Article

Deep Learning for Integrated Analysis of Insulin Resistance with Multi-omics Data

Eunchong Huang1, Sarah Kim2 and Taejin Ahn2\*

1Department of Advanced Green Energy and Environment, Handong Global University,   
Pohang-si, Gyeongbuk 37554, South Korea;

2Department of Life Science, Handong Global University, Pohang-si, Gyeonbuk 37554, South Korea;

hec1324@gmail.com (E.H.); [22032001@handong.edu](mailto:22032001@handong.edu) (S.K)

**\*** Correspondence: taejin.ahn@handong.edu; Tel: +82-54-260-1360

Received:

**Abstract:** Technological advances in next-generation sequencing (NGS) have made it possible to uncover extensive and dynamic alterations in diverse molecular components and biological pathways across healthy and diseased conditions. Large amounts of multi-omics data originating from emerging NGS experiments require feature engineering, which is a crucial step in the process of predictive modeling. The underlying relationship among multi-omics features in terms of insulin resistance is not well understood. In this study, using the multi-omics data of type II diabetes from the Integrative Human Microbiome Project, from 10,783 features, we conducted a data analytic approach to elucidate the relationship between insulin resistance and multi-omics features, including microbiome data. To better explain the impact of microbiome features on insulin classification, we used a developed deep neural network interpretation algorithm for each microbiome feature’s contribution to the discriminative model output in the samples.

**Keywords:** Feature reduction, Microbiome, Multi-omics, Prediction model, Feature engineering

1. Introduction

Advances in high-throughput DNA sequencing platforms have become essential in the field of gene expression profiling, epigenomics, genomics, and transcriptomics over the past ten years[1-3]. The technical developments and decreasing cost of sequencing platforms have made a dramatic contribution to large-scale projects. In particular, human microbiome studies have been accelerated by the advent of next-generation sequencing (NGS) and aim to unravel the association of microbial abundance with health or disease outcomes.

To address the link of humans and their microbiomes to health-related outcomes, the NIH Human Microbiome Project (HMP) and the second phase, the Integrated Human Microbiome Project (iHMP) observed the dynamic alterations in hosts and their microbiomes under particular conditions[4]. The HMP was one of the first large-scale initiative projects to address the linked interactions between hosts and their microbiomes. The first phase project sought to determine whether there were common healthy microbiomes in the absence of overt disease. The ten year NIH HMP project characterized the microbial communities from numerous body sites and correlated them with phenotypes to determine healthy and disease variations. However, one of the main findings of the HMP was that the microbial composition alone was not correlated with the host phenotype[5-8]. This finding led to the development of iHMP project, which was designed to gain a more holistic view of host-microbiome interactions over time.

The iHMP project expanded the repertoire of biological properties by providing not only microbial dynamics but also multi-omic analyses, including immunity, metabolism, and dynamic molecular activity, to address the relationship between host and microbiome mechanistically. Within the iHMP project, three sub-projects, comprising pregnancy and preterm birth, inflammatory bowel diseases, and stressors that affect individuals with prediabetes, were included to underpin the mechanisms of human and microbial activity longitudinally. Previous iHMP diabetes-related projects have mainly focused on the time series analysis of insulin sensitivity and resistance in prediabetes patients. The study profiled several molecular patterns, which show that a few markers are sufficient for predicting stress events (i. e., respiratory viral infection versus healthy time points) [9]. Moreover, it also showed that an individual progression of type II diabetes mellitus (T2D) could be predicted before its actual onset via multi-omics analysis. However, the previous study did not address multi-omic factors in classifying insulin sensitivity (IS) and insulin resistance (IR), despite it being feasible.

Now, the emphasis has moved from data generation to effective analysis of data. Substantial challenges are presented, including sample quality control, pre-processing, normalization, and integration of datasets across platforms and techniques. It is worth noting that feature engineering truly reflects the intrinsic relation with the attribute to be predicted, which can significantly affect the performance of any resulting models[10-11]. For instance, typical laboratories facilitating the generation of high-dimensional multi-omic datasets can produce more than 100 gigabytes of information. These high-dimensional multi-omic features may contain noise and misleading features that are detrimental to model performance and may also increase redundant information[12-13]. At the same time, however, the question of obtaining accurate molecular signatures from the biological processes of these complex datasets is complicated. While classifying IR and IS with a small number of biomarkers is very challenging, we aim to do so by identifying biomarkers that make it possible to distinguish IR from IS.

Considering these converging challenges within the biomedical field, especially with respect to clinical translation, we evaluated whether disease-specific multi-omic variables are present in patients with IR, identified the microbiome-based diagnostic signatures to a classifier setting, and interpreted how selected features contributed to the model output.

2. Materials and Methods

2.1. iHMP Type 2 diabetes mellitus data description

Processed analytes, such as clinical measurements, cytokine profiles, gut microbial taxa, proteomics, and RNA sequencing (RNA-seq) were obtained directly from the iHMP Stanford School of Medicine (<http://med.stanford.edu/ipop.html>). The T2D of iHMP was designed to understand the physiological changes that occur in the microbiome and host during viral infection and during changes in glucose levels and insulin resistance. This project established a cohort of approximately 60 individuals at risk for diabetes. Under diabetes progression, iHMP T2D performed longitudinal multi-omic analysis to obtain global microbiome-host changes.

In this study, we conducted a cross-sectional study of all the samples disregarding the longitudinal profiling of patients. The study population was classified by the steady state plasma glucose level (SSPG). Subjects with an SSPG greater than 150 mg/dL are classified as IR and below 150 mg/dL are classified as IS. The data consisted of 205 patients diagnosed with IR and 223 patients with insulin sensitivity (IS). The total number of combined multi-omic analytes was 10,783 features. To develop a model capable of distinguishing IS from IR, subjects were divided randomly in an 8:2 model training dataset to holdout dataset ratio. The holdout dataset was used only to verify the model performance and the training set is randomly divided into 8:2 ratio to obtain training dataset and validation dataset. All the data are scaled to a fixed range by *MinMax* scaling. The overall workflow of this study is demonstrated in Supplementary Data S1.

2.2. Microbiome data pre-processing

Working with the microbiome regularly involves the issue of data containing many zero values. Therefore, we first removed bacterial taxa with a mean relative abundance <0.005%. After filtering, the abundance of each bacterial taxon was normalized with a variance stabilizing arcsine square root transformation[14]. The resulting microbiome data were used with other metadata to develop the prediction model.

2.3. Predictive models for insulin resistance and insulin sensitivity (IRIS)

The appropriateness of the reduced features was validated using ensemble classifiers and a deep neural network (DNN) classifier. In order to obtain optimal model performance, hyperparameter optimization of ensemble models was performed for the learning algorithm using Gridsearch (Supplementary Data S2). The code for DNN hyperparameter optimization was built internally to aggregate the best parameters. Random forest yielded an optimal with max\_depth = 5, min\_samples\_split = 4, and n\_estimators = 350. AdaBoost yielded an optimal with learning\_rate = 0.05, and n\_estimators = 250. Gradient boosting yielded an optimal with min\_samples\_split = 5, learning\_rate = 0.05, and n\_estimators = 350. Xgboost yielded an optimal with learning\_rate = 0.05, max\_depth = 5, and n\_estimators = 100. Finally, DNN yielded an optimal with learning\_rate = 0.0005, layers = [10783, 8626, 6469, 4313, 2156], batch\_size = 25, epoch = 100, with the adam algorithm to adjust the learning rate dynamically, sigmoid for calculating predictions, and the remaining layers are activated with the ReLU function. The comparisons of all developed predictive models were based on the area under the receiver operating characteristic (ROC) curve.

2.4. Feature reduction and selection

Feature selection is performed to reduce the dimensionality of the multi-omic data and is used for selecting important attributes for classification. The metadata were divided into five folds (index), and in each fold, feature reduction was performed. We selected features from the gradiant boosting classifier because it showed the highest AUC in the training set among the classifiers of different methods. Feature importance from the gradient boosting classifier was obtained, and features with no importance magnitudes were eliminated beforehand. Then, we sequentially erased features in the feature list and re-evaluated the performance drop. Each time, we removed one feature, re-trained the model, and evaluated the performance with the test data. The features of each index are eliminated until the AUC of a model drops less than 0.98, and the final features are determined by selecting all the intersecting features across all indices.

2.5. Predictive models for IRIS with selected features

A new predictive model was built and used to recalculate the model performance to examine the effect of feature reduction. Out of 10,783 features from five different sets of metadata, only 16 features were used to train a new classifier. Hyperparameter tuning of all the five classifiers was performed to optimize the model performance. Random forest yielded an optimal with max\_depth = 5, min\_samples\_split = 2, and n\_estimators = 150. AdaBoost yielded an optimal with learning\_rate = 0.05, and n\_estimators = 350. Gradient boosting yielded an optimal with min\_samples\_split = 2, learning\_rate = 0.01, n\_estimators = 400. Xgboost yielded an optimal with learning\_rate = 0.05, max\_depth = 5, and n\_estimators = 250. Lastly, DNN yielded an optimal with learning\_rate = 0.0005, layers = [16, 12, 9, 5, 2], batch\_size = 25, epoch = 500, with the adam algorithm to adjust the learning rate dynamically, sigmoid for calculating predictions, and the remaining layers are activated with the ReLU function.

2.6. Predictive models for IRIS with microbiome feature substitution

The correlation network was created using the software qgraph version 1.6.5. First, the Pearson correlation of the analytes was obtained using the cor function. Argument use = “pairwise.complete.obs” was used to delete pairwise missing data. Finally, the network was drawn using the qgraph package in combination with graph = “pcor” and the threshold argument was given to remove edges that were not significant. The false discovery rate (FDR) was used to compute significance without correction for multiple testing with a combination of alpha = 0.001 to plot the significant edges.

Pairwise correlation networks between microbiome features and 16 significant features were calculated, and the corresponding features were replaced with the microbiome features. Predictive models were rebuilt and the model performance was re-calculated to highlight the microbisome data.

Hyperparameter tuning of all the five classifiers was performed to optimize the model performance. Random forest yielded an optimal with max\_depth = 20, min\_samples\_split = 3, and n\_estimators = 200. AdaBoost yielded an optimal with learning\_rate = 0.05, and n\_estimators = 250. Gradient boosting yielded an optimal with min\_samples\_split = 2, learning\_rate = 0.01, n\_estimators = 400. Xgboost yielded an optimal with learning\_rate = 0.05, max\_depth = 10, and n\_estimators = 250. Lastly, DNN yielded an optimal with learning\_rate = 0.0005, layers = [17, 13, 9, 6, 2], batch\_size = 15, epoch = 500, with adam algorithm to adjust the learning rate dynamically, sigmoid for calculating predictions, and the remaining layers are activated with the ReLU function.

2.7. Random sample permutation

Using the selected 17 features, the DNN model learns randomly splitted train samples from the training set and generates the test AUC using the holdout dataset. Each number of a permutation sequence uses the randomly splitted train samples and this is repeated for 100 times. For all the given node values of DNN model in Supplementary Data S2, DNN builds a optimal model for every permutation with the optimal combinations of learning rate, batch size and epoch. The histogram for the holdout dataset AUC scores is drawn using matplotlib.pyplot.hist.

2.8. Deep neural network interpretation algorithm

One of the disadvantages of the DNN is the complexity in understanding the precise contribution of a particular feature to the result. One way to address this issue is to input a range of expression values of a feature for a given sample and observe the alterations in the DNN outcome [15]. Specifically, we substituted the value of a feature from its minimum to maximum value across all samples in the dataset and observed the changes in the DNN outcome while the other features were left unchanged. This process was repeated until all features were considered for each sample. The pseudo code for the deep neuranl network interpretation algorithm is listed in the Supplementary Data S9.

2.8. Statistical Analysis

Statistical analysis was performed using RStudio (version 3.6.1, http://www.R-project.org). Unless otherwise indicated, the significance tests for the differences between IR and IS within training set and holdout datasets are performed using an independent two-sample t-test. Graphical analysis was performed using the GraphPad Prism 8 program (GraphPad Software Inc., USA). A p-value <0.05 was considered significant.

3. Results

3.1. Baseline characteristics of the iHMP dataset

The study population consisted of 223 patients with IS and 205 patients with IR. To develop a model capable of distinguishing IS from IR, subjects were divided randomly in an 8:2 model training to holdout dataset ratio. On the other hand, a total of 10,784 features subsisted in five multi-omic analyses, whereas the overall proportion of RNAseq features accounted for 95.9% of the total dataset features, while the ratio of the microbiome features represented 0.3% of the total dataset. A brief overview of the baseline features of the iHMP dataset is presented in Table 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **iHMP Type 2 diabetes mellitus data description** | | | | | |
| **RNAseq** | 10346 (95.9%) | | | | | |
| **Proteomics** | 301 (2.8%) | | | | | |
| **Cytokine** | 66 (0.6%) | | | | | |
| **Clinical** | 41 (0.4%) | | | | | |
| **Microbiome** | 29 (0.3%) | | | | | |
|  | **IS (n = 25)** | | | **IR (n = 32)** | | ***P*** |
| **Age** | 56.144 ± 8.098 | | | 57.223 ± 7.062 | | 0.6 |
| **BMI** | 27.474 ± 3.614 | | | 29.953 ± 3.558 | | 0.013 |
| **Gender** | Male (n = 11) | | | Male ( n = 16 ) | | |
|  | **Training set** | | ***p*** | **Holdout dataset** | | ***p*** |
| **IS (n = 179)** | **IR (n = 164)** | **IS (n = 44)** | **IR (n = 41)** |
| **SSPG** | 101.083 ± 29.354 | 199.402 ± 35.19 | <0.001 | 105.645 ± 27.412 | 203.422 ± 31.473 | <0.001 |
| **GLU** | 101.994 ± 17.672 | 92.543 ± 11.954 | <0.001 | 93.386 ± 11.429 | 89.537 ± 10.46 | 0.109 |
| **HbA1c** | 5.713 ± 0.423 | 5.558 ± 0.359 | <0.001 | 5.666 ± 0.557 | 5.498 ± 0.313 | 0.088 |

Table 1. Baseline characteristics of the iHMP dataset and the model training and holdout datasets. SSPG = Steady state plasma glucose; GLU = Fasting glucose; HbA1c = Hemoglobin A1c. Data are represented as means ± standard deviation. For statistical analysis, Welch’s unequal variances *t-test* is used to analyze the significance between two groups.

3.2. Predictive models of IRIS with full features

Using all the features from the data, the discriminative models of IR from IS were built using training sets. All features were used to build the discriminative model. The model performances of the five different classifiers are given in Figure 1. The gradient boosting classifier had an AUC of 0.972 in the validation set and 0.919 in the holdout dataset. Because the gradient boosting classifier showed the best AUC in the validation set, the feature importance from the gradient boosting classifier was obtained to construct a better model for IRIS.

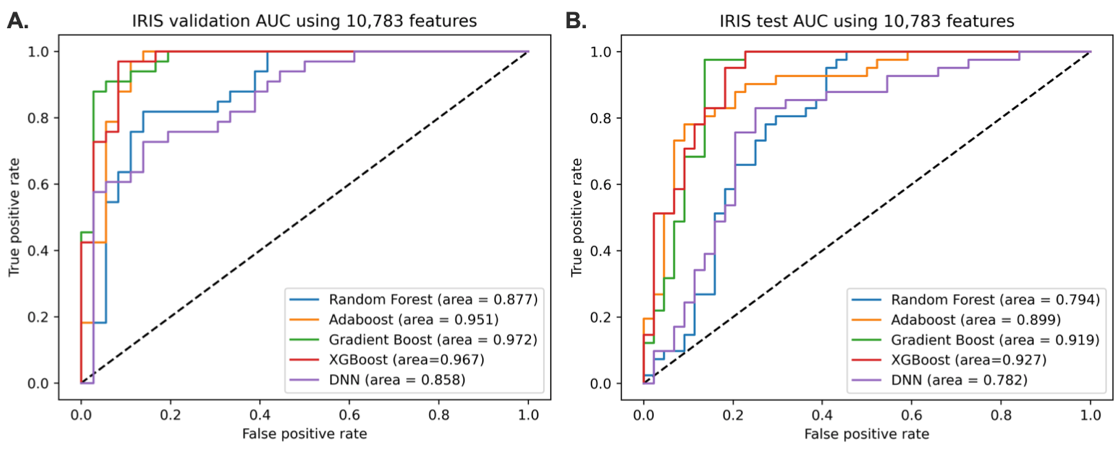


Figure 1. Performance of the models without feature selection. The discriminative models are built using 10,783 features. (A). AUC scores of 5 different classifiers using the validation set. (B). AUC scores of 5 different classifiers using the holdout dataset. Gradient boosting classifier is selected for showing the highest AUC in validation.

3.3. Feacture reduction and selection

Feature reduction was performed to increase the model performance. Feature importance magnitudes of the reduced features from the gradient boosting classifier were extracted and arranged from highest to lowest magnitude. Features with no feature importance magnitudes were discarded from the training model, and from the lowest magnitudes, features were erased sequentially and each time, the alteration of test AUC was observed.

Features were erased until the test AUC dropped to less than 0.98, and the remaining features were selected for each index. An overview of the method of feature selection is shown in Figure 2A. There were 24 features in index 1, 72 features in index 2, 155 features in index 3, 34 features in index 4, and 466 features in index 5. More information about the number of features is shown in Figure 2B. From all these indices, intersecting features were selected as the final features. The statistical analysis of the selected features in both the training and holdout dataset is shown in Supplementary Data S3.

Of the 10,783 features, 16 features were selected. Among the selected features, six were from clinical measurements, six were from cytokine profiles, three were from proteomics, one was from the microbiome, and none were from RNAseq.



Figure 2. Feature reduction method overview and Venn diagram of features from each index. (A). Schematic diagram of the feature reduction. Feature importance of 10,783 features from gradient boosting was obtained, and from the lowest magnitudes, the features were erased sequentially and each time, the alteration of test AUC was observed. Intersecting features with AUC lower than 0.98 were used for the further research. (B). Venn diagram of the features from each index. In total, 16 features were intersected in all 5 indices.

3.4. Predictive models of IRIS based on feature reduction

After the features were extracted and selected, the classification step using five different methods was performed on the resulting features. The performance of the models after feature selection is shown in Figure 3. Compared to the predictive models without feature reduction, some models showed improvements in performance. Notably, the test AUC of all the classifiers except Adaboost have increased.

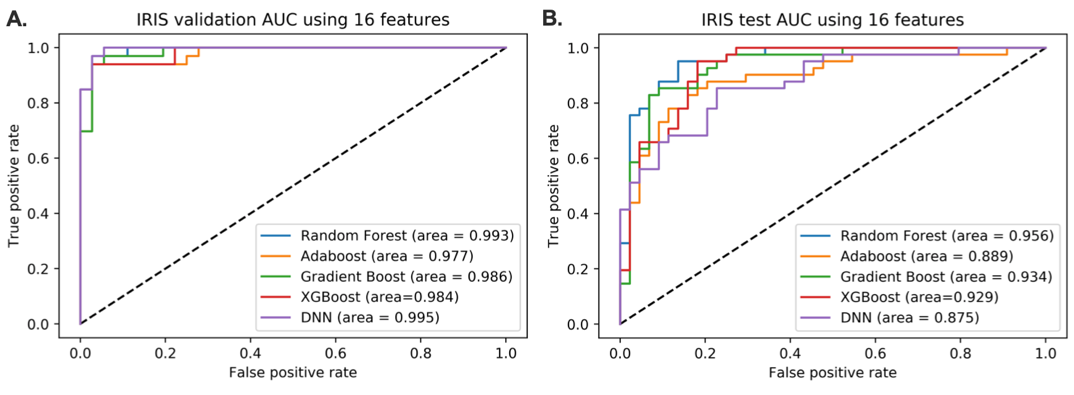


Figure 3. Performance of the models after the feature selection. The discriminative models were built using 16 features. (A). AUC scores of 5 different classifiers using the validation set. (B). AUC scores of 5 different classifiers using the holdout dataset.

3.5. Pairwise correlation network between the microbiome and extracted features

We subsequently constructed a pairwise correlation network over the extracted features and microbiome variables to discover the relationships between various biomedical characteristics. The pairwise correlation network of extracted features and the microbiome is shown in Figure 4. Based on Figure 4A, red lines connected between nodes represent positive correlation, and the green lines represent negative correlation. The thickness of the lines indicates a strong correlation between two nodes.



Figure 4. Pairwise correlation network of selected features with microbiome variables. (A). Pairwise correlation network of 16 selected features and microbiome variables were plotted without any additional options given. (B). Pairwise correlation network of 16 selected features and microbiome variables were plotted with threshold=false discovery rate and alpha=0.001. *HDL*= high density lipoprotein; c\_C= class\_*Clostridia*; o\_C= order\_*Clostridales*. Connected nodes of the selected features and microbiome variables are circled in blue. The red line represents positive correlation and the green line represents negative correlation between the nodes.

As shown in Figure 4B, high-density lipoprotein (HDL) was connected to class\_*Clostridia* and order\_*Clostridales*. These applicable features were trained on the classifiers by substituting the following features with microbiome variables.

3.6. Predictive model of IRIS with the replacement of corresponding features with microbiome variables

After applying the pairwise correlation network, we tested whether the model performances were sustainable with the replacement of the extracted features with the microbiome variables. Of the connected nodes between the microbiome variables and the extracted features, the corresponding features were replaced and used to build the disciminative model.

From 16 features, HDL was replaced with two microbiome variables, and a total of 17 features were used to build the discriminative model. The performances of these models are shown in Figure 5. To guarantee that the selected 17 features are well-extracted, random feature sampling permutation was performed (Supplementary Data S4).

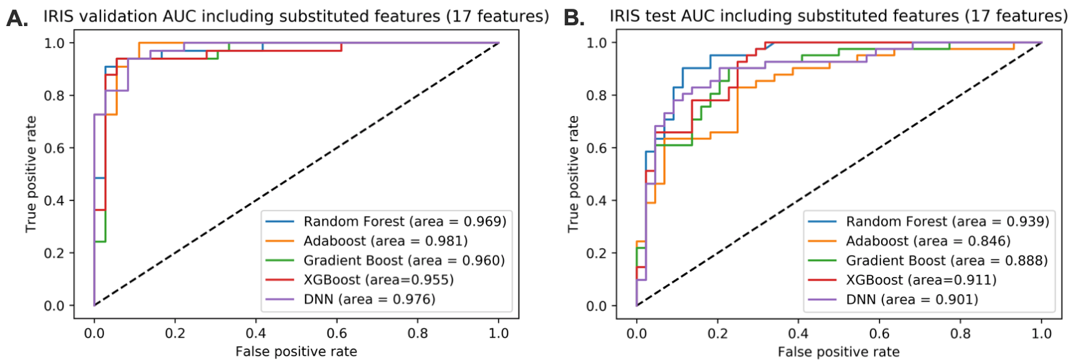


Figure 5. Model performance after substituting the corresponding features with microbiome variables. The discriminative models were built using 17 features. (A). AUC scores of 5 different classifiers using the validation set. (B). AUC scores of 5 different classifiers using the holdout dataset.

As a result, despite slight alterations in each model, substitution of selected features with microbiome features can sustain model performance (Figure 6). On the other hand, we compared our model result with other T2D models presented by different studies. Even though the characteristics of the data used in different studies were not the same, applied modern learning-machine techniques dealth with the issue of identifying patients with T2D or insulin resistance (Supplementary Data S5).



Figure 6. Overall test AUC scores including all previous models. Graphical analysis of model AUC score alteration from no feature reduction to feature substitution. *Before reduction* = model holdout dataset AUC without any feature reduction (10,783 features); *Feature selection* = model holdout dataset AUC with feature selection (16 features); *Feature substitution* = model holdhout dataset AUC after substituting corresponding features with microbiome variables (17 features).

3.7. Acquire the optimal model with the highest frequency in the random permutation

The method proposed by this paper is a fast feature selection method for features with large-scale datasets and a proper choice of an approach for handling a problem with a selection of suboptimal or diminished features after the classification step. Random sample permutation is performed with the selected 17 features to achieve the optimal model that elucidates the IRIS dataset. DNN was chosen for this analysis because it showed a consistent increment in AUC across other strategies in the holdout dataset among the training methods compared (Figure 6). The histogram for the test AUC scores of all permutations is shown in Figure 7. The advantage of selecting the central tendeny as the final IRIS model identifies as the representative of an entire distribution. DNN model with validation AUC of 0.9924 is a model that represents a case that is not unusual, but that occurs frequently. A model with validation AUC of 0.9924 had the AUC score of 0.9440 for the holdout dataset.

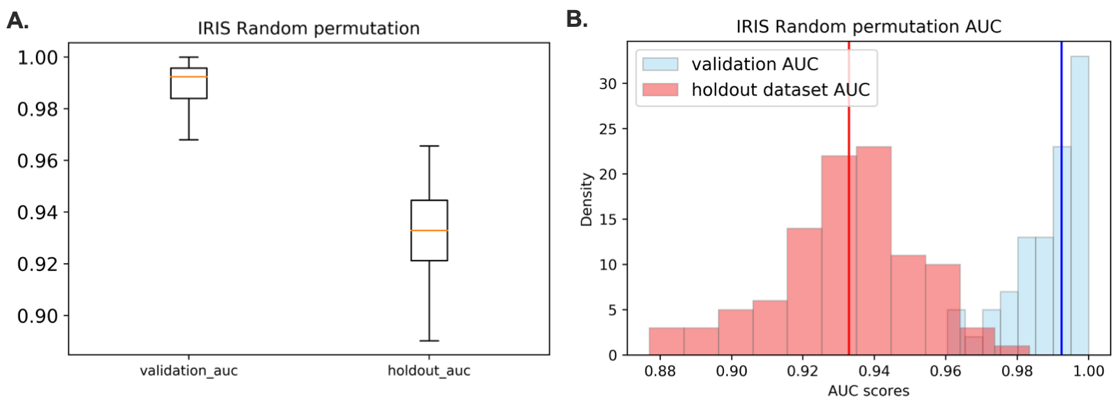
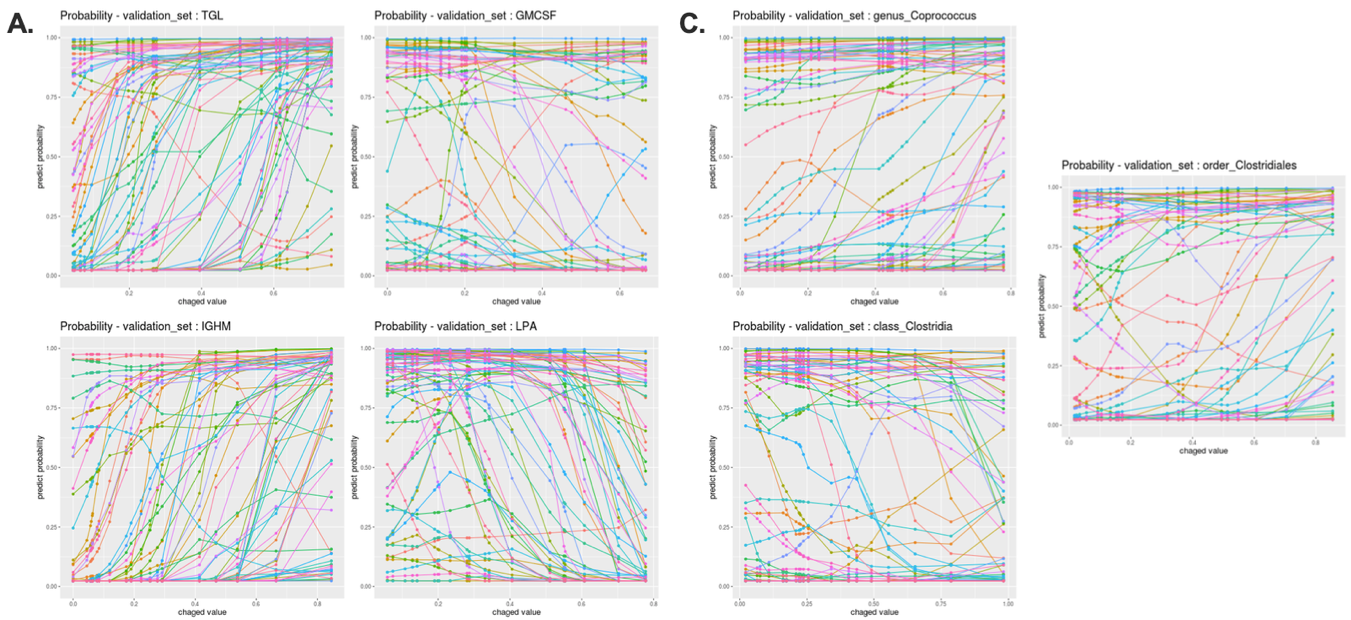


Figure 7. Graphical illustration of DNN model AUC scores after 101 times of random permutation. (A). Box and whisker plot of validation and holdout dataset AUC after 101 times of random permutation.

(B). The histogram shows the distribution of all test AUC scores for every permutation with the optimal combinations of learning rate, batch size and epoch. The vertical red line represents the median holdout dataset AUC which is at 0.9329 and the vertical blue line represents the median validation AUC which is at 0.9924.

3.8. Contribution of microbiome features to the DNN classification model

We applied the interpretation algorithm to determine the contribution of a single feature to the outcome of DNN model. We tested both in the validation and test datasets of the DNN model trained with 17 features. Figure 8 depicts the results of the top four features with the highest Shapley values (Supplementary Data S6) and microbiome features in both the validation and test datasets. The contribution of other features to the classification model is shown in Supplementary Data S7-8.



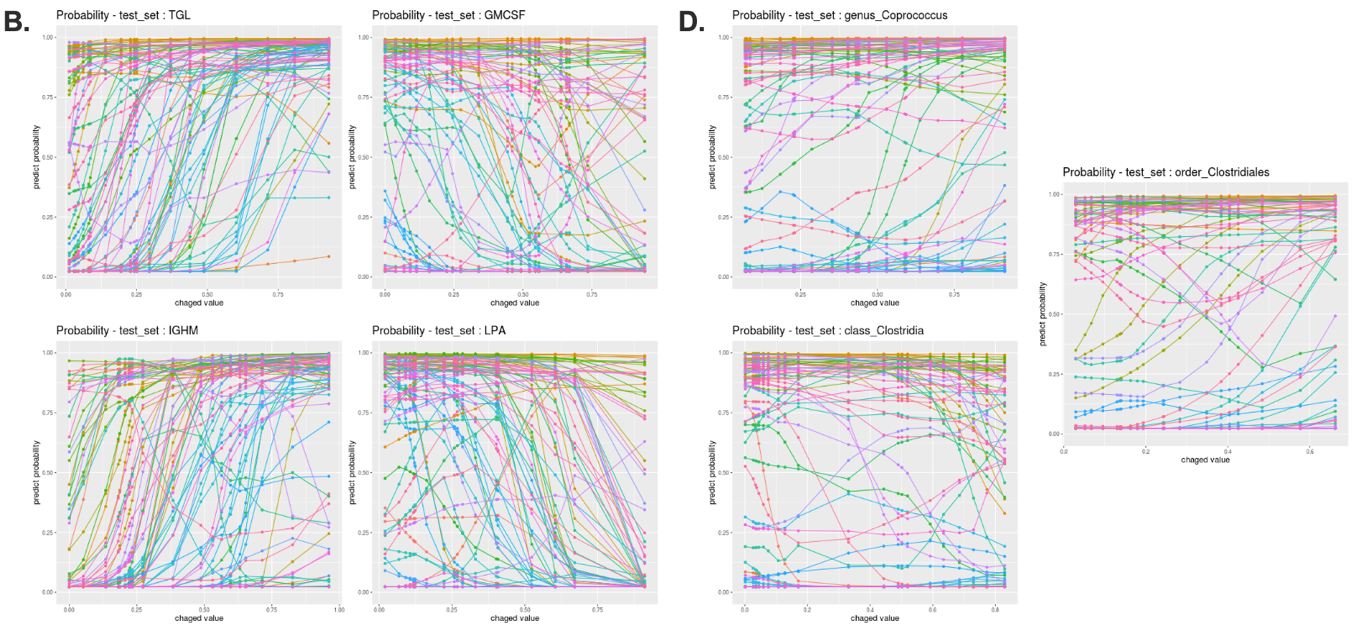


Figure 8. Graphical illustrations of features affecting the DNN classification. Graphical illustration of the contribution of a feature on the outcome of the DNN model. The x-axis represents the range of expression values of a feature and the y-axis represents the predict probability of the DNN model. If the predict probability is close to 1, it indicates that the sample is IR. (A). Probability outcome of the top 4 features with highest Shapley values of the DNN model from the validation set, (B). Probability outcome of top 4 features with the highest Shapley values of the DNN model from the holdout dataset, (C). Probability outcome of the microbiome features from the validation set and (D). Probability outcome of the microbiome features from the holdout dataset.

4. Discussion

The human microbiome plays an important role in human health, and there is growing evidence that the microbiome can be used as a predictor of various diseases. However, microbiome data pose a huge challenge due to uneven sampling depth, over-dispersion, and zero-inflation under high-dimensional microbiome profiles. Without several steps of careful data engineering, this imbalance induces the data to be highly sparse, which is detrimental to the model performance and may also increase redundant information. A new challenge is presented when the microbiome and other high-dimensional multi-omic datasets are compressed into low-dimensional features. In our study, when the model was generated with reduced features (16 features), the model performance of the five classifiers improved compared to that of the classifiers without feature reduction; however, the feature representation of the microbiome had almost no impact compared to other multi-omic profiles. Except for genus\_*Coprococcus*, other microbiome profiles had no influence on the classifier model.

To find meaningful feature representations of the microbiome, we substituted microbiome features with 16 reduced features using a pairwise correlation network. According to Figure 4B, only HDL had a significant connection with c\_C= class\_*Clostridia* and o\_C= order\_*Clostridiales.* The best performing classifiers were built and several model performances were sustained while the feature importance of the substituted microbiome features increased. These results indicate that some of the meaningful information may drop to prioritize the learning process. We attempted to substitute for 16 reduced features with more microbiome variables by using a pairwise correlation network with the combination of alpha = 0.01 and alpha = 0.05, to plot the significant edges. With alpha = 0.05, 8 features can be substituted with 16 microbiome variables and with alpha = 0.01, 2 features can be substituted with 5 microbiome variables, but all of the classifiers’ test AUCs from alpha = 0.05 and 0.01 dropped lower than 0.75 (data not shown). This does not mean that the dropped features (features with no feature importance magnitudes) are meaningless. We argue that additional features can improve the prediction performance, especially when a well-balanced set of features is augmented.

An interpretation algorithm was applied to evaluate the contribution of a single feature to the outcome of the DNN model. Interpreting the specific contribution of an individual feature is important because the identification of IR driving features in an individual may provide important information for treatment and prognosis. When the values of a feature increase, the probability outcomes of the stem cell factor (SCF) [16], lysophosphatidic acid (LPA) [17], granulocyte-macrophage colony-stimulating factor (GMCSF) [18], interleukin 7 (IL7) [19] and creatinine (CR) [20], and apolipoprotein E (APOE) [21-23] clearly switched from IR to IS, as proven by previous studies as having an inverse relationship with insulin resistance. On the other hand, when the values of a feature decrease, the probability outcomes of triglycerides (TGL) [24], monocyte absolute value (MONOAB) [25], and immunoglobulin heavy constant mu (IGHM) [26,27] switch from IS to IR, which have been previously reported to have a direct relationship with insulin resistance. In certain cases, the DNN's likelihood outcome was barely influenced by changes in the expression value from a single feature. For these samples, DNN was not significantly influenced by a single feature, but by the multiple expression values of the features. In other words, multiple features could adequately classify the samples as IR or IS, but no single feature was able to do so.

More research should be conducted to investigate the mean corpuscular volume (MCV), Fas ligand (FasL) [28], leptin [29,30], eotaxin [31], order\_*Clostridales*, class\_*Clostridia*, and monocyte chemoattractant protein-1 (MCP-1). Although MCV is irrelevant to insulin resistance, a study observed a positive correlation between the diabetes and prediabetes groups [32]. Unclear observations within patients are observed for FasL, leptin and eotaxin. Although these features are relevant to insulin resistance, patient characteristics could have caused different outcomes. Disease progression may vary greatly, which may enhance the implementation of precision medicine at the individual or a sub population level.

Though variation exists between people’s microbiomes, alteration in the host-microbiota is involved in the progress and development of human disease. Several reports observed that genus\_*Coprococcus* was enhanced in gestational diabetes mellitus patients [33-35]. In general, class\_*Clostridia* is reported to have an inverse relationship with insulin resistance and were reduced in the diabetic group compared to the control group [36]. On the other hand, one study reported that order\_*Clostridiales* was positively related to insulin sensitivity [37], which the result obtained in our study and the study reported previously clearly underline the link between the gut microbiome and insulin resistance. Although deeper research must be conducted to elucidate the link between insulin amelioration and human microbiome, but microbiome have the potential to be a good discriminant biomarker for IR.

Lastly, many studies have identified MCP-1 as an insulin-responsive cytokine that promotes insulin resistance and glucose intolerance [38-41]. Contrarily, one of the studies argued that elevated MCP-1 levels in plasma do not influence insulin signaling and have no effect on insulin resistance and glucose tolerance *in vivo* [42]. Based on the probability outcome of MCP-1, a convex shape is observed, meaning that MCP-1 may have markedly different prognoses for insulin resistance. Thus, the action of MCP-1 on insulin resistance remains unclear, and future studies are necessary to clarify this.

As with any large study, multi-omic datasets appear to be associated with certain variables, but it is not experimentally clear whether such variables are sufficient or informative for their associated disease phenotypes. Moreover, with the continued availability of large samples of multi-omic data, careful consideration must be given for the necessary information not to fade out in downstream prediction.

**Supplementary Materials:** The supplementary materials are available online at [www.mdpi.com/xxx/s1](http://www.mdpi.com/xxx/s1).

**Author Contributions**:

**Funding:**

**Acknowledgments**:

**Conflicts of Interest**: The authors declare no conflict of interest.

References

1. Sardaraz, M.; Tahir, M.; Ikram, A. A. Advances in High Throughput DNA Sequence Data Compression. *Journal of Bioinformatics and Computational Biology* **2016**, *14* (03), 1630002.
2. Lightbody, G.; Haberland, V.; Browne, F.; Taggart, L.; Zheng, H.; Parkes, E.; Blayney, J. K. Review of Applications of High-Throughput Sequencing in Personalized Medicine: Barriers and Facilitators of Future Progress in Research and Clinical Application. *Briefings in Bioinformatics* **2019**, *20* (5), 1795–1811.
3. Bansal, V.; Boucher, C. Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? *iScience* **2019**, *18*, 37–41.
4. Proctor; L M; Creasy; H H; Fettweis; J M. The Integrative Human Microbiome Project. *Nature* **2019**, *569* (7758), 641–648.
5. Fodor, A. A.; Desantis, T. Z.; Wylie, K. M.; Badger, J. H.; Ye, Y.; Hepburn, T.; Hu, P.; Sodergren, E.; Liolios, K.; Huot-Creasy, H.; Birren, B. W.; Earl, A. M. The “Most Wanted” Taxa from the Human Microbiome for Whole Genome Sequencing. *PLoS ONE* **2012**, *7* (7).
6. Nelson, K. E.; Weinstock, G. M.; Highlander, S. K.; Worley, K. C.; Creasy, H. H.; Wortman, J. R.; Rusch, D. B.; Mitreva, M.; Sodergren, E.; Chinwalla, A. T.; Feldgarden, M.; Gevers, D.; Haas, B. J.; Madupu, R.; Ward, D. V.; Birren, B. W.; Gibbs, R. A.; Methe, B.; Petrosino, J. F.; Strausberg, R. L.; Sutton, G. G.; White, O. R.; Wilson, R. K.; Durkin, S.; Giglio, M. G.; Gujja, S.; Howarth, C.; Kodira, C. D.; Kyrpides, N.; Mehta, T.; Muzny, D. M.; Pearson, M.; Pepin, K.; Pati, A.; Qin, X.; Yandava, C.; Zeng, Q.; Zhang, L.; Berlin, A. M.; Chen, L.; Hepburn, T. A.; Johnson, J.; Mccorrison, J.; Miller, J.; Minx, P.; Nusbaum, C.; Russ, C.; Sykes, S. M.; Tomlinson, C. M.; Young, S.; Warren, W. C.; Badger, J.; Crabtree, J.; Markowitz, V. M.; Orvis, J.; Cree, A.; Ferriera, S.; Fulton, L. L.; Fulton, R. S.; Gillis, M.; Hemphill, L. D.; Joshi, V.; Kovar, C.; Torralba, M.; Wetterstrand, K. A.; Abouellleil, A.; Wollam, A. M.; Buhay, C. J.; Ding, Y.; Dugan, S.; Fitzgerald, M. G.; Holder, M.; Hostetler, J.; Clifton, S. W.; Allen-Vercoe, E.; Earl, A. M.; Farmer, C. N.; Liolios, K.; Surette, M. G.; Xu, Q.; Pohl, C.; Wilczek-Boney, K.; Zhu, D. A Catalog of Reference Genomes from the Human Microbiome. *Science* **2010**, *328* (5981), 994–999.
7. Wylie, K. M.; Truty, R. M.; Sharpton, T. J.; Mihindukulasuriya, K. A.; Zhou, Y.; Gao, H.; Sodergren, E.; Weinstock, G. M.; Pollard, K. S. Novel Bacterial Taxa in the Human Microbiome. *PLoS ONE* **2012**, *7* (6).
8. Li, K.; Bihan, M.; Yooseph, S.; Methé, B. A. Analyses of the Microbial Diversity across the Human Microbiome. *PLoS ONE* **2012**, *7* (6).
9. Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., . . . Snyder, M. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature* **2019**, 569(7758), 663-671. doi:10.1038/s41586-019-1236-x
10. Mwangi, B.; Tian, T. S.; Soares, J. C. A Review of Feature Reduction Techniques in Neuroimaging. *Neuroinformatics* **2013**, *12* (2), 229–244.
11. Kondo, M.; Bezemer, C.-P.; Kamei, Y.; Hassan, A. E.; Mizuno, O. The Impact of Feature Reduction Techniques on Defect Prediction Models. *Empirical Software Engineering* **2019**, *24* (4), 1925–1963.
12. Li, J.; Lu, Q.; Wen, Y. Multi-Kernel Linear Mixed Model with Adaptive Lasso for Prediction Analysis on High-Dimensional Multi-Omics Data. *Bioinformatics* **2019**.
13. Coretto, P.; Serra, A.; Tagliaferri, R. Robust Clustering of Noisy High-Dimensional Gene Expression Data for Patients Subtyping. *Bioinformatics* **2018**.
14. Ho, N. T.; Li, F.; Wang, S.; Kuhn, L. MetamicrobiomeR: an R Package for Analysis of Microbiome Relative Abundance Data Using Zero-Inflated Beta GAMLSS and Meta-Analysis across Studies Using Random Effects Models. *BMC Bioinformatics* **2019**, *20* (1).
15. Ahn, T., Goo, T., Lee, C., Kim, S., Han, K., Park, S., & Park, T. Deep Learning-based Identification of Cancer or Normal Tissue using Gene Expression Data. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* **2018**. doi:10.1109/bibm.2018.8621108.
16. Li, J.-W.; Li, L.-L.; Chang, L.-L.; Wang, Z.-Y.; Xu, Y. Stem Cell Factor Protects against Neuronal Apoptosis by Activating AKT/ERK in Diabetic Mice. *Brazilian Journal of Medical and Biological Research* **2009**, *42* (11), 1044–1049.
17. D'Souza, K.; Paramel, G. V.; Kienesberger, P. C. Lysophosphatidic Acid Signaling in Obesity and Insulin Resistance. *Nutrients* **2018**, *10* (4), 399.
18. Kim, D.-H.; Sandoval, D.; Reed, J. A.; Matter, E. K.; Tolod, E. G.; Woods, S. C.; Seeley, R. J. The Role of GM-CSF in Adipose Tissue Inflammation. *American Journal of Physiology-Endocrinology and Metabolism* **2008**, *295* (5).
19. Lucas, S.; Taront, S.; Magnan, C.; Fauconnier, L.; Delacre, M.; Macia, L.; Delanoye, A.; Verwaerde, C.; Spriet, C.; Saule, P.; Goormachtigh, G.; Héliot, L.; Ktorza, A.; Movassat, J.; Polakowska, R.; Auriault, C.; Poulain-Godefroy, O.; Santo, J. D.; Froguel, P.; Wolowczuk, I. Interleukin-7 Regulates Adipose Tissue Mass and Insulin Sensitivity in High-Fat Diet-Fed Mice through Lymphocyte-Dependent and Independent Mechanisms. *PLoS ONE* **2012**, *7* (6).
20. Moon, J. S.; Lee, J. E.; Yoon, J. S. Variation in Serum Creatinine Level Is Correlated to Risk of Type 2 Diabetes. *Endocrinology and Metabolism* **2013**, *28* (3), 207.
21. Gao, J., Katagiri, H., Ishigaki, Y., Yamada, T., Ogihara, T., Imai, J., . . . Oka, Y. Involvement of Apolipoprotein E in Excess Fat Accumulation and Insulin Resistance. *Diabetes* 2006, 56(1), 24-33. doi:10.2337/db06-0144
22. Lyngdorf, L. Paradoxical reduction of atherosclerosis in apoE-deficient mice with obesity-related type 2 diabetes. *Cardiovascular Research* **2003**, 59(4), 854-862. doi:10.1016/s0008-6363(03)00506-6
23. Schreyer, S. A., Vick, C., Lystig, T. C., Mystkowski, P., & Leboeuf, R. C. LDL receptor but not apolipoprotein E deficiency increases diet-induced obesity and diabetes in mice. *American Journal of Physiology-Endocrinology and Metabolism* **2002**, 282(1). doi:10.1152/ajpendo.2002.282.1.e207
24. Lee, E. Y.; Yang, H. K.; Lee, J.; Kang, B.; Yang, Y.; Lee, S.-H.; Ko, S.-H.; Ahn, Y.-B.; Cha, B. Y.; Yoon, K.-H.; Cho, J. H. Triglyceride Glucose Index, a Marker of Insulin Resistance, Is Associated with Coronary Artery Stenosis in Asymptomatic Subjects with Type 2 Diabetes. *Lipids in Health and Disease* **2016**, *15* (1).
25. Kawarabayashi, R.; Motoyama, K.; Nakamura, M.; Yamazaki, Y.; Morioka, T.; Mori, K.; Fukumoto, S.; Imanishi, Y.; Shioi, A.; Shoji, T.; Emoto, M.; Inaba, M. The Association between Monocyte Surface CD163 and Insulin Resistance in Patients with Type 2 Diabetes. *Journal of Diabetes Research* **2017**, *2017*, 1–8.
26. Harmon, D. B., Srikakulapu, P., Kaplan, J. L., Oldham, S. N., Mcskimming, C., Garmey, J. C., . . . Mcnamara, C. A. Protective Role for B-1b B Cells and IgM in Obesity-Associated Inflammation, Glucose Intolerance, and Insulin Resistance. Arteriosclerosis, Thrombosis, and Vascular Biology **2016**, 36(4), 682-691. doi:10.1161/atvbaha.116.307166
27. Winer, D. A., Winer, S., Shen, L., Wadia, P. P., Yantha, J., Paltser, G., . . . Engleman, E. G. B cells promote insulin resistance through modulation of T cells and production of pathogenic IgG antibodies. *Nature Medicine* **2011**, 17(5), 610-617. doi:10.1038/nm.2353
28. Kumar, H., Mishra, M., Bajpai, S., Pokhria, D., Arya, A. K., Singh, R. K., & Tripathi, K. Correlation of insulin resistance, beta cell function and insulin sensitivity with serum sFas and sFasL in newly diagnosed type 2 diabetes. *Acta Diabetologica* **2011**, 50(4), 511-518. doi:10.1007/s00592-011-0307-8
29. Wang, J.; Obici, S.; Morgan, K.; Barzilai, N.; Feng, Z.; Rossetti, L. Overfeeding Rapidly Induces Leptin and Insulin Resistance. *Diabetes* **2001**, *50* (12), 2786–2791.
30. Osegbe, I.; Okpara, H.; Azinge, E. Relationship between Serum Leptin and Insulin Resistance among Obese Nigerian Women. *Annals of African Medicine* **2016**, *15* (1), 14.
31. Vasudevan, A. R., Wu, H., Xydakis, A. M., Jones, P. H., Smith, E. O., Sweeney, J. F., . . . Ballantyne, C. M. Eotaxin and Obesity. *The Journal of Clinical Endocrinology & Metabolism* **2006**, 91(1), 256-261. doi:10.1210/jc.2005-1280
32. Ziaee, A., Ghorbani, A., Kalbasi, S., Hejrati, A., & Moradi, S. Association of hematological indices with prediabetes: A cross-sectional study. *Electronic Physician* **2017**, 9(9), 5206-5211. doi:10.19082/5206
33. Naderpoor N, Mousa A, Gomez-Arango LF, Barrett HL, Dekker Nitert M, de Courten B. Faecal Microbiota Are Related to Insulin Sensitivity and Secretion in Overweight or Obese Adults. *J Clin Med*. **2019**, 8(4):452. doi:10.3390/jcm8040452
34. Wang, J.; Zheng, J.; Shi, W.; Du, N.; Xu, X.; Zhang, Y.; Ji, P.; Zhang, F.; Jia, Z.; Wang, Y.; Zheng, Z.; Zhang, H.; Zhao, F. Dysbiosis of Maternal and Neonatal Microbiota Associated with Gestational Diabetes Mellitus. *Gut* **2018**, 67 (9), 1614–1625.
35. Kuang, Y.-S.; Lu, J.-H.; Li, S.-H.; Li, J.-H.; Yuan, M.-Y.; He, J.-R.; Chen, N.-N.; Xiao, W.-Q.; Shen, S.-Y.; Qiu, L.; Wu, Y.-F.; Hu, C.-Y.; Wu, Y.-Y.; Li, W.-D.; Chen, Q.-Z.; Deng, H.-W.; Papasian, C. J.; Xia, H.-M.; Qiu, X. Connections between the Human Gut Microbiome and Gestational Diabetes Mellitus. *GigaScience* **2017**, 6 (8).
36. Larsen, N.; Vogensen, F. K.; Frans W. J. Van Den Berg; Nielsen, D. S.; Andreasen, A. S.; Pedersen, B. K.; Al-Soud, W. A.; Sørensen, S. J.; Hansen, L. H.; Jakobsen, M. Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults. *PLoS ONE* **2010**, 5 (2).
37. Allin, K. H.; Tremaroli, V.; Caesar, R.; Jensen, B. A. H.; Damgaard, M. T. F.; Bahl, M. I.; Licht, T. R.; Hansen, T. H.; Nielsen, T.; Dantoft, T. M.; Linneberg, A.; Jørgensen, T.; Vestergaard, H.; Kristiansen, K.; Franks, P. W.; Hansen, T.; Bäckhed, F.; Pedersen, O. Aberrant Intestinal Microbiota in Individuals with Prediabetes. *Diabetologia* **2018**, 61 (4), 810–820.
38. Chacón, M. R., Fernández-Real, J. M., Richart, C., Megía, A., Gómez, J. M., Miranda, M., . . . Vendrell, J. Monocyte Chemoattractant Protein-1 in Obesity and Type 2 Diabetes. Insulin Sensitivity Study\*. *Obesity* **2007**, 15(3), 664-672. doi:10.1038/oby.2007.578
39. Kanda, H. MCP-1 contributes to macrophage infiltration into adipose tissue, insulin resistance, and hepatic steatosis in obesity. *Journal of Clinical Investigation* **2006**, 116(6), 1494-1505. doi:10.1172/jci26498
40. Lin, Y., Ye, S., He, Y., Li, S., Chen, Y., & Zhai, Z. Short-term insulin intensive therapy decreases MCP-1 and NF-κB expression of peripheral blood monocyte and the serum MCP-1 concentration in newlydiagnosed type 2 diabetics. *Archives of Endocrinology and Metabolism* **2018**. doi:10.20945/2359-3997000000029
41. Westerbacka, J., Cornér, A., Kolak, M., Makkonen, J., Turpeinen, U., Hamsten, A., . . . Yki-Järvinen, H. Insulin regulation of MCP-1 in human adipose tissue of obese and lean women. *American Journal of Physiology-Endocrinology and Metabolism* **2008**, 294(5). doi:10.1152/ajpendo.00653.2006
42. Gogh, I. J., Oteng, A., Alex, S., Hamers, N., Catoire, M., Stienstra, R., . . . Kersten, S. Muscle-specific inflammation induced by MCP-1 overexpression does not affect whole-body insulin sensitivity in mice. *Diabetologia* **2015**, 59(3), 624-633. doi:10.1007/s00125-015-3822-2