## 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). However, insulin resistance 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 FFAs and 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 still remain unclear. For this reason, this study aims to investigate the differences in the microbiome, host proteome and host metabolome between insulin resistant and insulin sensitive (IS) pre-diabetics. 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.

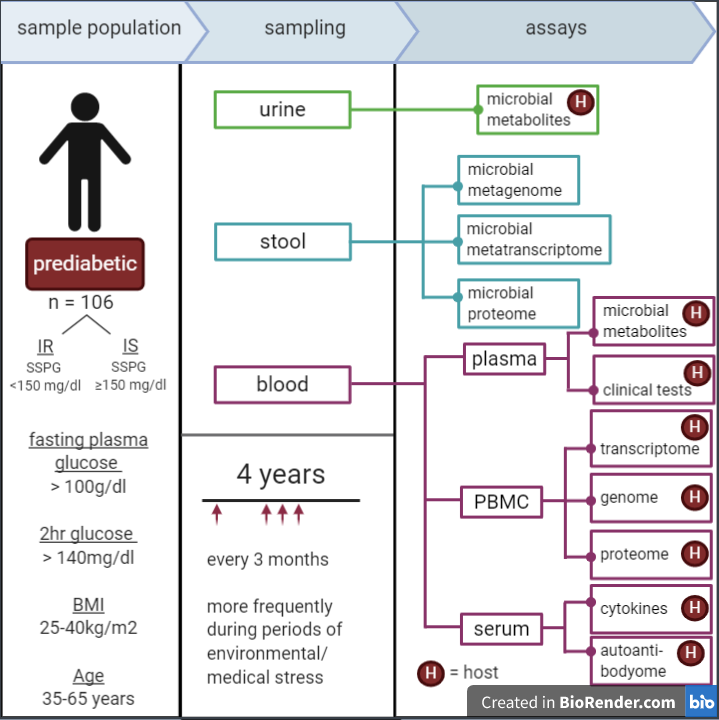
In order to test this hypothesis, data from the integrative Human Microbiome Project’s (iHMP) study concerning T2D was used. This study aims to research 106 individuals at high risk for diabetes over a period of 4 years in order to determine the ‘physiological changes that occur in the microbiome and 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 2 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. Samples were taken every 3 months but this frequency was increased during periods of environmental/medical stress. At each visit blood, urine and fecal 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.nature.com/articles/s41467-019-13036-1>).



*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>.) EXPLANATION OF PHYLOSEQ

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 1). 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 (<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217>) 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 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.

All analyses of the metagenomic data were executed using R-3.6.3. The corresponding script can be downloaded from … (github link).

### *Metabolomic data*

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

The metabolomic data was filtered to only include metabolites annotated with an HMDB (human metabolite database) identifier ([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 required for follow-up analysis. 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.

#### 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 changes*

#### 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 from one well-known database, 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. In some instances, 1 metabolite corresponded to 2 HMDB identifiers. In this case, only 1 of the 2 HMDB identifiers was used. Furthermore, some of the HMDB identifiers in the aforementioned metabolite annotation datafile were not able to be recognized by PathVisio. These identifiers were then changed to HMDB identifiers that could be recognized by PathViso. (Appendix 2)

As for the proteome DA results, most identifiers corresponded to HGNC (<https://academic.oup.com/nar/article/47/D1/D786/5124600>) annotations, however, some did not. As there was no datafile related to the protein identifiers on the iHMP website, a Google search of the non-HGNC identifiers was done to replace then with an equivalent HGNC identifier (Appendix 3).

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

##### Analysis

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 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 2 combined pathway analyses using different programmes allowed for a more detailed understanding of the pathways changed. 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 the biological pathways involving the significantly altered proteins that may have been lost by the overrepresentation analysis, these proteins were run through 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.>). This was done by first creating a protein-protein interaction network of the significantly altered proteins generated by the DA using stringApp (<https://pubs.acs.org/doi/abs/10.1021/acs.jproteome.8b00702?casa_token=USD2zPrA-5QAAAAA%3AYyafljBu-Nc16fPtSZTe6iUPJs-9Osr3IunM7yHnmRqfAMPsnYE-OCnVtE63ZJN9MbDSHO4XSTxDSww-&>) (confidence score = 0.4). The resulting network was then extended to include the corresponding altered biological pathways by applying the WikiPathways link set using CyTargetLinker (<https://f1000research.com/articles/7-743>).

A network analysis of this extended network was then executed 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 aforementioned PCA 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 allow for a more sound 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 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).