## 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 increase in cases from 2019 to 2045 (International Diabetes Federation. *IDF Diabetes Atlas, 9th edn.* Brussels, Belgium: 2019. Available at: [https://www.diabetesatlas.org](https://www.diabetesatlas.org/)). Type 2 diabetes will likely comprise 90-95% of these cases (American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2019. Diabetes Care. 2019 Jan 1;42(Supplement 1):S13-28), 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 characterized by insulin resistance (IR).

IR is 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

(<https://www-jci-org.ezproxy.ub.unimaas.nl/articles/view/77812>). Ectopic fat storage refers to the storage of fat in non-adipose tissues and can induce IR in the corresponding tissue via interference with insulin signalling (<https://www.hindawi.com/journals/ije/2012/983814/>). In this way, IR can develop in range of tissues separately (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6131567/>) but is mainly discussed in regards to skeletal muscle, the liver and adipose tissue due to their critical involvement in metabolism (<https://www-ncbi-nlm-nih-gov.ezproxy.ub.unimaas.nl/books/NBK507839/>). Development of IR in these tissues results in decreased glucose uptake and increased glycogenolysis/gluconeogenesis (<https://www-sciencedirect-com.ezproxy.ub.unimaas.nl/science/article/pii/S0025712510002051#sec1>), 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 (<https://www.emjreviews.com/diabetes/article/editors-pick-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/>). In this way, although the exact degree to which IR plays a role in T2D development is still not yet fully understood, there is no doubt that IR is an important accelerating factor. Hence, great efforts are being made in order to investigate factors influencing 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 (<https://www.thelancet.com/pdfs/journals/ebiom/PIIS2352-3964(19)30800-X.pdf>, <https://nyaspubs-onlinelibrary-wiley-com.ezproxy.ub.unimaas.nl/doi/full/10.1111/nyas.14107>).

The associations can most likely be explained by the ability of the microbiome to influence the host metabolome and proteome (Gut microbiome-host interactions in health and disease James M Kinross, <https://www.nature.com/articles/s41467-018-05357-4>). 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 hypertriglyceridemia (<https://www-sciencedirect-com.ezproxy.ub.unimaas.nl/science/article/pii/S0002914999002118>). This dyslipidaemia is characterized by increased plasma free fatty acids (FFAs) and very low density lipoprotein (VLDLs) resulting from decreased insulin-dependent suppression of lipolysis in IR adipose tissue (<https://www.jci.org/articles/view/10762>). Other non-lipid metabolites, such as branched chain amino acids (BCAAs), have also been shown to define the IR metabolome with studies showing increased plasma levels of these compounds in IR individuals (Metabolomics and Type 2 Diabetes: Translating Basic Research into Clinical Application Matthias S. Klein, Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity Oded Shaham). These changes are likely to be the result of an alteration in the expression of genes involved in white adipose tissue BCAA catabolism and, therefore, provide supporting evidence for not only a metabolomic but also a proteomic signature of IR (The proteomic signature of insulin-resistant human skeletal muscle reveals increased glycolytic and decreased mitochondrial enzymes J. Giebelstein).

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) pre-diabetic 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 insulin sensitive and insulin resistant 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’ (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5109542/>). 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 individual microbes contribute most to this separation. Finally, microbes and metabolites likely

Materials and Methods

### *Patient data*

#### Data acquisition

Proteomic, metabolomic, metagenomic and the corresponding subject data of the integrative Human Microbiome Project’s (iHMP) T2DM (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5109542/>, <https://www.nature.com/articles/s41586-019-1236-x#MOESM3>) was used and is available from <http://hmp2-data.stanford.edu/>. The complete study design of the iHMP’s T2DM project is shown in figure 1. Subject data consisted of steady-state plasma glucose (SSPG, mg/dl) measurements, race, age, gender, classification as either IR or IS and BMI. Subjects were classified as either IR or IS based on their SSPG measurement: a SSPG < 150 mg/dl was considered as insulin-sensitive and a SSPG ≥ 150 mg/dl, insulin resistant.

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 (<https://www.nature.com/articles/s41467-017-00249-5>) and LC-MC/MS (Want EJ. LC-MS Untargeted Analysis. InMetabolic Profiling 2018 (pp. 99-116). Humana Press, New York, NY.), respectively. Microbial taxa from stool samples were identified using 16s sequencing (<https://www.pnas.org/content/82/20/6955>).



*Figure 1: the sample population and general sampling protocol of the iHMP’s T2DM project. IR = insulin resistant. IS = insulin sensitive. SSPG =* *steady-state plasma glucose. PBMC = peripheral blood monocytes.*

#### Pre-processing of subject data

Only subjects whose insulin sensitivity status had been recorded were included in present study. In this way, the original sample population was filtered down to only those classified as either insulin resistant (IR) or insulin sensitive (IS). 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 HMP 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*

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

A pre-existing phyloseq-class object included in the HMP2Data Bioconductor R package (Stansfield J, Dozmorov M (2019). HMP2Data: 16s rRNA sequencing data from the Human Microbiome Project 2. R package version 1.1.0, <https://github.com/jstansfield0/HMP2Data>.) 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 (<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217>) which contains many tools to analyse and visualise complex microbial data.

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 metagenomic 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 2). A large spread of samples was seen to be separated from the main cluster of samples, with the majority possessing an Axis1 value of less than -2.8. Hence, only samples with an Axis1 value of more than -2.8 were included in the metagenomic data.

Finally, the taxa of the phyloseq object were filtered. A prevalence threshold was 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 in order to prevent filtering of taxa that 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. Finally, 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.

#### 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 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 (Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin,R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2019). vegan:Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan).

From this, an investigation into the phylum abundances per sample in each group was executed using barplots and allowed for elaboration of the variance in phylum abundance of the 2 groups. Finally, in order to examine which taxa were more and less abundant in each group, box plots for each phylum illustrating the mean abundances of each group 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*

#### 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 ([https://pubmed.ncbi.nlm.nih.gov/29140435/?from\_term=HMDB+%5Bti%5D&from\_sort=date&from\_pos=1](https://www.google.com/url?q=https://pubmed.ncbi.nlm.nih.gov/29140435/?from_term%3DHMDB%2B%255Bti%255D%26from_sort%3Ddate%26from_pos%3D1&sa=D&ust=1591951941889000&usg=AFQjCNHee4PsR1KPRjL6tdxNhRF-i1s6ig)). 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 to identify metabolites whose abundances were significantly different between the IR and IS condition via application of 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* package (<https://academic.oup.com/bioinformatics/article/34/19/3417/4987147>) in R-3.6.3.

#### Differential analysis of the metabolomic data

A DA of the metabolomic data was executed using the MetaboDiff and allowed for identification of metabolites likely to differentially abundant between the 2 conditions.

### *Proteomic data*

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

Like 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 (<https://academic.oup.com/nar/article/47/D1/D786/5124600>).

#### Differential analysis of the proteomic data

Similarly, to the analysis of the metabolomic data, a DA of the proteomic data in R-3.6.3 was done to achieve the same end goal but in relation to proteins. However, instead of MetaboDiff, the limma package (Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K.(2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.) was used.

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

#### 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 (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004085>) and MetaboAnalyst v.3.0 (<https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.86>). This allowed for the most likely altered biological pathways in the IR subjects compared to the IS subjects to be determined.

##### Data pre-processing

Since the compounds in both DA output files were not annotated with identifiers 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 corresponding HMDB identifiers using the metabolite annotation datafile available from the iHMP website. Furthermore, some of the HMDB identifiers in the metabolite annotation datafile were not able to be recognized by PathVisio. These identifiers were then changed to HMDB identifiers that could be recognized by PathVisio. (Appendix 3)

As for the proteome DA results, most identifiers corresponded to HGNC annotations, however, some did not correspond directly. As there was no datafile related to the protein identifiers on the iHMP website, a BioMart search of the non-HGNC identifiers was done to replace then with an equivalent HGNC identifier (Appendix 4).

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

##### Pathway analysis

Pathway analysis using PathVisio was done using the combined DA results. An expression criterion of p-value < 0.05 was specified and pathways were sourced from the WikiPathways database of curated human pathways (<https://academic.oup.com/nar/article/46/D1/D661/4612963>). Statistical analysis involved an overrepresentation analysis of the metabolites and proteins in these pathways and allowed for identification of pathways most significantly altered in the IR condition compared to the IS condition.

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) (<https://pubmed.ncbi.nlm.nih.gov/31441146/>) was applied.

Performing combined pathway analyses using different programmes allowed for a more detailed understanding of the pathways altered based on differences protein and metabolite abundances. This is especially true since MetaboAnalyst is more adept concerning metabolites but focuses more on enzymes, compared to PathVisio.

#### 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 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403769/#:~:text=Cytoscape%20is%20an%20open%20source,into%20a%20unified%20conceptual%20framework.>) was performed. This was done by first running the significantly altered proteins (p < 0.05) generated by the DA through stringApp (<https://pubs.acs.org/doi/abs/10.1021/acs.jproteome.8b00702?casa_token=USD2zPrA-5QAAAAA%3AYyafljBu-Nc16fPtSZTe6iUPJs-9Osr3IunM7yHnmRqfAMPsnYE-OCnVtE63ZJN9MbDSHO4XSTxDSww-&>) (confidence score = 0.4) to create a protein-protein interaction network. The resulting network was then extended with protein-pathway interactions using the WikiPathways linkset in the Cytoscape CyTargetLinker app (<https://f1000research.com/articles/7-743>).

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*

#### 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 2) with only those corresponding to an Axis1 value > -2.8 being included, as explained before. 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.

#### 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 recognition of the corresponding features linking the 2 datasets. The PMA (penalyzed multivariate analysis) package (Daniela Witten and Rob Tibshirani (2020). PMA: Penalized Multivariate Analysis. R package version 1.2.1. https://CRAN.R-project.org/package=PMA) in R was used to execute this analysis and a penalty of 0.15 was applied to both the metagenome and metabolome matrices.

The R scripts used to run the metagenomic analyses, the DAs and the integrative analysis of the metabolomic and metagenomic data can all be downloaded from … (github link).

## Results (1343 words without figures/tables + their descriptions)

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 specified.

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 of less than -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) based on Bray-Curtis dissimilarity illustrating the separation in the microbial composition of the insulin resistant (IR) and insulin sensitive (IS) samples. 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 distinction 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 the IR samples compared to the IS samples (Figure 4.B) while the Verrucomicrobia were more abundant in the IR samples (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 microbes contributing the most to the metagenomic differences between the 2 groups. The majority (70%) of these microbes 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 taxa generated by the multivariate ANOVA with permutations (PERMANOVA) deemed to contribute the most to the separation between the insulin resistant and insulin sensitive groups. Taxa with negative values are less abundant in the IR condition and taxa with positive values are more abundant in the IS condition. Blue bars represent taxa belonging to the Bacteroidetes phylum, orange bars represent taxa belonging to the Firmicutes phylum and green bars represent taxa 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 (p < 0.05) in the IR group compared to the IS group: 21 were less abundant (logFC < 0) and 19 more abundant in the IR condition (logFC > 0) (Appendix 5). Based on a logFC > 0.05 or < -0.05, 9 out of the 21 less abundant metabolites and 3 out of the 19 more abundant metabolites were sufficiently changed (Table 1). 11 of these 12 sufficiently changed metabolites were lipids/lipid-like molecules with the only exception being glutaric acid. 7 metabolites possessed a significant adjusted p-value with 5 also having a logFC suggesting a sufficient change in abundance (Table 1). The remaining 2 metabolites with a logFC > 0.05 or < -0.05 but a non-significant adjusted p-value were androsterone sulphate (logFC = -0.327) and hexanoylcarnitine (logFC = 0.212).

*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 |
| Glutaric 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 p-value <0.05. 13 of these proteins were more abundant in the IR subjects (logFC > 0) and 10, less abundant (logFC < 0) (Appendix 6). 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 compounds 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: Significantly altered pathways (1.96 < Z-score < -1.96) and the associated significantly altered compounds (p < 0.05) from the combined pathway analysis using PathVisio.*

|  |  |  |
| --- | --- | --- |
| Pathway | Z-score | Proteins and Metabolites |
| Gastric acid production | 3.05 | ↑ Acetylcholine |
| Heroin metabolism | 3.05 | ↑ Pseudo-ChE |
| IL-1 signaling pathway | 3.05 | ↓ IL1RAP (CHECK) |
| 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 signaling 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 |

*The common name for each metabolite is given and proteins are annotated using HGNC symbols. 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), hence, meaning 39 metabolites and 23 proteins were used in the corresponding joint pathway analysis. 6 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 significantly altered pathways ( p < 0.05) and the associated significantly altered compounds (p < 0.05) from the combined pathway analysis using MetaboAnalyst.*

|  |  |  |
| --- | --- | --- |
| 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 |  |
|  |

*The common name for each metabolite is given. Arrows indicate compounds more abundant (*↑) *and less abundant (*↓) *in the IR condition.*

Since overrepresentation analysis can bury information concerning pathways that it deems to not be significantly altered, an extended network analysis of the significantly altered proteins was performed to uncover alterations in biological pathways that may have been lost in this analysis. The results revealed the human complement system and the complement and coagulation cascades 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 cascades 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 wich were less abundant in the IR condition (logFC < 0) (Appendix 6).

### *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. 8 of these metabolites were monoacylglycerols with the remaining metabolites being arachidonic acid and 2-arachidonoylglycerol (2-AG). One of these monoacylglycerols (MG(0:0/14:1(9Z)/0:0)) as well as 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.

## Appendix

### *Appendix 1:* Figure A1

Study design.



### *Appendix 2:* 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 3*: Table A1

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 4*: Table A2

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 5:* 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 (DHA) | -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 6*: Table A3

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.