemistry. Cambridge, UK: The Royal Society of Chemistry*.

Mills, P. B., Mills, K., Mian, N., Winchester, B. G., Clayton, P. T. (2003): **Mass spectrometric analysis of glycans in elucidating the pathogenesis of CDG type *Iix***. *J. Inherit. Metab. Dis.* 26: 119–134.

Muggeo, V. M. R. (2003): **Estimating regression models with unknown break-points**. *Statistics in Medicine* 22: 3055-3071.

Nairn, A. V., York, W. S., Harris, K., Hall, E. M., Pierce, J. M. et al. (2008): **Regulation of glycan structures in animal tissues: transcript profiling of glycan-related genes**. *J Biol Chem.* 283: 17298-17313. doi:10.1074/jbc.M801964200.

Ohtsubo, K., Marth, J. D. (2006): **Glycosylation in cellular mechanisms of health and disease.** *Cell* 126: 855-867. doi: 10.1016/j.cell.2006.08.019.

Paajanen, H., Lehto, I., Alanen, A., Erkintalo, M., Komu, M. (1994): **Diurnal fluid changes of lumbar discs measured indirectly by magnetic resonance imaging**. *J Orthop Res*. 12: 509–514.

Papageorgiou, A. C., Croft, P. R., Ferry, S., Jayson, M. I., Silman, A. J. (1995): **Estimating the prevalence of low back pain in the general population: evidence from the South Manchester Back Pain Survey**. *Spine* 20: 1889–1894.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2015): **nlme: Linear and Nonlinear Mixed Effects Models**. R package version 3.1-121.

Presson, A. P., Sobel, E. M., Papp, J. C., Suarez, C. J., Whistler, T., Rajeevan, M. S., Vernon, S. D., Horvath, S. (2008): **Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome**. *BMC Syst Biol*. 2: 95.

Pucić, M., Knezević, A., Vidic, J., Adamczyk, B., Novokmet, M. 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: 90-100.

Pucic, M., Pinto, S., Novokmet, M., Kneževid, A., Gornik, O., Polašek, O., Vlahoviček, K., Wang, W., Rudd, P. M., Wright, A. F., Campbell, H., Rudan, I., Lauc, G. (2010): **Common aberrations from the normal human plasma N-glycan profile**. Glycobiology 254: 1–6.

Royle, L., Campbell, M. P., Radcliffe, C. M., White, D. 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.

Sambrook, P. N., MacGregor, A. J. and Spector, T. D. (1999): **Genetic influences on cervical and lumbar disc degeneration: a magnetic resonance imaging study in twins.** Arthritis Rheum. 42(2): 366-372.

Saris, C. G., Horvath, S., van Vught, P. W., van Es, M. A., el al (1998): **Natural history of low back pain: a longitudinal study in nurses**. *Spine* 23: 2422–2426.

Tan, Q. et al. (2010): **Dissecting complex phenotypes using the genomics of twins.** Funct Integr Genomic 10(3): 321-327. doi: 10.1007/s10142-010-0160-9.

Thorn, B. E. (2004): **Cognitive therapy for chronic pain: A step-by-step guide**. *New York, NY: Guilford Press.*

Tibshirani, R., (1997): **The lasso method for variable selection in the Cox model.** *Statistics in medicine* 16(4): 385-395.

Turk, D. C., Melzack, R. (2001): **The measurement of pain and the assessment of people experiencing pain.** *In D. C. Turk, R. Melzack (Eds.), Handbook of pain assessment (2nd ed.). New York, NY: Guilford Press.*

van Eijk K. R, de Jong S, Boks M. P, Langeveld T, Colas F, Veldink J. H, et al. (2012): **Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects**. *BMC Genomics* 13: 636.

Vasudevan, S. V. (1992): **Impairment, disability, and functional capacity assessment**. *In D. C. Turk, R. Melzack (Eds.), Handbook of Pain Assessment (1st ed.). New York, NY: Guilford Press.*

Von Korff, M. (1999): **Pain management in primary care: an individualized self-care approach.** *In R. J. Gatchel, D. C. Turk (Eds.), Psychosocial factors in pain: critical perspectives. New York, NY: Guilford Press.*

Wager, T. D. (2005): **Expectations and anxiety as mediators of placebo effects in pain.** *Pain* 115: 225-226.

White, K. P., Harth, M., Speechley, M. et al. (1999). **Testing an instrument to screen for fibromyalgia syndrome in general population studies: the London Fibromyalgia Epidemiology Study Screening Questionnaire**. *J Rheumatol* 26(4): 880-884.

Wynne-Jones, G., Dunn, K. M., Main, C. J. (2008): **The impact of low back pain on work: A study in primary care consulters.** *European Journal of Pain* 12: 180-188.

Curriculum vitae

PERSONAL INFORMATION

Dunja Vučenović

Kornatska 27, 22000 Šibenik (Croatia)

(+385) 98 921 6886

dunja.vucenovic@gmail.com

Skype dunja.vucenovic

Sex Female | Date of birth 28/06/1991 | Nationality Croatian

WORK EXPERIENCE

02/02/2015–02/06/2015 Internship in Statistical analysis of Omics data for complex deseases

King's College London, Department of Twin research, London (United Kingdom)

Supervisors: Frances Williams and Tim Spector

06/2013–08/2013 Internship student researching chimerical transcripts

Structural and computational biology lab at CNIO, Madrid (Spain)

Supervisors: Milana Frenkel-Morgenstern and Alfonso Valencia

02/2013–06/2013 Internship in biochemistry lab describing properties of ribosomal proteins L10 and L12

Department of Biochemistry, Faculty of Science, Zagreb (Croatia)

Supervisors: Vlatka Godinic-Mikulcic and Ivana Weygand-Durasevic

EDUCATION AND TRAINING

01/09/2013–09/2015 Master of Molecular Biology

Faculty of Science, Zagreb (Croatia)

08/2014–08/2014 Summer School of Statistical Omics

MedILS, Split (Croatia)

09/2010–09/2013 Bachelor of Molecular Biology

Faculty of Science, Zagreb (Croatia)

PERSONAL SKILLS

Mother tongue(s): Croatian

Other language(s):

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken interaction	Spoken production	
English	C1	C1	C1	C1	C1
Italian	B1	B1	A2	A2	A2
German	A2	A2	A2	A2	A2

Organisational / managerial skills: I organized workshops with scientific background for pre-school kids in kindergarten, and for high school students in Summer school of Science in Višnjan, Croatia.

Project leader in Summer school of Science in Pozega, Croatia

Computer skills: Advanced user of Microsoft Office programs. I do program in Python and R, and I have basic skills in C++, C# and Octave

ADDITIONAL INFORMATION

Publications: ChiTaRS 2.1—an improved database of the chimeric transcripts and RNA-seq data with novel sense–antisense chimeric RNA transcripts: Milana Frenkel-Morgenstern, Alessandro Gorohovski, Dunja Vucenovic, Lorena Maestre, Alfonso Valencia; Nucleic Acids Res. 2015 January 28; 43(Database issue): D68–D75. Published online 2014 November 20. doi: 10.1093/nar/gku1199