## 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 (1). Since type 2 diabetes (T2D) is responsible for 90-95% of all diabetes (2), the vast majority of these cases will be type 2. In this way, T2D is considered a much more pressing public-health concern than type 1. Despite this high prevalence in the population, the biological mechanism underlying its development and the physiological changes that occur as a result are not entirely understood. Nevertheless, it is generally accepted that T2D is characterized by a widespread insufficient insulin response known as insulin resistance. The onset of insulin resistance has been proposed to be the instigator of pancreatic beta cell failure and, hence, decreased insulin secretion by placing beta cells under increased pressure to produce insulin (3). This is especially relevant considering that a portion of type 2 diabetics possess alleles associated with compromised beta-cell function (<https://link-springer-com.ezproxy.ub.unimaas.nl/article/10.1007/s00125-004-1338-2>) and that new beta cells are unable to be produced past the age of 30 in the human pancreas (<https://pubmed.ncbi.nlm.nih.gov/20660050/>). Despite the exact degree to which insulin resistance plays a role in the development of T2D not being entirely understood, it is certainly an important accelerating factor in T2D development. Hence, great efforts are being made in order to investigate factors influencing insulin sensitivity. For instance, the 2nd phase of the National Institutes of Health’s 10 year long Human Microbiome Project includes a study concerning T2D. 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/>). However, changes in insulin resistance are not only important due to it being strongly associated with diabetes risk, but also due to its metabolic significance.

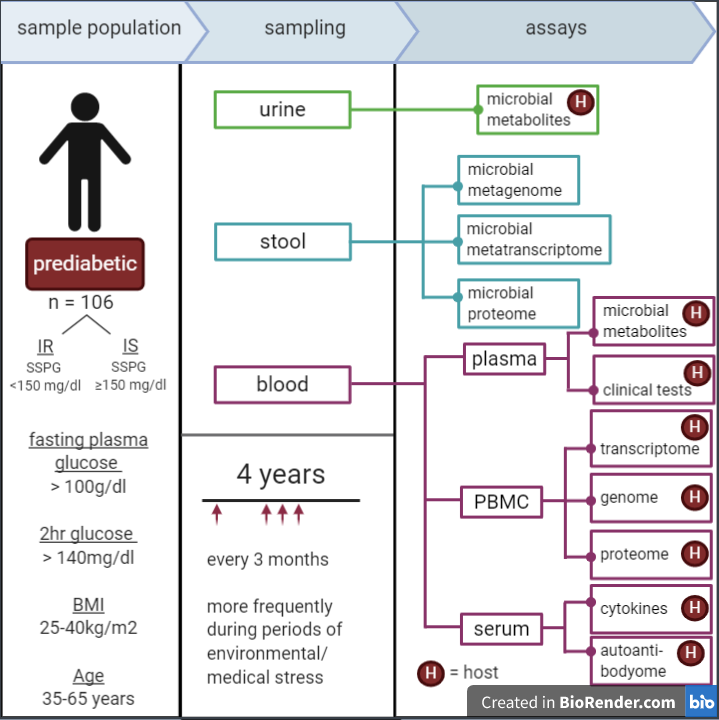
The inability of peripheral tissues to correctly respond to insulin results in a large array of metabolic consequences such as dyslipidemia and unrestrained gluconeogenesis (<https://www-sciencedirect-com.ezproxy.ub.unimaas.nl/science/article/pii/S0025712510002051#sec1>). In this way, the metabolome of IR individuals can be expected to be quite different from IS individuals. For example, hypertriglyceridemia is commonly associated with insulin resistance (<https://www-sciencedirect-com.ezproxy.ub.unimaas.nl/science/article/pii/S0002914999002118>). This is most likely due to the inability of insulin resistant adipose tissue to receive the insulin signals necessary to suppress lipolysis, hence, resulting in the release of more and uptake of less free fatty acids (FFAs) by adipocytes. The subsequent increase in plasma FFAs causes the liver to increase its production and secretion of VLDL, eventually resulting in hypertriglyceridemia (<https://www.jci.org/articles/view/10762>).

Other non-lipid metabolites have also been shown to characterize the insulin resistant metabolome. Due to insulin also playing a role in protein metabolism, certain key amino acids and associated intermediary metabolites have become a popular topic in diabetes research. This is especially true in respect to branched chain amino acids (BCAAs) whose plasma levels are highly sensitive to insulin action (<https://link.springer.com/article/10.1007/s00726-011-1088-7>) and which have been shown to be increased insulin-resistant pre-diabetics (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). One way that insulin signaling is able to influence BCAA level is via the expression of genes involved in white adipose tissue BCAA catabolism (Mechanisms of human insulin resistance and thiazolidinedione-mediated insulin sensitization D. D. Sears). This association between insulin resistance and gene expression has also been observed for other genes: for instance, genes involved in PPAR-gamma’s activity have been shown to be downregulated in the skeletal muscle of pre-diabetics (Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of *PGC1* and *NRF1* Mary Elizabeth Patti (<https://www-pnas-org.ezproxy.ub.unimaas.nl/content/100/14/8466.long#ref-12>)). Extending upon these discoveries, the proteome of insulin resistant individuals has too been demonstrated to be distinct from insulin sensitive individuals (The proteomic signature of insulin-resistant human skeletal muscle reveals increased glycolytic and decreased mitochondrial enzymes J. Giebelstein). However, the proteome and metabolome are not the only signatures of insulin resistance. Recent advancements in high throughput sequencing technologies have allowed for detailed studies on the microbiome to become more realisable. The field of T2D research is no exception to this scientific revolution, with different microbes being correlated with T2D (<https://www.thelancet.com/pdfs/journals/ebiom/PIIS2352-3964(19)30800-X.pdf>). These findings are pivotal in the understanding of T2D due to the influence of the microbiome on the host metabolome and proteome (Gut microbiome-host interactions in health and disease James M Kinross). Nevertheless, this interaction is still poorly understood and the changes in the omics of type 2 diabetic individuals remain debatable, especially in the time-period prior to diabetes onset. For this reason, this study aims to investigate the differences in the microbiome, host proteome and host metabolome between insulin resistant and insulin sensitive pre diabetics. The majority of analyses quantifying these differences will be done using R. Additional analyses to understand the biological implications will use pathway analysis, network analysis and taxon enrichment set analysis. From the aforementioned research identifying alterations in the omics resulting from insulin resistance, it is hypothesised 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.

Materials and Methods

### *Patient data*

Proteomic, metabolomic, metagenomic and the corresponding subject data of the integrative Human Microbiome Project’s (iHMP) T2DM was used and is available from <http://hmp2-data.stanford.edu/>. Subject data consisted of steady-state plasma glucose (SSPG) 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 deemed 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 by 16s sequencing and urine samples were solely used to investigate the host microbial metabolite abundances. (Figure 1). Further information can be found in the publications concerning this research project (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5109542/>, <https://www.nature.com/articles/s41586-019-1236-x#MOESM3>).



*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.*

### *Data pre-processing*

#### Subject data

Only subjects whose insulin sensitivity status had been recorded were included in this 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 3 omics datasets downloaded from the HMP website.

#### Proteomic and metabolomic data

Samples within the proteomic and metabolomic data were filtered using the subject data to only include samples from classified subjects. These samples from each subject were then averaged together per compound in order to obtain one abundance value for each subject-compound combination.

The metabolomic data was further filtered to only contain metabolites for whom an HMDB identifier was specified since these identifiers are needed in the follow-up analysis. The resulting dataset was normalized by variance stabilization using the MetaboDiff package (<https://academic.oup.com/bioinformatics/article/34/19/3417/4987147>)in R. This allowed for the identification of metabolites whose abundances were significantly different between the IR and IS condition.

#### 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. The sample data of this object was exported from R and compared to the subject data. This allowed for the creation of a vector of sample IDs which matched those in the phyloseq object and which belonged to classified subjects. This vector of sample IDs was then used to include only classified individuals in the metagenomic data.

#### Metagenomic and metabolomic data

In order to perform an integrative analysis of the metabolomic and metagenomic data, the metagenomic phyloseq object and metabolomic data set were made consistent with each other. This was done by comparing the sample IDs in the metagenomic and metabolomic datafiles.

Before exclusion of any samples, the shorter sample IDs in the metabolomic datafile were adjusted to match the longer sample IDs found in the metagenomic data. Most metabolomic sample IDs had 1 corresponding metagenomic sample ID, in which case the longer metagenomic sample ID was substituted in place for the shorter sample ID in the metabolomic datafile. However, 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.

The resulting metabolomic data was then used for the integrative analysis and to subset the samples of the metagenomic phyloseq object in R. This ensured that the 2 datasets contained the same samples with matching IDs.

Additional samples were excluded based on a depiction of an integrative metagenomic and metabolomic analysis using a Principal Component Analysis (PCA) plot (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 metabolomic and metagenomic data.

Further filtering concerned the taxa of the phyloseq object. Any taxa for whom family and genus were not specified were excluded. Furthermore, only taxa with an abundance sum of more than 4 across all samples and which were present in at least 3 samples possessing an abundance of more than 2, were included. Finally, a prevalence filter of 10% was applied to the 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. This final phyloseq was then processed so that any abundance value greater than 50 was reduced to a value of 50.

The metabolomic data used for the integrative analysis of the metagenomic and metabolomic data was also further preprocessed by only including metabolites that had an abundance of 0 in no more than 3 samples as well as by log(x+1) transforming the abundances to weaken any heavy tails.

### *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 log2 transformed mean abundances in each group were created. The resulting difference in mean abundance of each phylum between the 2 groups was investigated using a Wilcoxon test. 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).

### *Statistical analysis of the proteomic and metabolomic data*

Other than the microbiome, protein and metabolite abundances have been seen to differ depending on insulin sensitivity. For this reason, changes in the proteome and metabolome between the IR and IS subjects as well as the influence of these changes at pathway level were investigated.

#### Differential analysis

A differential analysis (DA) was done for both omic data sets but a different R package was used for the proteomic and metabolomic data. This analysis allowed for the most likely altered proteins and metabolites between the IR and IS condition to be determined.

Concerning the packages used, the proteomic data was analysed using 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.) while the metabolomic data was analyzed using the MetaboDiff package (Andreas Mock (2020). MetaboDiff: An R package for differential metabolomic analysis. R package version 0.9.3.).

Analysis used R-3.6.3. The corresponding R scripts can be downloaded from …

#### Combined pathway analysis

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 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 elaborate upon as well as visualize the involvement of the significantly altered proteins in biological pathways, 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 network of protein-protein interactions using stringApp (<https://pubs.acs.org/doi/abs/10.1021/acs.jproteome.8b00702?casa_token=USD2zPrA-5QAAAAA%3AYyafljBu-Nc16fPtSZTe6iUPJs-9Osr3IunM7yHnmRqfAMPsnYE-OCnVtE63ZJN9MbDSHO4XSTxDSww-&>) with a confidence score of 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.

### *Extension of the metagenomic analysis*

The microbiome has been shown to influence the host metabolome through its involvement in digestion as well as through its involvement in gut hormone release. Furthermore, since metabolic disruption is associated with many diseases, the microbiome has also been proposed to be implicated in host health. For this reason, the metabolites and microbes which together best explain the covariation between the metabolomic and metagenomic data, were identified. The taxa contributing the most to the separation of the IR and IS group were then used to establish what physiological characteristics were associated with the microbiome of the IR group.

#### Metabolomic and metagenomic integrative analysis

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 corresponding R script can be downloaded from … (github link).

#### Taxon set enrichment analysis

A Taxon Set Enrichment analysis (TSEA) was applied to both the most and least abundant taxa in the IR condition generated by the PERMANOVA. This was done using MicrobiomeAnalyst (<https://www.nature.com/articles/s41596-019-0264-1>) and allowed for associated host intrinsic factors for each group of taxa to be determined.

## Appendix

### *Appendix 1:* Figure A1

Principal Component Analysis (PCA) plot of the metagenomic data with the features most likely to explain the covariation between the metabolomic and metagenomic abundances as input. The explained variance of the metagenomic principal components are given as a percentage by the corresponding axis. OTU = taxa. IR = insulin resistant. IS = insulin sensitive.

### 

### *Appendix 2*: 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 3*: 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 4*

## 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. 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.