­­

Title of your bachelor’s thesis

Investigating the small to target the big

Oliveira, S. de

ID: i6255125

Bachelor Biomedical Sciences

Faculty of Health, Medicine and Life Sciences (FHML)

Maastricht University

Bachelor’s thesis

Supervisor: Dr. Susan Coort

Dr. Martina Summer-Kutmon

Department of Bioinformatics (BiGCAT)

Internship period (BBS3006): 6th April – 3rd July

3rd of July, 2020

Table of Contents

[Abstract 2](#_Toc44335337)

[Introduction  3](#_Toc44335338)

[Materials and Methods 5](#_Toc44335339)

[1. *Patient data* 5](#_Toc44335340)

[2. *Metagenomic data* 7](#_Toc44335341)

[3. *Metabolomic data* 8](#_Toc44335342)

[4. *Proteomic data* 9](#_Toc44335343)

[5. *Analysis of the biological implications of metabolomic and proteomic differences* 9](#_Toc44335344)

[6. *Integrative analysis of the metabolic and metagenomic data* 10](#_Toc44335345)

[Results 11](#_Toc44335346)

[*Identification of differences in the gut microbiome of IR and IS subjects* 12](#_Toc44335347)

[*Identification of differences in metabolome of IR and IS subjects* 15](#_Toc44335348)

[*Identification of differences in proteome of IR and IS subjects* 16](#_Toc44335349)

[*Identification of altered biological pathways involving these metabolomic and proteomic changes* 17](#_Toc44335350)

[*Identification of associated microbes and metabolites* 19](#_Toc44335351)

[Discussion 20](#_Toc44335352)

[Conclusion 23](#_Toc44335353)

[References 24](#_Toc44335354)

[Appendix 29](#_Toc44335355)

[*Appendix A:* 29](#_Toc44335356)

[*Appendix B:* 30](#_Toc44335357)

[*Appendix C*: 30](#_Toc44335358)

[*Appendix D*: 31](#_Toc44335359)

[*Appendix E:* 33](#_Toc44335360)

[*Appendix F*: 35](#_Toc44335361)

[Add-on page: Description of work by student Bachelor BMS 36](#_Toc44335362)

## Abstract

*Introduction*: Insulin resistance (IR) is an important accelerating factor in the development of type 2 diabetes (T2D), making it imperative to understand factors influencing it. Recently, specific changes in the omics of IR individuals have been discovered and shown to affect one another. Nevertheless, general changes in the omics of IR individuals as well as how these changes are related, is still uncertain.

*Aim*: To determine differences in the microbiome, host proteome and host metabolome between IR and insulin sensitive (IS) pre diabetics as well as investigate any biological consequences of these changes.

*Methods*: Data from the integrative Human Microbiome Project’s study concerning T2D was used, with only individuals classified as IR or IS being included. Group separation concerning the microbiome was visualised and factors contributing to this separation were investigated. Significant changes in metabolite and protein abundance were determined and the possible biological implications investigated. Finally, potential host metabolome-microbiome interactions were explored using an integrative analysis.

*Results*: Significant separation in the metagenomes of the 2 groups was observed. A lower mean abundance of *Bacteroidetes*, *Firmicutes* and *Proteobacteria* in IR were found to be potential explanations for this observation. Some of the microbes contributing most to the metagenomic separation supported these changes, while others contradicted them.

40 metabolites and 23 proteins were found to be significantly altered. Many were implicated in immune function, lipid homeostasis or energy homeostasis and 2 of the significantly altered metabolites were also generated by the metagenomic-metabolomic integrative analysis.

*Conclusion*: The microbiome, host proteome and host metabolome were shown to be significantly different between the IR and IS subjects. Several of the significantly altered metabolites and proteins were associated with some biological consequences of IR. Future research into these proteins/metabolites and the functional implications of the microbiome alterations could uncover preventative strategies for T2D.

*Keywords*: pre diabetes, insulin resistance, metabolomics, proteomics, gut microbiome.

## Introduction

Diabetes is one of the major chronic diseases burdening the present-day healthcare systems and is predicted to continue to increase in prevalence, with a 51% expected rise in cases from 2019 to 2045 (1). Type 2 diabetes will likely comprise 90-95% of these cases (2), thus, making it a much more pressing public-health concern than type 1. Despite its high prevalence, the biological mechanism underlying T2D development and the physiological changes that occur as a result are not entirely understood. Nevertheless, it is generally accepted that T2D is commonly characterized by insulin resistance (IR) (<https://www.sciencedirect.com/science/article/pii/S0002934300003375?casa_token=-RG1YMAzA2cAAAAA:nRyO_ACwnjbxmnjIt9M0UxPvmREJH6SwtsV7puDp6IQeOunMWIRBecWi0fPjlndHbTE1tmFaNqk>, <https://www.sciencedirect.com/science/article/pii/S0002914902025547?casa_token=n4eCjP-n8RQAAAAA:jD7lmNnhx0s08Y01babrl5uVpVY_FRfuugbk43EKbvYgUEEmJiogCg5YRYzda1MVbn0vQr_v520>).

IR is a state in which the body’s cells respond in a less-than-adequate way to a given concentration of insulin. It can be caused by a variety of factors of which ectopic fat storage from excessive caloric intake is the most widely accepted (3). Ectopic fat storage refers to the storage of fat in non-adipose tissues and can induce IR in these tissues via interference with insulin signalling (4). In this way, IR can develop in a range of tissues separately (5) but is mainly discussed in regards to skeletal muscle, the liver and adipose tissue due to their critical involvement in metabolism (6). Development of IR in these tissues results in decreased glucose uptake and increased glycogenolysis/gluconeogenesis (7), both of which contribute to increased plasma glucose levels. Consequently, a need for increased insulin secretion ensues, placing strain on the pancreatic beta cells and eventually resulting in their dysfunction (8). In this way, although the exact degree to which IR plays a role in T2D development is not yet fully understood (<https://link.springer.com/article/10.1007/s00125-002-1009-0>), there is no doubt that IR is an important accelerating factor. Hence, great efforts are being made in order to investigate variables that are able to influence insulin sensitivity.

For instance, recent advancements in high throughput sequencing technologies have allowed for the identification of microbes and specific microbiome compositions associated with T2D and IR (9, 10). These associations can most likely be explained by the ability of the microbiome to influence the host metabolome and proteome (11, 12). However, IR itself – whether induced by the microbiome or not – has been characterized by its own metabolic and proteomic changes.

One of the well-known metabolic consequences linked to IR is hypertriglyceridaemia (13). This dyslipidaemia is characterized by an increase in plasma free fatty acids (FFAs) and very low density lipoproteins (VLDLs) and is most likely the result of decreased insulin-dependent suppression of lipolysis in IR adipose tissue (14). However, other non-lipid metabolites have also been shown to define the IR metabolome. Recently, branched chain amino acids (BCAAs) have become a topic of interest with some studies showing an association between increased plasma levels of these compounds and IR (15, 16). These changes are likely the result of an alteration in the expression of genes involved in BCAA catabolism (17) and, therefore, provide supporting evidence for not only a metabolomic but also a proteomic signature of IR (18).

Even though advances have been made in the identification of specific microbes, metabolites and gene products correlated with IR and T2D, general changes in the omics of insulin resistant individuals and the biological implications of these changes remain unclear. For this reason, the current study aims to investigate the differences in the microbiome, host proteome and host metabolome between insulin resistant and insulin sensitive (IS) prediabetic subjects. Changes in these variables will then be used to identify and examine the biological consequences at pathway level. Since previous research has managed to establish IR specific biological signatures, it is hypothesized that there will be differences between the IS and IR groups and that these differences will be able to explain some of the biological characteristics of each group.

In order to test this hypothesis, data from the integrative Human Microbiome Project’s (iHMP) study concerning T2D was used. This study aims to research 106 individuals at high risk for diabetes over a period of 4 years in order to determine the ‘physiological changes that occur in the microbiome and host during viral infection and during changes in glucose levels and insulin resistance’ (19). Differences in the corresponding proteomic and metabolic data of the IR and IS groups were analysed using differential analyses (DAs) and any biological implications of these changes were evaluated using a combined pathway analysis and a network analysis. Analysis of the metagenomic separation between the IR and IS groups was done using a Principal Coordinate Analysis (PCoA) as well as by determining which phyla and Operational Taxonomic Units (OTUs) contribute most to this separation. Finally, the microbes and metabolites able to best explain the covariation between the metabolomic and metagenomic datasets were identified.

Materials and Methods

### *Patient data*

#### 1.1 Data acquisition

Proteomic, metabolomic, metagenomic and corresponding subject data of the integrative Human Microbiome Project’s study on T2D (19, 20) was used and is available from <http://hmp2-data.stanford.edu/>. The complete study design of the iHMP’s T2D project is shown in figure 1. Subject data consisted of steady-state plasma glucose (SSPG, mg/dl) measurements, race, age, gender, BMI and classification as either IR or IS. Subject classification related to the SSPG measurements, with a SSPG < 150 mg/dl being considered IS and a SSPG ≥ 150 mg/dl being considered IR.

At each visit blood, urine and faecal samples were taken and clinical laboratory tests were performed. Blood samples were fractionated into peripheral blood monocytes (PBMCs), plasma as well as serum, with the plasma being used to quantify the host metabolome (involving microbial metabolites) and the PBMCs for the host proteome. Stool samples were used to profile the microbial metagenome and urine samples were solely used to investigate the host microbial metabolite abundances. Proteomics and metabolomics were performed using SWATH-MS (21) and LC-MC/MS (22), respectively. Microbial taxa from stool samples were identified using 16s sequencing (23).



*Figure 1:* ***the sample population and general sampling protocol of the iHMP’s T2D project.*** *The data utilized in this study is highlighted in yellow. IR = insulin resistant. IS = insulin sensitive. SSPG =* *steady-state plasma glucose. PBMC = peripheral blood monocytes.*

#### 1.2 Pre-processing of subject data

In order to include only subjects whose insulin sensitivity had been recorded in the analyses, the original sample population was filtered down to only these subjects. Furthermore, the resulting sample population was further filtered to only include subjects that were present in all metabolomic, proteomic and metagenomic data downloaded from the iHMP website. The final list of subject IDs was used in the subsequent pre-processing of the metabolomic, proteomic and metagenomic datasets so that only classified individuals were used in the analyses.

### *Metagenomic data*

#### 2.1 Data pre-processing of the metagenomic data

A pre-existing phyloseq object included in the HMP2Data *Bioconductor* R package (24) was used for the analysis of the metagenomic data. This object was constructed in such a way to allow for use of the *phyloseq* R package (25) which contains many tools to analyse and visualise complex microbial data clustered into OTUs.

In order to perform the integrative analysis of the metabolomic and metagenomic data, the metagenomic data contained in the phyloseq object had to made consistent with the metabolomic data. This was done by 1) comparing the sample IDs in the metagenomic and metabolomic datafiles, 2) adjusting those in the metabolomic datafile to match the phyloseq and 3) using the processed metabolomic datafile to subset the samples of the phyloseq object. It should be noted that during the adjustment process it was discovered that some sample IDs in the metabolomic dataset corresponded to multiple metagenomic sample IDs. In this scenario, to avoid making any incorrect assumptions, the sample ID was excluded from the metabolomic datafile and, hence, from the metagenomic data.

Additional sample filtration involved exclusion of supposed outliers based on a Principal Component Analysis (PCA) plot comparing the IS and IR metagenomic data (Appendix C). Some samples were seen to be separated from the main cluster of samples. Since most of these samples possessed an Axis1 value < -2.8, only samples with an Axis1 value **>** -2.8 were included in the metagenomic data.

Finally, the taxa of the phyloseq object were filtered. First, only taxa with family and genus specified were included. A prevalence threshold was then applied to a subset of the IR and IS samples separately to remove the taxa that were not present in at least 10% of these samples. The remaining taxa in each group were then used to prune the taxa of the entire phyloseq object. This prevented the removal of taxa which could be separating the 2 groups. Taxa were also filtered to only include those which were present in at least 40 samples possessing an abundance of 2. This number of samples was chosen to complement the previously applied prevalence threshold. Lastly, any taxa with an abundance sum of 4 or less across all samples were excluded and any abundance values greater than 50 were reduced to a value of 50.

#### 2.2 Statistical analysis of the metagenomic data

To determine whether the microbiomes of the IR and IS group are distinct, the *phyloseq* R package was used to perform a Principal Coordinate Analysis (PCoA) with the Bray-Curtis dissimilarity on the log(1+x) transformed metagenomic data. A multivariate ANOVA with permutations (PERMANOVA) was then carried out on the phyloseq relative abundances to investigate whether the differences suggested by the PCoA were significant. Finally, a multivariate homogeneity check of the group dispersions was performed to determine whether a significant difference in the variance of the 2 groups could be an explanation for any separation seen in the PCoA. These last 2 analyses were done using the *vegan* R package (26).

From this, the variation in phylum abundance between the samples of each group was investigated using bar plots. Finally, in order to examine which taxa were more and less abundant in the groups, box plots illustrating the mean abundance of each phylum were created. All differences in mean abundance were investigated using Wilcoxon tests. Moreover, the top taxa separating the 2 groups were able to be identified using the results of the previously run PERMANOVA.

### *Metabolomic data*

#### 3.1 Data pre-processing of the metabolomic data

For some of the measured metabolites no identifier was given, however, a majority were annotated with HMDB (Human Metabolite Database) identifiers (27). Since a well-recognized annotation system was needed for follow-up analysis, the metabolomic data was filtered to only include metabolites for which an HMDB identifier was given. One abundance value for each metabolite-subject combination was then determined by averaging together the corresponding samples per metabolite. This allowed for the metabolomic data to be used along with the subject data in a differential analysis (DA). However, before the DA could be run, these averages needed to be normalized. This was done by variance stabilization using the *MetaboDiff* R package (28).

#### 3.2 Differential analysis of the metabolomic data

A DA of the metabolomic data was executed using *MetaboDiff* by applying a student’s t-test and the Benjamini-Hochberg (BH) multiple testing correction procedure. This allowed for identification of metabolites likely to be differentially abundant between the 2 conditions.

### *Proteomic data*

#### 4.1 Data pre-processing of the proteomic data

As with the metabolomic data, the proteomic data had to be made consistent with the subject data file in order to run the DA. This was achieved in the same way: the samples from each subject in the proteomic data were averaged together per protein in order to obtain one abundance value for each subject-protein combination.

All proteins were annotated using HGNC (HUGO Gene Nomenclature Committee) identifiers (29).

#### 4.2 Differential analysis of the proteomic data

Similar to the analysis of the metabolomic data, a DA of the proteomic data was done to achieve the same end goal but in relation to proteins. However, instead of *MetaboDiff*, the *limma* R package (30) was used to apply empirical Bayes statistics and the BH multiple testing correction procedure.

### *Analysis of the biological implications of metabolomic and proteomic differences*

#### 5.1 Combined pathway analysis of the metabolomic and proteomic data

A combined pathway analysis was executed on the output of both the metabolomic DA and proteomic DA and was carried out using PathVisio v3.3.0 (31) and MetaboAnalyst v.3.0 (32). This allowed for the identification of the biological pathways most likely altered in the IR subjects compared to the IS subjects.

##### 5.1.A. Data pre-processing

Since the compounds in both DA output files were not annotated with identifiers from commonly used databases, the compound IDs were changed so that they could be recognized by the utilized pathway analysis tools.

In respect to the metabolome DA results, identifiers unique to the T2D iHMP study were replaced with the corresponding HMDB identifiers found in the metabolite annotation datafile. Like with the other data, this datafile is available on the iHMP website. However, some of the HMDB identifiers in the metabolite annotation datafile were not able to be recognized by PathVisio. These identifiers were changed to HMDB identifiers that could be recognized by PathVisio (Appendix D).

As for the proteome DA results, most identifiers corresponded to HGNC annotations but some did not correspond directly. As there was no datafile related to the protein identifiers on the iHMP website, a search of the non-HGNC identifiers using well-known databases (such as KEGG (33) and ChEBI (34)) was done to replace them with an equivalent HGNC identifier (Appendix E).

The resulting metabolomic and proteomic data containing the correct identifiers were then combined into one Excel file in order to be used with PathVisio.

##### 5.1.B. Pathway analysis

The PathVisio pathway analysis was done using the combined DA results and using the BridgeDb mapping databases for metabolites (metabolites\_20192025) (35) (download date: 17/04/20) and human gene/protein IDs (Hs\_Derby\_Ensembl\_91) (36) (download date: 19/04/20). An expression criterion of p-value < 0.05 was specified and pathways were sourced from the WikiPathways database of human pathways (download date: 07/04/2020) (37). Statistical analysis involved an overrepresentation analysis of the metabolites and proteins in these pathways. This analysis used Fisher’s exact test as its statistical method and produced Z-scores for each pathway in the WikiPathways database. As a result, pathways containing more proteins/metabolites meeting the expression criterion than expected were able to be recognized.

An additional joint pathway analysis using MetaboAnalyst was executed using only the identifiers of the compounds deemed to be significantly altered by the DAs (p-value < 0.05). For this analysis, the latest KEGG pathway database (2019) was applied.

#### 5.2 Network Analysis

In order to further investigate any biological pathways that involve the significantly altered proteins, a network analysis of the proteomic data using Cytoscape v. 3.7.2 (38) was performed. This was done by first running the significantly altered proteins (p < 0.05) generated by the DA through the Cytoscape stringApp (39) (confidence score = 0.4) to create a protein-protein interaction network. The resulting network was then extended with protein-pathway interactions using the WikiPathways link set in the Cytoscape CyTargetLinker app (40).

This extended network was then analysed to determine the out-degree per pathway node. The pathways with the highest out-degree were of interest as they involved the most significantly altered proteins. To more easily visualize the relevancy of the altered pathways, out-degree was linked to the size of the corresponding pathway node.

### *Integrative analysis of the metabolic and metagenomic data*

#### 6.1 Data pre-processing of the metabolomic data

No further filtering of the metagenomic data was performed in order to carry out the integrative analysis of the metabolomic and metagenomic data, however, the metabolomic data needed slight adjustments in order to be suitable.

Like with the metagenomic datafile, samples were excluded based on the PCA (Appendix C) with only those corresponding to an Axis1 value > -2.8 being included. Furthermore, the metabolites were filtered to only include those that had an abundance of 0 in no more than 3 samples. The final abundances in the metabolomic datafile were then log(x+1) transformed to weaken the heavy tails and so allow for a sounder statistical analysis.

#### 6.2 Sparse Canonical Correlation Analysis of the metabolomic and metagenomic data

The filtered phyloseq object and matching metabolomic datafile were used to perform a sparse Canonical Correlation Analysis (sparse CCA) which allowed for the recognition of metabolites and microbes able to best explain the covariation between the datasets. The *PMA* (Penalized Multivariate Analysis) R package (41) was used to execute this analysis and a penalty of 0.15 was applied to both the metagenome and metabolome matrices.

The metagenomic analyses, the DAs and the integrative analysis of the metabolomic and metagenomic analysis were all performed using R-3.6.3 (42). The corresponding R scripts can all be downloaded from <https://github.com/sabdeolive/Sabrina_Bachelor-Thesis_BiGCAT>.

## Results

From the pre-processing of the subject datafile, 60 subjects remained of which 32 were IR and 28 were IS. The corresponding subject IDs were used to filter the proteome and metabolome data for the DAs, combined pathway analysis and network analysis. This resulted in 686 samples (316 IR and 370 IS) being included in the proteome data and 726 samples (425 IR and 302 IS) in the metabolome data. Moreover, the proteome data included 302 proteins annotated with HGNC symbols while the metabolome data was reduced from 724 metabolites to 323 metabolites after excluding those with no HMDB identifier.

After making the metagenomic and metabolomic datafiles consistent in order to run the integrative analysis, both datasets contained 49 subjects (26 IR and 23 IS) and 441 samples. This number of samples was reduced to 402 (200 IR and 202 IS) after excluding those with a PCA Axis1 value < -2.8. Despite excluding metabolites that had an abundance of 0 across many samples, the number of metabolites used in this integrative analysis remained the same (323 metabolites). Finally, filtering of the metagenomic data to only include sufficiently classified and abundant taxa, resulted in the phyloseq object containing 362 taxa.

### *Identification of differences in the gut microbiome of IR and IS subjects*

To determine whether differences in microbiome composition exist between the 2 groups, Bray-Curtis dissimilarities were illustrated using a PCoA. This PCoA demonstrated some clustering of the 2 groups, with 2 clusters being able to be distinguished per group (Figure 2). The separation seen in this plot was deemed to be significant by the PERMANOVA (p = 0.010) and was found to likely not be due to a significant difference in the variance of the groups by the multivariate analysis of the group dispersion homogeneity (p = 0.171). Nevertheless, there did exist a large variation in the absolute microbial abundances and absolute phyla abundances across both the IR and IS samples (Figure 3).



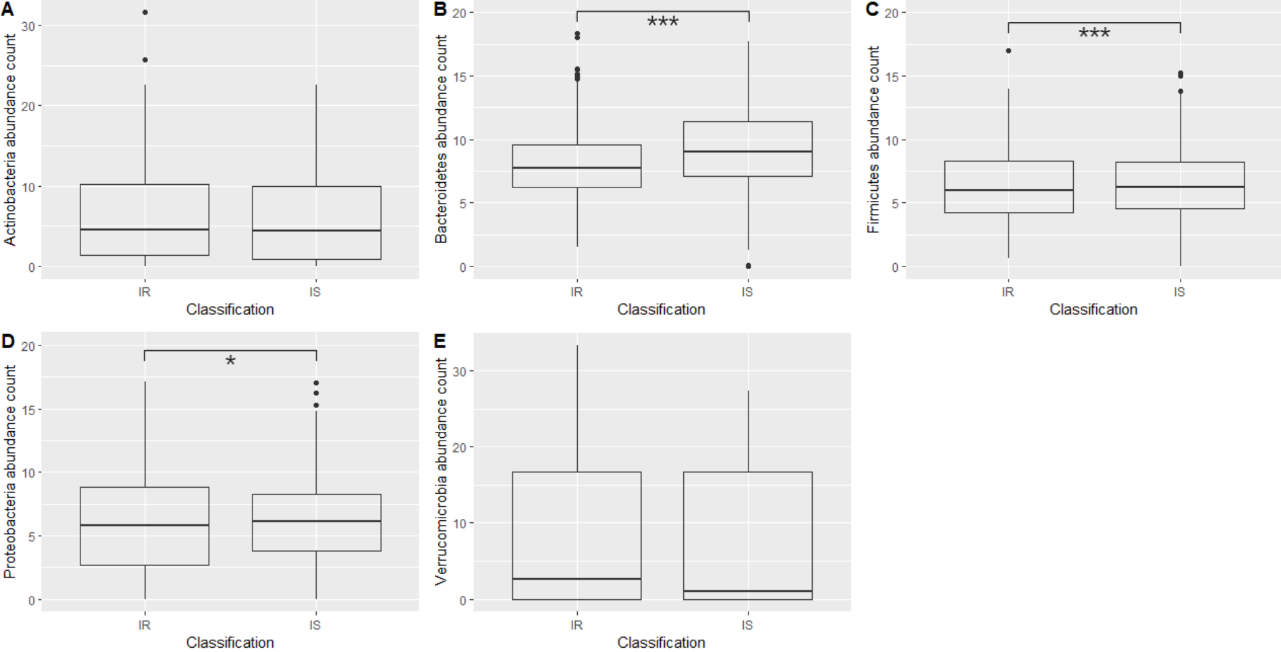
*Figure 2:* ***Principal Coordinate Analysis (PCoA) illustrating the separation in the microbial composition of the insulin resistant (IR) and insulin sensitive (IS) samples.*** *The PCoA is based on Bray-Curtis dissimilarity and the variation explained by the corresponding principal coordinates are given in %.*



*Figure 3:* ***Bar plots depicting the variation in the absolute microbial abundance and in the absolute phyla abundances across the A) insulin resistant (IR) samples and B) the insulin sensitive (IS) samples****.*

The mean abundance of each phyla in the different groups was investigated to determine which phyla were most likely to contribute to the separation between the IR and IS metagenomes. Small differences in mean abundance between the 2 conditions were observed for the *Actinobacteria*, *Firmicutes* and *Proteobacteria* phyla, with the *Firmicutes* (Figure 4C) and *Proteobacteria* (Figure 4D) both being slightly less abundant and the *Actinobacteria* (Figure 4A) slightly more abundant in the IR condition. However, only the differences in the *Firmicutes* and *Proteobacteria* phyla were significant (p< 0.0001 and p = 0.009, respectively).

More notable differences were seen for the remaining phyla. The *Bacteroidetes* were less abundant in IR compared to IS (Figure 4.B) while the *Verrucomicrobia* were more abundant in IR (Figure 4.E). However, only the difference in *Bacteroidetes* was significant (p < 0.0001).

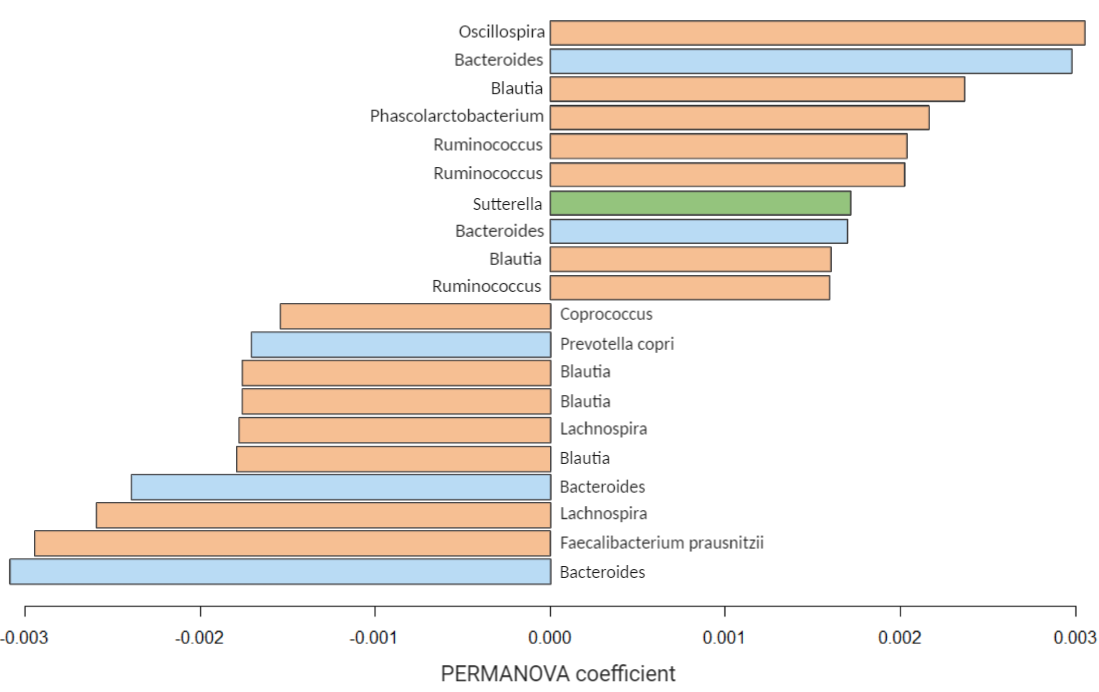


*Figure 4:* ***box plots representing the differences in mean phylum abundance between the insulin resistant (IR) and insulin sensitive (IS) group****. Significant p-values (p < 0.05) are indicated by \*\*\* = p < 0.0001, \*\* = p < 0.001 or \* = p < 0.01.*

To further investigate the taxa separating the groups, the top 20 taxa generated by the PERMANOVA were investigated in order to identify the OTUs contributing most to the metagenomic differences between the 2 groups. The majority (70%) of these OTUs belonged to the *Firmicutes* phylum.

Consistent with the *Bacteroidetes* mean abundance being lower in the IR group (Figure 4B), the microbe least abundant in the IR condition compared to the IS condition was of the *Bacteroidetes* phylum (Figure 5). However, the microbe most abundant in the IR samples was found to belong to the *Firmicutes* phylum (Figure 5). This contradicted the observation of a lower *Firmicutes* mean abundance in the IR group (Figure 4C).

Three of the 10 microbes more abundant in IR and none of those less abundant in IR were *Ruminococcus*. On the other hand, 2 out of the 10 microbes less abundant in IR and none of those more abundant in IR were Lachnospira. (Figure 5)



*Figure 5:* ***bar plot of the Operational Taxonomic Units (OTUs) contributing most to the metagenomic separation between the insulin resistant (IR) and insulin sensitive (IS) group****. All results included in the plot are from the multivariate ANOVA with permutations (PERMANOVA). For each OTU, the corresponding genus and species (if available) is given. OTUs with negative values are less abundant in the IR condition and OTUs with positive values are more abundant in the IS condition. Blue bars represent OTUs belonging to the Bacteroidetes phylum, orange bars represent OTUs belonging to the Firmicutes phylum and green bars represent OTUs belonging to the Proteobacteria phylum.*

### *Identification of differences in metabolome of IR and IS subjects*

The DA using MetaboDiff generated 40 metabolites whose abundances were significantly altered in the IR group compared to the IS group (p < 0.05): 21 were less abundant (logFC < 0) and 19 more abundant in the IR condition (logFC > 0) (Appendix F). Based on a logFC > 0.50 or < -0.50, 9 out of the 21 less abundant metabolites and 3 out of the 19 more abundant metabolites were sufficiently changed (Table 1). Eleven of these 12 sufficiently changed metabolites were lipids/lipid-like molecules with the only exception being glucaric acid. Seven metabolites possessed a significant adjusted p-value (Appendix F) with 5 also having a logFC suggesting a sufficient change in abundance (Table 1).

*Table 1: table specifying the sufficiently less abundant (logFC < -0.50, blue cells) and sufficiently more abundant (logFC > 0.05, red cells) metabolites in insulin resistance (IR).*

|  |  |  |
| --- | --- | --- |
| Metabolite (metabolite type) | logFC | Adjusted p-value |
| LysoPE(0:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z)) (Lysophospholipid) | -1.336 | 0.005 |
| LysoPE(0:0/16:0) (Lysophospholipid) | -1.010 | 0.281 |
| LysoPE(P-16:0/0:0) (Phospho-ether lipid) | -1.089 | 0.042 |
| Ethylmalonic acid | -0.795 | 0.040 |
| 2-arachidonoylglycerol | -0.758 | 0.005 |
| 5-dodecenoic acid | -0.720 | 0.162 |
| MG(0:0/14:1(9Z)/0:0) (Monoacylglycerol) | -0.683 | 0.026 |
| LysoPE(0:0/20:3(11Z,14Z,17Z)) (Lysophospholipid) | -0.591 | 0.158 |
| LysoPE(0:0/22:0) (Lysophospholipid) | -0.553 | 0.326 |
| Glucaric acid | 0.556 | 0.326 |
| 2-trans,4-cis-Decadienoylcarnitine (Fatty acid ester) | 0.604 | 0.106 |
| LysoPC(20:0/0:0) (Lysophospholipid) | 0.867 | 0.326 |

*Data was generated from a differential analysis (DA) using MetaboDiff. All listed metabolites listed were significantly changed in the IR condition compared to the IS condition based on p value (p < 0.05). The common name, logFC and adjusted p-value are given for each metabolite.*

### *Identification of differences in proteome of IR and IS subjects*

From the DA using *limma*, 23 proteins were significantly altered in the IR condition compared to the IS condition based on a p-value < 0.05. Thirteen of these proteins were more abundant in the IR subjects (logFC > 0) and 10, less abundant (logFC < 0) (Appendix G). When a sufficient change was deemed to be a logFC > 0.50 or < -0.50, 3 of the significantly altered proteins remained: LPA and SHBG were both less abundant in the IR condition (logFC = -1.074 and -0.522, respectively) while APOC4 was more abundant in the IR condition (logFC = 0.624). Despite these proteins having significant p-values, it should be noted that none possessed a significant adjusted p-value.

### *Identification of altered biological pathways involving these metabolomic and proteomic changes*

Combined pathway analysis of the proteome and metabolome data was done using PathVisio and MetaboAnalyst.

Out of all the 323 metabolites and 302 proteins analysed in PathVisio, only 15 metabolites and 25 proteins met the expression criterion (p < 0.05). The overrepresentation analysis of these compounds found 18 pathways to contain significantly more changed compounds than expected (Z-score > 1.96). All these pathways possessed 1-2 proteins/metabolites that had significantly different levels in the IR subjects compared to in the IS subjects (p < 0.05). The majority of the significantly altered pathways involved metabolites that met the expression criteria, but none contained both proteins and metabolites (Table 2).

*Table 2: List of the pathways determined by PathVisio to contain more changed proteins/metabolites than expected (Z > 1.96).*

|  |  |  |
| --- | --- | --- |
| Pathway | Z-score | Proteins and Metabolites |
| Gastric acid production | 3.05 | ↑ Acetylcholine |
| Heroin metabolism | 3.05 | ↑ Pseudo-ChE |
| IL-1 signalling pathway | 3.05 | ↓ IL1RAP |
| Irinotecan pathway | 3.05 | ↑ BCHE |
| Monoamine GPCRs | 3.05 | ↑ Acetylcholine |
| Oligodendrocyte Specification and differentiation (including remyelination), leading to Myelin Components for CNS | 3.05 | ↓ Sphingomyelin (**d18:0/18:1(11Z))** |
| Phosphodiesterases in neuronal function | 3.05 | ↑ L-Glutamate |
| Secretion of Hydrochloric Acid in Parietal Cells | 3.05 | ↑ Acetylcholine |
| Signal transduction through IL1R | 3.05 | ↓ IL1RAP |
| Sphingolipid pathway | 3.05 | ↓ Sphinganine |
| Structural Pathway of Interleukin 1 (IL-1) | 3.05 | ↓ IL1RAP |
| Thermogenesis | 3.05 | ↓ 2-Arachidonoylglycerol |
| Vitamin D-sensitive calcium signalling in depression | 3.05 | ↑ Acetylcholine |
| Ebola Virus Pathway on Host | 2.3 | ↓ GSN |
| Phosphatidylcholine catabolism | 2.3 | ↑ LysoPC(20:0/0:0)  ↓ Sphingomyelin (**d18:0/18:1(11Z))** |
| PPAR Alpha Pathway | 2.3 | ↓ APOA1  ↓ PLTP |
| Senescence and Autophagy in Cancer | 2.3 | ↑ VTN  ↓ GSN |
| Glutathione metabolism | 1.97 | ↑ L-Glutamate  ↑ 5-Oxoproline |

*For each pathway, the associated significantly altered compounds (p < 0.05) are given. The common name and HMDB identifier of each metabolite are provided, while only the HGNC protein identifiers are mentioned. Arrows indicate the compound being more abundant in insulin resistance (↑) or less abundant in insulin resistance (↓), in comparison to the insulin sensitive condition.*

MetaboAnalyst was unable to recognise one metabolite out of the 40 significantly altered metabolites (HMDB61112), meaning 39 metabolites and 23 proteins were used in the corresponding combined pathway analysis. Six pathways were found to be significantly changed (p < 0.05) in the IR subjects compared to the IS subjects. Like PathVisio, all these pathways concerned 1-2 significantly altered compounds, however, all were metabolites. Furthermore, the glutathione metabolism pathway as well as pathways associated with sphingolipids were present in the results of both programmes (Table 2-3).

*Table 3: List of the pathways determined by MetaboAnalyst to contain more changed proteins/metabolites than expected (Z > 1.96).*

|  |  |  |
| --- | --- | --- |
| Pathway | p-value | Proteins and Metabolites |
| Porphyrin and chlorophyll metabolism | 0.019 | **↓** Biliverdin (variant 1)  **↑** Biliverdin (variant 2)  **↑** L-Glutamate |
|  |
| Glutathione metabolism | 0.021 | **↑** L-Glutamate  **↑** 5-Oxoproline |  |
|  |
| Sphingolipid metabolism | 0.022 | **↓** Sphinganine  **↓** Sphingomyelin (d18:0/18:1(11Z)) |  |
|  |
| Nitrogen metabolism | 0.040 | **↑** L-Glutamate |  |
| D-Glutamine and D-glutamate metabolism | 0.040 | **↑** L-Glutamate |  |
| Glycerophospholipid metabolism | 0.046 | **↓** 1-Acyl-sn-glycero-3-phosphocholine  **↑** Acetylcholine |  |
|  |

*For each pathway, the associated significantly altered compounds (p < 0.05) are given. The common name and HMDB identifier of each metabolite are provided. Arrows indicate the compound being more abundant in insulin resistance (↑) or less abundant in insulin resistance (↓), in comparison to the insulin sensitive condition.*

Overrepresentation analysis can bury information concerning pathways that it deems to not be significantly altered. For this reason, an extended network analysis of the significantly altered proteins was performed to uncover alterations in biological pathways that may have been lost in the overrepresentation analysis. The results revealed the human complement system and the complement and coagulation cascade pathways to possess the highest out-degree values: the first of these pathways having a value of 5 and the second a value of 4. Both pathways involved CFH and F10, with the 3 remaining proteins involved in the human complement system being C4A, APOA1 and VTN and the 2 remaining proteins in the complement and coagulation cascade pathway being SERPIND1 and C1QB. All of these proteins were more abundant in the IR condition compared to the IS condition (logFC > 0), except for C4A and APOA1 which were less abundant in the IR condition (logFC < 0) (Appendix G).

### *Identification of associated microbes and metabolites*

The CCA used to determine which features best explain the covariation between the metabolomic and metagenomic data generated 13 microbes and 11 metabolites. 10 of these metabolites were monoacylglycerols (MAGs) with one of these being a unique MAG known as 2-arachidonoylglycerol (2-AG). The remaining 1 metabolite was arachidonic acid.

Two of the MAGs (MG(0:0/14:1(9Z)/0:0) and 2-AG) were also shown to have significantly different abundances between the 2 groups by the DA (p < 0.05).

All 13 of the covariance-explaining microbes were *Firmicutes*. This coincides with the absolute phylum abundances in each sample since a large proportion of all samples contained microbes belonging to the Firmicutes phylum (Figure 3). 6 of the 13 *Firmicutes* were *Faecalibacterium* *prausnitzii*, 3 were *Oscillospira* and 2 *Coprococcus*, with the remaining 2 taxa being *Dorea* and *Blautia* *producta*.

However, it should be noted that there did not exist a strong relationship between the metagenomic and metabolomic datasets (correlation coefficient = 0.455).

## Discussion

The current study used systems biology approaches on iHMP T2D data to investigate the differences in the microbiome, host proteome and host metabolome between insulin sensitive and insulin resistant pre-diabetics. The results were extended upon using functional and integrative analyses which allowed for interpretation within a biological context. As hypothesized, changes in all 3 of the investigated omics data types were observed. Twenty-three of the 302 measured proteins and 40 of the 323 HMDB-annotated metabolites were seen to have significantly different abundances in the IR and IS group. Many of these changes were associated with immune function, lipid homeostasis or energy homeostasis and more than half of the significantly altered metabolites (23/40) were lipids/lipid-like molecules.

Concerning the metagenome, there existed a significant separation between the groups potentially due to differences in the abundance of certain phyla and OTUs. Some of these distinguishing OTUs were linked to the metabolomic differences, hence, suggesting an involvement in the host (patho)physiology.

General differences in the IR and IS microbiome significantly related to the *Bacteroidetes*, *Firmicutes* and *Proteobacteria* phylum which coincides with the fact that these phyla are 3 of the most abundant phyla in the human gut (43). All were lower in IR condition but the *Bacteroidetes* phylum had a considerably greater difference between the groups compared to the other 2 phyla. The *Bacteroidetes*/*Firmicutes* ratio has been found to be decreased in T2D (44). This coincides with the present results even though this ratio change was not very large. However, a decreased *Bacteroidetes*/*Firmicutes* has also been associated with obesity and a high-fat diet (45). Since all participants included in the iHMP study had a BMI classifying them as overweight or obese, the ratio change seen in this study may not be as large as it could have been if all participants were lean. Furthermore, the fact that diet was not recorded could also have had an influence on this ratio change. Nevertheless, the small change observed in the present study is still consistent with previous literature and, hence, suggests that IR is able to influence the microbiome independent of BMI and diet.

However, it should be noted that metagenomic separation was unexpectedly observed within the IR and IS group themselves. This could have resulted from the large variation in absolute phyla abundances between the samples within each group and may have been caused by factors not taken into account. For instance, both ethnicity and diet have been shown to affect microbiome composition (46, 47) but were not considered in the current study and, thus, could have influenced the results. Nonetheless, investigation of the microbiome at OTU level rather than phylum level, produced results more consistent with previous literature.

For instance, the OTU more abundant in the IR condition that had the greatest discrepancy in abundance compared to the IS condition (*Oscillospira*), belongs to the *Firmicutes* phylum. This coincides with the finding that *Firmicutes* tend to be more abundant in IR individuals when compared to IS controls of similar BMI (48). Hence, considering the present study population, it may be more beneficial to investigate differences at a level of classification more specific than phylum. For instance, an OTU classified as *Faecalibacterium prausnitzii* was found to be less abundant in the IR condition and to be the microbe contributing the 2nd most to the metagenomic separation of the groups. *F. prausnitzii* is a species of butyrate producing bacteria whose presence has been shown to improve insulin sensitivity by protecting against inflammation-induced insulin resistance (49). Hence, a lower abundance of this species in IR subjects could potentially be an explanation for their insulin resistance. Moreover, almost half the OTUs that best explained the covariation between the metabolomic and metagenomic data were *F.prausnitzii*. This could suggest a possible link between the host metabolome and the presence of *F.prausnitzii*. For instance, one of the metabolites generated from the same analysis and which was deemed to be significantly less abundant in IR was 2-AG. Increased levels of 2-AG have been associated with administration of *Akkermansia muciniphila* (50), a bacteria that has been shown to cross-feed *F.prausnitzii (51)*. Therefore, the lower abundance of *F.prausnitzii* in IR may explain the lower 2-AG levels since it may be an indication of low *A. muciniphila*. Yet, this is more of a suggestion rather than a conclusion from the utilized analyses, especially considering the mediocre relationship between the  metagenomic and metabolic datasets.

Concerning 2-AG function, it is an endocannabinoid able to bind receptors involved in thermogenesis (52) as evidenced by the PathVisio results. Thermogenesis and other processes involved in energy homeostasis have been shown to be altered in IR states (53, 54).

In addition to thermogenesis, other pathways involved in energy homeostasis were also identified with pathway analysis. One of these being the D-glutamine and D-glutamate metabolism pathway. This pathway involved one significantly altered metabolite called L-glutamate, which was found to be significantly more abundant in the IR subjects of this study. This change in abundance coincides with previous literature showing an association between high glutamate levels and T2D (55), potentially due to the fact that it can be used as an alternative energy source (56).

However, this adaptive increase in glutamate levels may result in decreased insulin secretion by accelerating beta-cell dysfunction through excessive NMDA receptor activation (57). This reduced insulin secretion along with the presence of insulin resistance in the IR group would result in a further decreased glucose uptake. Consequently, hypertriglyceridemia would be intensified in the IR group, as explained in the introduction. Higher levels of lipoproteins in the blood may be an explanation for the alterations in the abundance of some apolipoproteins seen in this study. For instance, APOC4 was seen to be more abundant in the IR condition compared to the IS condition. APOC4 is mainly found on VLDL (58) and its overexpression has been associated with hypertriglyceridemia (59). Furthermore, two apolipoproteins found on HDL, APOA1 (60) and APOD (61), were decreased in the IR subjects compared to the IS subjects. This could be evidence of low HDL levels which too have been previously associated with insulin resistance (62, 63).  Other non-APO proteins have also been found on HDL particles, some of which are involved in the complement system such as C4A (64). Like APOA1 and APOD, C4A was less abundant in the IR group compared to the IS group of this study and so could indicate dyslipidaemia as seen with the apolipoproteins. Any alterations in the levels of complement proteins obviously have consequences on immune function. This is demonstrated by the fact that multiple pathways generated by the overrepresentation analyses were associated with the immune system. However, all these pathways possessed only a small number of significantly altered proteins/metabolites compared to the total number of components involved in the pathway. Therefore, it is possible that these pathways are not altered enough to produce a noticeable difference between the IR and IS group. Yet, the low count of significantly altered compounds involved in the pathway analysis results could be more of an indication of low coverage of the utilized tools rather than of a low-biological impact. Nevertheless, multiple tools were used to produce these results, hence, allowing for a more detailed representation of the potentially altered pathways. However, since these 2 analyses were unable to be performed simultaneously in one programme, the problem of coverage still exists within each of the toolsets. This problem could be minimized by the creation of one tool that encompasses the coverage of WikiPathways and KEGG.

Despite these issues concerning the interpretation of the pathway analysis, network analysis of the protein data was able to suggest pathways most likely modified in IR by identifying which pathways involved the largest number of significantly altered proteins. The 2 pathways identified using this analysis were both associated with the complement system, therefore, potentially indicating that the proteomic alterations in IR are most likely associated with noticeable differences in immune function.

Another explanation for the possibly low biological effect of the observed metabolite and protein changes may have to do with the fact that baseline samples from which insulin sensitivity classification was determined, were not exclusively used. This could mean that some of the subjects in the IS group may have transitioned to become insulin resistant but were still classified as IS in the present study. Consequently, some of the changes between the classifications may not be as strong as they could have been if only the baseline samples were used. This may also explain the small change in the previously discussed *Bacteroidetes*/*Firmicutes* ratio.

Regardless of these limitations, this study did produce results consistent with the pathophysiology of insulin resistance especially in relation to the metabolomic and proteomic analyses. Due to time limitations and limitations within the field of metagenomics, analysis of the metagenomic data was more superficial. Integrative analysis of this data and the metabolomics data of both the IR and IS group would allow for a better understanding of the host-microbiome interactions driving the differences between these groups. Furthermore, any biological implications of the observed metagenomic changes could be better determined by the development of reliable functional analysis methods/tools.

## Conclusion

In conclusion, a separation in the metagenomes of IR and IS subjects was observed. This separation seemed to mainly be explained by changes in *Bacteroidetes* abundance. Direct association between the microbiome composition and the host metabolome was unable to be determined by the present analyses, however, certain OTUs and metabolites were found to be possible explanatory factors of the covariation between the datasets.

Concerning the proteomic and metabolomic data, the abundance of several metabolites and proteins was significantly different in the IR condition compared to IS. Many of these compounds were associated with the immune system, lipid homeostasis or energy homeostasis. This confirms the presently accepted knowledge concerning the pathophysiology of IR and could potentially be used to target future research towards biomarkers and therapeutic strategies for IR.

## References

1. IFD. IDF Diabetes Atlas, 9th edn. Brussels, Belgium: International Diabetes Federation 2019. 2019.

2. Association AD. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2019. Diabetes Care. 2019;42(Supplement 1):S13-S28.

3. Samuel VT, Shulman GI. The pathogenesis of insulin resistance: integrating signaling pathways and substrate flux. The Journal of clinical investigation. 2016;126(1):12-22.

4. Snel M, Jonker JT, Schoones J, Lamb H, de Roos A, Pijl H, et al. Ectopic fat and insulin resistance: pathophysiology and effect of diet and lifestyle interventions. International journal of endocrinology. 2012;2012.

5. Trouwborst I, Bowser SM, Goossens GH, Blaak EE. Ectopic fat accumulation in distinct insulin resistant phenotypes; targets for personalized nutritional interventions. Frontiers in nutrition. 2018;5:77.

6. Freeman AM, Soman-Faulkner K, Pennings N. Insulin resistance. StatPearls [Internet]: StatPearls Publishing; 2019.

7. Ferrannini E, Gastaldelli A, Iozzo P. Pathophysiology of prediabetes. Medical Clinics. 2011;95(2):327-39.

8. Saisho Y. How Can We Develop More Effective Strategies for Type 2 Diabetes Mellitus Prevention? A Paradigm Shift from a Glucose-Centric to a Beta Cell-Centric Concept of Diabetes. DIABETES. 2018.

9. Gurung M, Li Z, You H, Rodrigues R, Jump DB, Morgun A, et al. Role of gut microbiota in type 2 diabetes pathophysiology. EBioMedicine. 2020;51:102590.

10. Lee CJ, Sears CL, Maruthur N. Gut microbiome and its role in obesity and insulin resistance. Annals of the New York Academy of Sciences. 2020;1461(1):37-52.

11. Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome medicine. 2011;3(3):14.

12. Zhang X, Deeke SA, Ning Z, Starr AE, Butcher J, Li J, et al. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. Nature communications. 2018;9(1):1-14.

13. Grundy SM. Hypertriglyceridemia, insulin resistance, and the metabolic syndrome. The American journal of cardiology. 1999;83(9):25-9.

14. Ginsberg HN. Insulin resistance and cardiovascular disease. The Journal of clinical investigation. 2000;106(4):453-8.

15. Klein MS, Shearer J. Metabolomics and type 2 diabetes: translating basic research into clinical application. Journal of diabetes research. 2016;2016.

16. Shaham O, Wei R, Wang TJ, Ricciardi C, Lewis GD, Vasan RS, et al. Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. Molecular systems biology. 2008;4(1):214.

17. Sears D, Hsiao G, Hsiao A, Yu J, Courtney C, Ofrecio J, et al. Mechanisms of human insulin resistance and thiazolidinedione-mediated insulin sensitization. Proceedings of the National Academy of Sciences. 2009;106(44):18745-50.

18. Giebelstein J, Poschmann G, Højlund K, Schechinger W, Dietrich J, Levin K, et al. The proteomic signature of insulin-resistant human skeletal muscle reveals increased glycolytic and decreased mitochondrial enzymes. Diabetologia. 2012;55(4):1114-27.

19. Integrative H. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell host & microbe. 2014;16(3):276.

20. Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, Leopold SR, et al. Longitudinal multi-omics of host–microbe dynamics in prediabetes. Nature. 2019;569(7758):663-71.

21. Collins BC, Hunter CL, Liu Y, Schilling B, Rosenberger G, Bader SL, et al. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. Nature communications. 2017;8(1):1-12.

22. Want EJ. LC-MS untargeted analysis. Metabolic Profiling: Springer; 2018. p. 99-116.

23. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proceedings of the National Academy of Sciences. 1985;82(20):6955-9.

24. Stansfield J, Dozmorov M. HMP2Data: 16s rRNA sequencing data from the Human Microbiome Project 2. version 1.0.0 ed2019.

25. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS one. 2013;8(4):e61217.

26. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. version 2.5-6 ed2019.

27. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic acids research. 2018;46(D1):D608-D17.

28. Mock A, Warta R, Dettling S, Brors B, Jäger D, Herold-Mende C. MetaboDiff: an R package for differential metabolomic analysis. Bioinformatics. 2018;34(19):3417-8.

29. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, et al. Genenames. org: the HGNC and VGNC resources in 2019. Nucleic acids research. 2019;47(D1):D786-D92.

30. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43(7):e47-e.

31. Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, et al. PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol. 2015;11(2):e1004085.

32. Chong J, Wishart DS, Xia J. Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. Current protocols in bioinformatics. 2019;68(1):e86.

33. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Science. 2019;28(11):1947-51.

34. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic acids research. 2016;44(D1):D1214-D9.

35. De Sl. Metabolite BridgeDb ID Mapping Database (20191025): figshare; 2019 [Available from: <https://doi.org/10.6084/m9.figshare.10048508.v1>.

36. BiGCaT. Gene/Protein BridgeDb ID Mapping Database (Ensembl 91) Zenodo; 2020 [Available from: <https://zenodo.org/record/3667670>.

37. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic acids research. 2018;46(D1):D661-D7.

38. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003;13(11):2498-504.

39. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: network analysis and visualization of proteomics data. Journal of proteome research. 2018;18(2):623-32.

40. Kutmon M, Ehrhart F, Willighagen EL, Evelo CT, Coort SL. CyTargetLinker app update: A flexible solution for network extension in Cytoscape. F1000Research. 2018;7.

41. Witten D, Tibshirani R. PMA: Penalized Multivariate Analysis. version 1.2.1 ed2020.

42. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.

43. Khanna S, Tosh PK, editors. A clinician's primer on the role of the microbiome in human health and disease. Mayo clinic proceedings; 2014: Elsevier.

44. Zhang Y, Zhang H. Microbiota associated with type 2 diabetes and its related complications. Food Science and Human Wellness. 2013;2(3-4):167-72.

45. Tilg H, Kaser A. Gut microbiome, obesity, and metabolic dysfunction. The Journal of clinical investigation. 2011;121(6):2126-32.

46. Gaulke CA, Sharpton TJ. The influence of ethnicity and geography on human gut microbiome composition. Nature medicine. 2018;24(10):1495-6.

47. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505(7484):559-63.

48. Moreno-Indias I, Sánchez-Alcoholado L, García-Fuentes E, Cardona F, Queipo-Ortuño MI, Tinahones FJ. Insulin resistance is associated with specific gut microbiota in appendix samples from morbidly obese patients. American journal of translational research. 2016;8(12):5672.

49. Ganesan K, Chung SK, Vanamala J, Xu B. Causal relationship between diet-induced gut microbiota changes and diabetes: a novel strategy to transplant Faecalibacterium prausnitzii in preventing diabetes. International journal of molecular sciences. 2018;19(12):3720.

50. Baptista LC, Sun Y, Carter CS, Buford TW. Crosstalk Between the Gut Microbiome and Bioactive Lipids: Therapeutic Targets in Cognitive Frailty. Frontiers in Nutrition. 2020;7:17.

51. Lopez-Siles M, Enrich-Capó N, Aldeguer X, Sabat-Mir M, Duncan SH, Garcia-Gil LJ, et al. Alterations in the abundance and co-occurrence of Akkermansia muciniphila and Faecalibacterium prausnitzii in the colonic mucosa of inflammatory bowel disease subjects. Frontiers in cellular and infection microbiology. 2018;8:281.

52. Krott LM, Piscitelli F, Heine M, Borrino S, Scheja L, Silvestri C, et al. Endocannabinoid regulation in white and brown adipose tissue following thermogenic activation. Journal of lipid research. 2016;57(3):464-73.

53. Camastra S, Bonora E, Del Prato S, Rett K, Weck M, Ferrannini E. Effect of obesity and insulin resistance on resting and glucose-induced thermogenesis in man. International journal of obesity. 1999;23(12):1307-13.

54. Goodpaster BH, Sparks LM. Metabolic flexibility in health and disease. Cell metabolism. 2017;25(5):1027-36.

55. Liu X, Zheng Y, Guasch-Ferré M, Ruiz-Canela M, Toledo E, Clish C, et al. High plasma glutamate and low glutamine-to-glutamate ratio are associated with type 2 diabetes: Case-cohort study within the PREDIMED trial. Nutrition, Metabolism and Cardiovascular Diseases. 2019;29(10):1040-9.

56. Roberts LD, Koulman A, Griffin JL. Towards metabolic biomarkers of insulin resistance and type 2 diabetes: progress from the metabolome. The lancet diabetes & endocrinology. 2014;2(1):65-75.

57. Huang X-T, Li C, Peng X-P, Guo J, Yue S-J, Liu W, et al. An excessive increase in glutamate contributes to glucose-toxicity in β-cells via activation of pancreatic NMDA receptors in rodent diabetes. Scientific reports. 2017;7:44120.

58. Kotite L, Zhang L, Yu Z, Burlingame A, Havel R. Human apoC-IV isolation, characterization, and immunochemical quantification in plasma and plasma lipoproteins. Journal of lipid research. 2003;44(7):1387-94.

59. Allan C, Taylor J. Expression of a novel human apolipoprotein (apoC-IV) causes hypertriglyceridemia in transgenic mice. Journal of lipid research. 1996;37(7):1510-8.

60. Chistiakov DA, Orekhov AN, Bobryshev YV. ApoA1 and ApoA1-specific self-antibodies in cardiovascular disease. Laboratory Investigation. 2016;96(7):708-18.

61. Perdomo G, Dong HH. Apolipoprotein D in lipid metabolism and its functional implication in atherosclerosis and aging. Aging (Albany NY). 2009;1(1):17.

62. Karhapää P, Malkki M, Laakso M. Isolated low HDL cholesterol: an insulin-resistant state. Diabetes. 1994;43(3):411-7.

63. Li N, Fu J, Koonen DP, Kuivenhoven JA, Snieder H, Hofker MH. Are hypertriglyceridemia and low HDL causal factors in the development of insulin resistance? Atherosclerosis. 2014;233(1):130-8.

64. von Zychlinski A, Kleffmann T. Dissecting the proteome of lipoproteins: New biomarkers for cardiovascular diseases? Translational Proteomics. 2015;7:30-9.

## Appendix

### *Appendix A:*

Table A1: List of abbreviations

|  |  |
| --- | --- |
| T2D | Type 2 diabetes |
| IR | Insulin resistance/insulin resistant |
| IS | Insulin sensitive |
| iHMP | Integrative Human Microbiome Project |
| FFA | Free fatty acids |
| VLDL | Very low density lipoproteins |
| BCAA | Branched chain amino acids |
| PBMC | Peripheral blood monocytes |
| PCoA | Principal Coordinates Analysis |
| PCA | Principal Component Analysis |
| DA | Differential Analysis |
| CCA | Canonical Correlation Analysis |
| OTU | Operational Taxonomic Unit |
| PERMANOVA | Multivariate ANOVA with permutations |
| HMDB | Human Metabolite Database |
| BH | Benjamini-Hochberg |
| HGNC | HUGO Gene Nomenclature Committe |
| PMA | Penalized Multivariate Analysis |
| MAG | Monoacylglycerol |
| 2-AG | 2-arachidonoylglycerol |

### *Appendix B:*



Figure A1: **study design**. iHMP = integrative human microbiome project. T2D = type 2 diabetes. PCoA = Principle Coordinate Analysis. PERMANOVA = Multivariate ANOVA with Permutations. Sparse CCA = sparse Canonical Correlation Analysis.

### *Appendix C:*



Figure A2: **Principal Component Analysis (PCA) plot of the metagenomic data prior to outlier exclusion**. The explained variance of the principal components are given as a percentage by the corresponding axis. OTU = taxa. IR = insulin resistant. IS = insulin sensitive.

### *Appendix D*:

Table A2: List of the HMDB metabolite identifiers originally found in the iHMP T2D metabolomic abundance data and of the corresponding HMDB identifiers used in order to run the combined pathway analysis.

|  |  |
| --- | --- |
| Original HMDB Identifier(s) | HMDB identifier used for analysis |
| HMDB03736 | HMDB0000491 |
| HMDB01918 | HMDB0000248 |
| HMDB02994|HMDB04136 | HMDB02994 |
| HMDB00172|HMDB00687 | HMDB00172 |
| HMDB11756|HMDB61684 | HMDB11756 |
| HMDB28783|HMDB29014| | HMDB28783 |

### *Appendix E*:

Table A3: List of non-HGNC protein identifiers found in the iHMP T2D proteome abundance data and of the corresponding HGNC identifiers used in order to run the combined pathway analysis.

|  |  |
| --- | --- |
| Original identifier | HGNC identifier used for analysis |
| PRG4.1 | PRG4 |
| IL1RAP.1 | IL1RAP |
| KNG1\_2 | KNG1 |
| HV169 | IGHV1-69 |
| IGJ | JCHAIN |
| HV439 | IGHV4-39 |
| KV320.2 | IGKV3-20 |
| HV307\_2 | IGHV3-7 |
| KV116 | IGKV1-16 |
| HV333 | IGHV3-33 |
| HV353 | IGHV3-53 |
| HV323 | IGHV3-23 |
| KV320 | IGKV3-20 |
| LV147 | IGLV1-47 |
| HV102 | IGHV1-2 |
| KV133 | IGKV1-33 |
| LV151 | IGLV1-51 |
| LV657\_2 | IGLV6-57 |
| GPR116 | ADGRF5 |
| LV321.1 | IGLV3-21 |
| KVD16 | IGKV1D-16 |
| KVD33\_2 | IGKV1D-33 |
| LV657 | IGLV6-57 |
| CLU.1 | CLU |
| KVD28 | IGKV2D-28 |
| LV140 | IGLV1-40 |
| KV315 | IGKV3-15 |
| HV146 | IGHV1-46 |
| LV211 | IGLV2-11 |
| HV333\_2 | IGHV3-33 |
| KVD33\_3 | IGKV1D-33 |
| LV144 | IGLV1-44 |
| KV139 | IGKV1-39 |
| ATRN.1 | ATRN |
| HV270 | IGHV2-70 |
| HV313 | IGHV3-13 |
| HV330 | IGHV3-30 |
| SEPP1 | SELENOP |
| HV434 | IGHV4-34 |
| LV743 | IGLV7-43 |
| HV307 | IGHV3-7 |
| HV330\_2 | IGHV3-30 |
| KV310 | IGKV3-20 |
| KVD33\_4 | IGKV1D-33 |
| HV348 | IGHV3-48 |
| KVD33 | IGKV1D-33 |
| KV320.1 | IGKV3-20 |
| LV319 | IGLV3-19 |
| FBLN1.1 | FBLN1 |
| KV311 | IGKV3-11 |
| LV325 | IGLV3-25 |
| KV230 | IGKV2-30 |
| IGHM.1 | IGHM |
| KV320\_2 | IGKV3-20 |
| LV321 | IGLV3-21 |

### *Appendix F:*

Table A4: List of metabolites with significantly different abundances (p < 0.05) between the insulin resistant and insulin sensitive subjects.

|  |  |  |  |
| --- | --- | --- | --- |
| Metabolite | logFC | p-value | Adjusted p-value |
| HMDB04666  2-arachidonoylglycerol (2-AG) | -0.758 | 0.000 | 0.005 |
| HMDB11496  **LysoPE(0:0/22:6(4Z,7Z,10Z,13Z,16Z,19Z))** | -1.336 | 0.000 | 0.005 |
| HMDB00705  Hexanoylcarnitine | 0.212 | 0.000 | 0.023 |
| HMDB11531  MG(0:0/14:1(9Z)/0:0 | -0.683 | 0.000 | 0.026 |
| HMDB02759  Androsterone sulfate | -0.327 | 0.000 | 0.026 |
| HMDB00622  Ethylmalonic acid | -0.795 | 0.001 | 0.040 |
| HMDB11152  **LysoPE(P-16:0/0:0)** | -1.089 | 0.001 | 0.042 |
| HMDB12088  Sphingomyelin (d18:0/18:1(11Z)) | -0.332 | 0.001 | 0.054 |
| HMDB00387  **3-Hydroxydodecanoic acid** | 0.198 | 0.003 | 0.103 |
| HMDB13325 2-trans,4-cis-decadienoylcarnitine | 0.604 | 0.003 | 0.106 |
| HMDB11474  **LysoPE(0:0/16:1(9Z))** | -0.485 | 0.006 | 0.158 |
| HMDB10395  LysoPC(20:4(5Z,8Z,11Z,14Z)) | -0.395 | 0.007 | 0.158 |
| HMDB11484 LysoPE(0:0/20:3(11Z,14Z,17Z)) | -0.591 | 0.007 | 0.158 |
| HMDB00152  **Gentisic acid** | 0.409 | 0.007 | 0.158 |
| HMDB12881  **Acetylcarnosine** | 0.129 | 0.007 | 0.158 |
| HMDB61112  3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid (CMPF) | -0.313 | 0.008 | 0.162 |
| HMDB00529 5Z-Dodecenoic acid | -0.720 | 0.009 | 0.162 |
| HMDB02183 Docosahexaenoic acid (65) | -0.115 | 0.017 | 0.281 |
| HMDB11473 LysoPE(0:0/16:0) | -1.100 | 0.017 | 0.281 |
| HMDB01008  **Biliverdin (isoform 2)** | 0.165 | 0.017 | 0.281 |
| HMDB01008  **Biliverdin (isoform 1)** | -0.161 | 0.021 | 0.319 |
| HMDB11490  **LysoPE(0:0/22:0)** | -0.553 | 0.022 | 0.326 |
| HMDB01015  N-Formyl-L-methionine | 0.290 | 0.025 | 0.326 |
| HMDB0000895  **Acetylcholine** | 0.146 | 0.026 | 0.326 |
| HMDB00269  **Sphinganine** | -0.155 | 0.027 | 0.326 |
| HMDB13124  Propenoylcarnitine | -0.360 | 0.027 | 0.326 |
| HMDB13034  Palmitoylglycine | -0.199 | 0.027 | 0.326 |
| HMDB13302  Phenylalanylphenylalanine (Phe-Phe) | 0.262 | 0.030 | 0.326 |
| HMDB00421  2,3-Dihydroxyvaleric acid | 0.145 | 0.030 | 0.326 |
| HMDB11487 LysoPE(0:0/20:4(5Z,8Z,11Z,14Z)) | -0.269 | 0.034 | 0.326 |
| HMDB12342  PS(14:1(9Z)/14:1(9Z)) | 0.104 | 0.034 | 0.326 |
| HMDB00267  Pyroglutamic acid | 0.157 | 0.035 | 0.326 |
| HMDB00148  **L-Glutamic acid (L-Glutamate)** | 0.288 | 0.035 | 0.326 |
| HMDB06695  Prolylhydroxyproline | 0.172 | 0.036 | 0.326 |
| HMDB10390  LysoPC(20:0/0:0) | 0.867 | 0.036 | 0.326 |
| HMDB37844  N-(1-Deoxy-1-fructosyl)valine | 0.163 | 0.037 | 0.326 |
| HMDB00663  **Glucaric acid** | 0.556 | 0.037 | 0.326 |
| HMDB00651 Decanoylcarnitine | 0.262 | 0.044 | 0.376 |
| HMDB01931  Gamma-CEHC | -0.425 | 0.049 | 0.397 |
| HMDB29377 Piperine | 0.186 | 0.049 | 0.397 |

For all metabolites, the associated common name, HMDB identifier, logFC, p-value and adjusted p-value are given.

### *Appendix G*:

Table A5: List of proteins with significantly different abundances (p < 0.05) between the insulin resistant and insulin sensitive subjects.

|  |  |  |  |
| --- | --- | --- | --- |
| Protein | logFC | p-value | Adjusted p-value |
| PLTP | -0.458 | 0.000 | 0.105 |
| APOC4 | 0.624 | 0.001 | 0.105 |
| APOD | -0.292 | 0.002 | 0.143 |
| PRG4 | 0.470 | 0.002 | 0.143 |
| CFH | 0.214 | 0.003 | 0.169 |
| SERPIND1 | 0.198 | 0.004 | 0.221 |
| HP | 0.389 | 0.009 | 0.298 |
| BCHE | 0.283 | 0.009 | 0.298 |
| F10 | 0.271 | 0.011 | 0.298 |
| PRG4 | 0.475 | 0.012 | 0.298 |
| APOA1 | -0.301 | 0.012 | 0.298 |
| APOF | -0.269 | 0.013 | 0.298 |
| LRG1 | -0.245 | 0.014 | 0.298 |
| VTN | 0.190 | 0.014 | 0.298 |
| SHBG | -0.522 | 0.021 | 0.427 |
| NPHP3 | 0.347 | 0.027 | 0.501 |
| IL1RAP | -0.302 | 0.033 | 0.560 |
| BTD | 0.186 | 0.033 | 0.560 |
| C1QB | 0.164 | 0.038 | 0.601 |
| PCOLCE | 0.170 | 0.041 | 0.608 |
| C4A | -0.402 | 0.042 | 0.608 |
| LPA | -1.074 | 0.049 | 0.609 |
| GSN | -0.166 | 0.050 | 0.609 |

For all proteins, the associated HGNC identifier, logFC, p-value and adjusted p-value are given.

## Add-on page: Description of work by student Bachelor BMS

**This add-on page to the Bachelor thesis BMS provides details on the role of the student in the experiments, data collection, and analyses described in the thesis.**

**Experiments and measurements**

/

[Please describe here which wet-lab experiments, clinical measurements, or other experiments and/or measurements were conducted by the student]

**Provenance of data**

The data used for analysis were obtained from the integrative Human Microbiome Project’s study on type 2 diabetes.

[Please describe here, for all data(sets) described in the results, whether it was generated by the student, generated by others during the project (e.g. project members, staff, other students), previously generated or existing in the lab, or taken from public resources]

**Data analysis and interpretation**

All analyses of the data were performed by Sabrina. For the metagenomic data, this included a PCoA, PERMANOVA, multivariate analysis of the homogeneity of group dispersions and identification of changes in mean phyla abundances. An adjusted R script of a Bioconductor tutorial (<https://www.bioconductor.org/help/course-materials/2017/BioC2017/Day1/Workshops/Microbiome/MicrobiomeWorkflowII.html#graph-based_analyses>) from another intern of Susan Coort, Anna Ritzen, was ­followed to complete these analyses. Moreover, interpretation of the results was discussed with a PhD student of John Penders, David Barnett, in order to better understand the analyses.

For the metabolomic and proteomic data, Sabrina performed the differential analyses for both datasets and used the output for pathway analysis as well as network analysis. Finally, integrative analysis of the metagenomic and metabolomic data using sparse CCA was also done by Sabrina.

[Please describe here which data analyses and interpretations were conducted by the student for the data(sets) described in the previous box]

**Integration with other analytical results**

/

[Please describe here which other analytical results were integrated with those of the analysis performed by the student, if any]

**Remarks**

*/*

[Room for any further remarks]