Combination of Principal Component Analysis and Genetic Algorithm for Microbial Biomarker Identification in Obesity

Ping Zhang
Menzies Health Institute QLD
Griffith University
Gold Coast, Australia
p.zhang@griffith.edu.au

Nicholas West
School of Medical Science
Griffith University
Gold Coast, Australia
n.west@griffith.edu.au

Pin-Yen Chen
Menzies Health Institute QLD
Griffith University
Gold Coast, Australia
Pin-yenfiona.chen@griffithuni.edu.au

Allan Cripps
School of Medicine
Griffith University
Gold Coast, Australia
allan.cripps@griffith.edu.au

Amanda Cox
School of Medical Science
Griffith University
Gold Coast, Australia
a.cox@griffith.edu.au

Abstract

Background: A large number of microbial species have been detected in human faecal samples, with many of the species having high correlations with each other. Principal components analysis (PCA) is often used to find characteristic patterns associated with certain diseases by reducing variable numbers before a predictive model is built, particularly when some variables are correlated. Usually, the first two or three components from PCA are used to see whether individuals can be clustered into two classification groups based on pre-determined criteria: control and disease group. However, there might be a combination of other components that better distinguish diseased individuals from healthy controls. Genetic algorithms (GA) can be useful and efficient for searching the best combination of variables to build a prediction model. This study aimed to develop a prediction model that combines PCA and GA for identifying sets of bacterial species associated with high body mass.

Results: GA has selected the subsets of the principal components (PCs) produced by PCA. The prediction models built with theses PCs produced much higher area under the curve (AUC) values compared to the models built using top PCs which explained the most variance in the sample. The combinatorial effect of the identified bacterial species that contributed the most to the PCs may be associated with body mass.

Conclusions: The proposed algorithm overcomes the limitation of using PCA for prediction modelling. The application of the algorithm on an obesity study has shown the value of applying GA for selecting PC subsets from PCA to improve prediction models. The variables included in the PCs that were selected by GA can be combined with flexibility for potential clinical applications. The algorithm can be useful for many biological studies where high dimensional data are collected with highly correlated variables.

Keywords— PCA, genetic algorithm, Obesity, biomarker

I. INTRODUCTION

A number of studies have shown an association between the human gut microbiome with a diverse range of health issues [1, 2]. Knight and colleagues [3] reviewed the practices in microbiome studies, including: experimental design, choice of molecular analysis technology, methods for data analysis, and the integration of multiple omics data sets. Different methods for surveying microbial communities include 16S ribosomal RNA, metagenomic and metatranscriptomic sequencing. Next-step data analyses are needed to search for overall patterns in microbiome variation. The association between obesity and gut microbiota from the phylum level to the species level has been studied and various results have been reported [4, 5, 6].

Quite a few well-known sequence data analysis pipelines for microbiota study have been published, for example Quantitative Insights into Microbial Ecology (QIIME) [7], MetaGenome Rapid Annotation using Subsystem Technology (MG-RAST) [8] and mothur [9]. These packages include the functions of sequencing alignment, operational taxonomic unit (OTU) identification, taxonomy classification, and alpha and beta diversity calculation. They have been widely used for different biological and medical research purposes, such as associating gut microbiota diversity with diseases [10, 11, 12, 13]. It is important to recognise that due to some possible pitfalls in sample processing, the abundance of specific bacterial species and overall community composition can be distorted, thus hampering the analysis and threatening the validity of the research findings [14]. In addition, a key limitation of using 16S rRNA gene analysis for genus and species level classification is that related bacterial species may be indistinguishable due to near identical 16S rRNA gene sequences [15]. The potential for different data analysis approaches to impact on outcomes has also been recognised. Plummer et al. [15] compared three pipelines commonly used for 16S rRNA gene analysis: QIIME, MG-RAST and mothur. Favourably, their results showed that the three pipelines assessed produce comparable results for analysis of faecal samples, in terms of alpha diversity analysis and usability. Although a difference was observed between the pipelines in terms of taxonomic classification of genera from the Enterobacteriaceae family, the three pipelines detected the same phylum in similar abundances. A statistically significant difference was observed between two bioinformatics pipelines, QIIME and MG-RAST, with regards to beta diversity measures. D'Argenio et al. also compared QIIME and MG-RAST, and observed a

statistically significant difference between these two bioinformatics pipelines with regards to beta diversity measures.

While bioinformaticians work hard to ensure the development of high quality analysis pipelines, it is recognised that the complexity and variability of human microbiota can be sensitive to various environmental factors [16]. Efforts for improved analytical pipelines have been complicated by the limitation of available sample material and relatively high cost of microbiome profiling. This has meant that most of the microbiota studies involved limited sample sizes. In addition to efforts to improve the accuracy of OTU detection and taxonomy classification, especially at the genus and species levels, researchers have been studying ways to characterise diseases based on microbial composition. Rather than simply associating diseases and individual microbial features, such as a phylum or species, studies have looked at defining microbial signatures for specific diseases. This includes the application of computational modelling and variable selection techniques. For example, Rivera-Pinto etal [17] presented a greedy stepwise algorithm for selection of microbial signatures that preserves the principles of compositional data analysis. Sze and Schloss [18] performed a meta-analysis on associations between specific microbiome-based markers and obesity, concluding that although there was support for a relationship between human faecal microbial communities and obesity status, this association was relatively weak and its detection is confounded by large interpersonal variation and insufficient sample sizes. They also tested random forest models for classifying individuals as obese on the basis of the composition of their microbiome and did not find obvious patterns that could separate the obese and healthy groups. Random forest models were built by Peters et al [19] to find a taxonomic signature of obesity. The models were evaluated with Receiver Operator Characteristic (ROC) curve, and the area under the curve (AUC) produced by the optimal model, which included 49 OTUs, was 0.81. When the repeated cross-validation was performed the AUC was 0.65. Other machine learning methods used for microbiota study have been reviewed by Knights et al [20].

With the potential for large numbers of microbial species to be identified in human faecal samples and the high correlation between many of the species detected, principal components analysis (PCA) is often used. Studies use PCA to find characteristic patterns associated with certain diseases by reducing variable numbers based on their correlation with a principal component (PC), before a predictive model is built. Usually, the first two or three components from PCA (which account for the greatest proportion of the variance in the dataset) are used to see whether individuals clustered into two classification groups based on pre-determined criteria: control and disease group. However, we have asked the following questions: (i) Is it possible that the proportion of variance captured by the first two or three PCs is unrelated to the disease groups, and that the variance explained by other components is able to better distinguish disease individuals from healthy controls? (ii) Are there different groups of bacterial species associated with individual obesity?

With these questions in mind, we developed a prediction method that combines PCA and a genetic algorithm (GA) for microbial biomarkers identification. We applied this approach to faecal microbial data collected from obese and healthy weight individuals to identify potential sets of bacterial species that may be associated with obesity.

II. Methodology

A. Principal Components Analysis

PCA is often used as a tool in exploratory data analysis for variable dimensionality reduction prior to building predictive models. It can be used to reduce a large number of predictor variables to a few principal components (PCs), particularly in datasets that are noisy or have strong correlated explanatory variables. The principal components can then be used to build the predictive models. The principal components are the linear combinations of the original variables that account for the variance in the data. PCA can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix. The coefficients corresponding to each variable in the linear combinations indicate the relative weight the variable in the component. The larger the absolute value of the coefficient, the more important the corresponding variable is in calculating the component. To make the coefficient value for each variable comparable, the data should be normalized to have the same units of measurement before PCA is used.

B. Genetic Algorithms

A genetic algorithm is a search heuristic to find optimal solutions by mimicking Charles Darwin's theory of natural evolution--fittest individuals are selected for reproduction of the next generation. In GA, the potential solutions compete and mate with each other to produce increasingly fitter individuals over multiple generations.

Genetic algorithms can be useful and efficient for searching a combination of variables for the best achievement (eg. accuracy of prediction). GAs have been developed and applied for biomarker profile identification in a range of settings including Alzheimer's disease progression and breast cancer diagnosis [21, 22]. The GA algorithms have also been modified and improved to adapt to different computational environments and for different applications [23, 24]. An application of GA for for selecting vaginal microbiome features associated with bacterial vaginosis was found in [25]. However, the actual features were not reported as authors explained that evaluation was needed from both microbial and clinical perspectives in the future.

In this study, GA is used to find the best subset of principal components produced from a PCA using gut microbial species data.

III. Proposed Method

The method proposed here uses normalized OTU abundance with taxonomy assigned across the sample as the input for PCA. The OTUs can be identified by any of the sequence analysis pipelines mentioned above or other software packages, such as "DADA2" [26] in R (https://cran.r-project.org/). A GA is then applied for selection of the best set of components created from the PCA to predict the individuals as obese or having healthy body mass. The scores of selected principal components calculated for each individual are used as the input for building a classification model. ROC curve analysis is used to evaluate

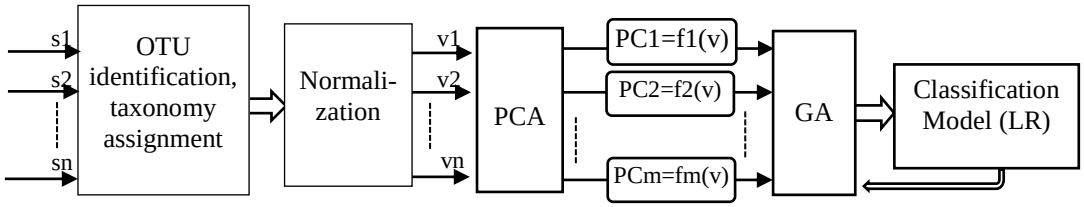

Fig. 1. A diagram of proposed method. s1, s2...sn are the 16S rRNA sequences for this study (can be from other sequencing). v1 to vn are the abundance (normalized) of each species detected in each individual. m=number of PCs created by PCA, n=number of individuals included in the sample. PCA is used to produce PC scores for each individual, and GA is used to select the best subset of PCs to distinguish obesity from healthy cases.

the classification models and is used as the fitness function for the GA. The method is shown diagrammatically in Fig. 1. In this research, logistic regression (LR) is used for building the classification models and more details about how to implement the GA can be found in reference [21].

IV. Experiments and Result

In this study, faecal samples from 22 obese and 105 healthy-weight individuals were sequenced with 16S-based approach. VSEARCH [27] was used for OTU identification and the Greengenes database was used as the reference for taxonomy classification. From the 68590 OTUs identified, 163 species were mapped to the Greengenes database. Once the OTUs were identified for each individual with taxa assigned, the species abundance values were normalized to the range of [0, 1] for each species, with the highest abundance across the individuals as 1 and the lowest 0. Species that were found to have zero abundance in 80% of both healthy and obesity groups were excluded, leaving 37 species for further analysis. The PCA and GA models were based on these 37 species.

For experiments, we performed 3 PCA using: the whole dataset, obese only samples, and healthy weight only samples. This idea is based on the possibility that for different populations the correlation between species might differ. The function "prcomp" from the "stats" package in R [28] was used to create the PCs and calculate the scores for each individual. These scores were then used to build the classification model with GA to select the best components for predicting obesity. The algorithm used by "prcomp" for creating the PCs can be found in [29]. Essentially, the PC calculation is performed by a singular value decomposition of the data matrix. If there are n observations with p variables, then the number of distinct PCs is min(n,p).

GA was completed with the fitness function of cross-validated AUC value created from logistic regression model. More explanation about AUC can be found in [21] and is also well documented elsewhere. Constraints for GA were put to include 1 to 6 PCs in the classification model. 10 times repeated five-fold cross-validation was used for testing the classification model with selected PCs. With each data set (all, healthy or obesity), GA was run 100 times repeatedly. The PC sets that were selected the most in the repeated runs were chosen as the final result. From the result (Table I) we can see the selection from GA was quite consistent with slight variation from each run.

The PCAs constructed from the whole data set and healthy individuals both created 37 principal components (PC1 to PC37) while the PCA from obese-only individuals created 22 components (PC1 to PC22). Table I listed the sets of PCs selected by GA and the cross-validated AUC produced from each prediction model built with the selected PC(s). The symbols "+" or "-" following the PC numbers indicate whether the coefficient of this PC is positive or negative value in the corresponding classification model. Positive coefficient means that increased score of this PC will increase the probability of the individual being characterised as obese. For example, PC1+ represents that the first PC created from the species abundance data will have a positive contribution to obesity.

Table II lists the top five species that have the highest contribution to each PC selected by GA. The symbols "+" or "-" following the species names indicate whether it has positive or negative contribution to the corresponding PC. For example, Prausnitzii- under PC1 represents that Prausnitzii has negative correlation with PC1. That means increased Prausnitzii abundance will decrease the PC1 value. As PC1 has positive correlation with being overweight, so that when Prausnitzii abundance increases the probability of being obese is reduced.

From the results presented in Table I and Table II, we analysed each of the species and categorized them into two groups, positive (indicated with asterisk (*) in Table II) or

TABLE I. GA selected PCs and the classification model performance (ROC)

Data for creating PCA	Result	Model _6 PCs	Model _5 PCs	Model _4 PCs	Model _3 PCs	Model _2 PCs	Model _1 PC
All	**PCs selected**	PC1+, PC2-, PC7+, PC11+, PC15-, PC27-	PC1+, PC2-, PC7+, PC11+, PC27-	PC1+, PC2-, PC7+, PC27-	PC1+, PC2-, PC7+	PC1+, PC7+ (orPC2-)	PC1+
	AUC (CV)	***0.87***	***0.85***	***0.84***	***0.81***	***0.77***	***0.69***
Obesity	**PCs selected**	PC2-, PC4-, PC14+, PC16-PC18+, PC19-	PC2-, PC4-, PC14+, PC18+ PC19-	PC2-, PC4-, PC14+, PC18+	PC2-, PC14+, PC18+	PC14+, PC18+	PC14+
	AUC (CV)	***0.92***	***0.92***	***0.90***	***0.87***	***0.84***	***0.80***
Healthy	**PCs selected**	PC1+,PC3+, PC5-,PC23+, PC28-,PC34+	PC1+,PC3+, PC23+,PC28-PC34+	PC1+,PC23+, PC28-,PC34+	PC1+,PC23+, PC34+	PC1+,PC34+	PC1+
	AUC (CV)	***0.92***	***0.90***	***0.88***	***0.87***	***0.83***	***0.72***

\+ Positive correlation coefficient in the model
\- Negative correlation coefficient in the model

TABLE II. TOP SPECIES INCLUDED IN THE SELECTED 1, 2, 3, 4, 5 OR 6 PCS PRODUCED WITH DIFFERENT DATA SETS

Data for creating PCA	High contribution variables (high coefficients in the corresponding PC) included in the most selected components					
	Comp1	***Comp2***	***Comp3***	***Comp4***	***Comp5***	***Comp6***
All (PC1, PC7, PC2, PC27, PC11, PC15)	Prausnitzii-* Eutactus-* Formicigenerans-* Catus-* Faecis-*	Gnavus+ Faecis-* Copri+ Muciniphila-* Adolescentis-*	Eutactus+* Prausnitzii+* Aerofaciens- Catus- Adolescentis-`	Moorei- Obeum- Lenta+* Animalis- Torques-	Eggerthii-* Dispar-* Adolescentis+ Mucilaginosa-* Aerofaciens+	Zeae+* Gnavus- Stutzeri+* Bromii+* Fragilis+*
Obesity (PC14, PC18, PC2, PC4, PC19, PC16)	Eutactus-* Bromii+ Adolescents-* Formicigenerans+ Producta-*	Uniformis+ Catus-* Dispar+ Faecis+ Distasonis-*	Dolichum- Lenta- Aerofaciens+* Producta- Gnavus-	Producta- Prausnitzii+* Aerofaciens- Fragilis- Faecis+*	Caccae+* Parainfluenzae+* Formicigenerans+* Adolescentis- Dispar-	Formicigenerans+* Bromii- Distasonis- Eutactus+* Perfringens+*
Healthy (PC1, PC34, PC23, PC28, PC3, PC5)	Prausnitzii-* Eutactus-* Catus-* Formicigenerans-* Faecis-*	Stutzeri-* Zeae+ Gnavus+ Dispar+ Lenta-*	Callidus-* Moorei+ Formigenes+ Prausnitzii+ Catus-*	Ovatus- Longum+* Distasonis+* Fragilis- Aerofaciens-	Copri+ Muciniphila-* Formigenes-* Catus+ Biforme+	Copri+* Muciniphila+* Prausnitzii- Formigenes+* Eutactus+*

*Species has a positive correlation with the probability of having healthy body mass.
+ Positive correlation with the corresponding PC
- Negative correlation with the corresponding PC

negative correlations with probability of having healthy body mass. The combination of having any one of the microbial species from each column can be a set of species that can have a high impact on health. For example, based on the result from the first set of the experiment which ran PCA on the sample set for both obesity and healthy combined (all), either "Prausnitzii, Faecis, Eutactus, Lenta, Eggerthii and Zeae", "Formicigenerans, Faecis, Eutactus, Lenta, Eggerthii and Zeae" or "Prausnitzii, Formicigenerans, Faecis, Eutactus, Lenta, Eggerthii and Zeae" can be a combination to have potential benefit on health. High values on Gnavus, Catus, Moorei and Aerofaciens together are associated with high probability with obesity.

A final classification model was built with each set of PCs selected by GA and first 1 to 6 PCs (which explain the most variance of the data) from the PCA. Again, the PCs were calculated from whole dataset, healthy set or obesity set. The AUCs produced from the GA selected PCs were quite obviously higher than the ones from the top PCs of PCA. Fig. 2 shows the ROCs created from the models built with the selected PCs and the first PCs of the PCA. The PCs in the graph were calculated with the healthy cohort (comparing with the result from the PCs calculated from whole dataset and obesity set, the first PCs from healthy data produced the highest AUC values).

V. DISCUSSION AND CONCLUSION

In this study, a computational method that combines PCA and GA has been proposed to produce more accurate prediction result and to find sets of features (variables) that contribute the most to the classification models. The model was applied for finding sets of bacterial species associated with high body mass. Due to the high correlation between many species of the gut bacteria, constructing PCA before the GA can improve the efficacy of GA for selecting multiple sets of microbial species associated with obesity. The result from this study showed that the prediction models built with the PCs selected by GA produced much higher AUC values than the models built with the top PCs that explained the greatest proportion of the variance in the sample. This demonstrates the value of applying GA for selection of subsets of PCs from PCA to improve the performance of prediction models. The features included in the PCs that were selected by GA can be combined with flexibility for potential clinical applications. With the flexible option of combining the features included

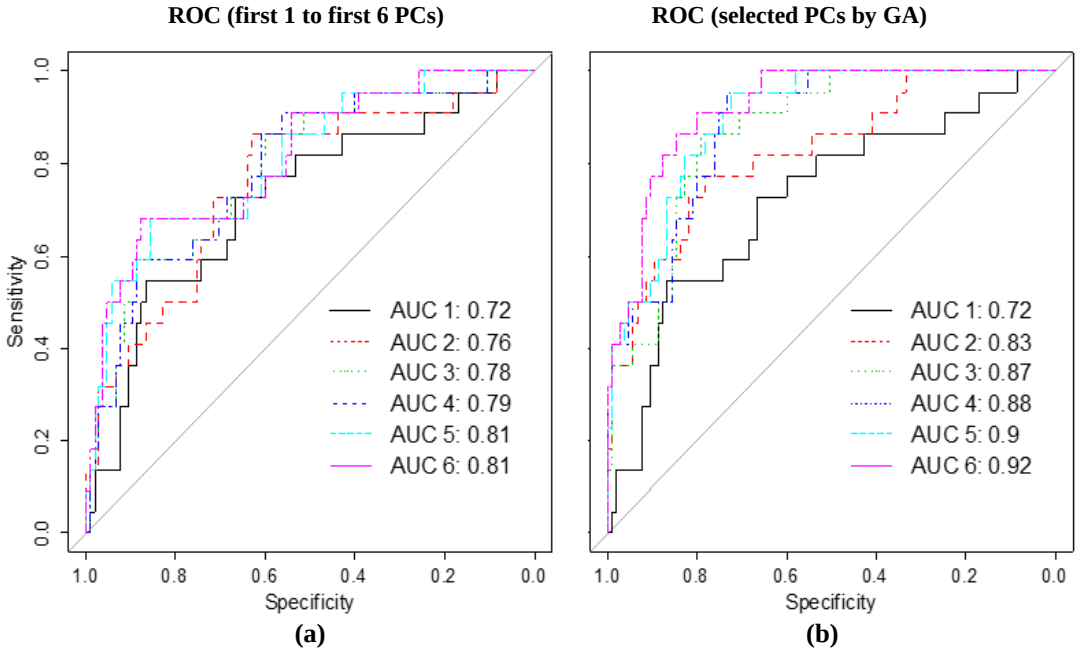

Fig. 2. ROC produced from the top PCs of PCA and from the PCs selected by GA

(a): 1--PC1, 2--PC1+PC2, 3--PC1+PC2+PC3, 4--PC1+PC2+PC3+PC4, 5-- PC1+PC2+PC3+PC4+PC5, 6-- PC1+PC2+PC3+PC4+PC5+PC6;
(b): 1--PC1, 2--PC1+PC34, 3--PC1+PC34+PC23, 4--PC1+PC23+PC28+PC34, 5-- PC1+PC3+PC23 +PC28+PC34, 6--PC1+PC3+PC5+PC23+PC28+PC34

in the PCs selected by the GA, different interventions can be recommended for different patients, which contributes to the practice of personalised medicine. The proposed algorithm was designed in a general way and was tested in the obesity study. It can be applied for any other classification or biomarkers identification study.

In the microbiome study, the presented result depends on the accuracy of the sequencing analysis. The microbial species identified here was based on the sample of 22 individuals with obesity and 105 healthy individuals. Assuming this result was validated in multiple datasets with bigger sample sizes, the results from Table 1 and Table 2 can suggest a few flexible combinations of microbial species groups that are beneficial to health. As described in the previous sections, the bacterial species detected can be different when applying different sequencing analysis and taxonomy classifications. To validate the findings from this study, the presented algorithm should be run with the outcomes from metagenomics sequencing and with other sequencing analysis pipelines. Different reference databases (e.g. NCBI) can be used for taxonomy classification of the OTUs identified. The proposed model takes into account correlations of the variables (bacteria species in this study) and the advantages of GA for feature selection. It overcomes the limitations of the ways in which PCAs are currently used for prediction modelling. The algorithm can be useful for many biological studies where high dimensional data are collected with strongly correlated variables.

ACKNOLEDGEMENT

The authors would like to thank the facility support from Queensland Facility for Advanced Bioinformatics (https://qfab.org/) and the participants of this study for their value contributions. Salary support for PZ and AJC was provided by the Griffith University Area of Strategic Investment in Chronic Disease Prevention. The microbial compositional profiling data used was generated as part of projects funded by the Australian Institute of Sport and the Gold Coast Hospital Foundation.

REFERENCES

[1] M. A. Jackson, S. Verdi, M.-E. Maxan, C. M. Shin, J. Zierer, R. C. E. Bowyer, T. Martin, F. M. K. Williams, C. Menni, J. T. Bell, T. Sector and J. J. Steves0, "Gut microbiota associations with common diseases and prescription medications in a population-based cohort," Nature Communication, vol. 9, p. Article number: 2655 , 2018.

[2] J. A. Gilbert, M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch and R. Knight, "Current understanding of the human microbiome," Nature Medicine, vol. 24, p. 392–400, 2018.

[3] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, L. R. Thompson, A. Tripathi, . Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso and . P. C. Dorrestein, "Best practices for analysing microbiomes," Nature Reviews Microbiology, vol. 16, no. 7, pp. 410-422, 2018.

[4] F. Ottosson, L. Brunkwall, U. Ericson, P. Nilsson, P. Almgren, C. Fernandez, O. Melander and M. Orho-Melander, "Connection Between BMI-Related Plasma Metabolite Profile and Gut Microbiota," The Journal of Clinical Endocrinology Metabolism, vol. 103, no. 4, pp. 1491-1501, 2018.

[5] M. Million, J.-C. Lagier, D. Yahav and M. Paul, "Gut bacterial microbiota and obesity," Clinical Microbiology and Infection, vol. 19, no. 4, pp. 305-313, 2013.

[6] C. K. Chakraborti, "New-found link between microbiota and obesity," World Journal of Gastrointestinal Pathophysiology, vol. 6, no. 4, pp. 110-119, 2015.

[7] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights and J. E. Koenig, "QIIME allows analysis of high-throughput community sequencing data," Nature Methods, p. doi:10.1038/nmeth.f.303, 2010.

[8] K. P. Keengan, E. M. Glass, F. Meyer. "MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function," Methods in Molecular Biology, vol. 1399, pp. 207-233,doi: 10.1007/978-1-4939-3369-3_13, 2016.

[9] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger and D. J. V. Horn, "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities," Applied and Environmental Microbiology, vol. 75, no. 23, pp. 7537-7541, 2009.

[10] G. G. Han, J.-Y. Lee, G.-D. Jin, J. Park, Y. H. Choi, B. J. Chae, E. B. Kim and Y.-J. Choi, "Evaluating the association between body weight and the intestinal microbiota of weaned piglets via 16S rRNA sequencing," Veterinary microbiology, vol. 196, pp. 55-62, 2016.

[11] J. Clemente, L. Ursell, L. Parfrey and K. R., "The impact of the gut microbiota on human health: an integrative view," Cell, vol. 148, no. 6, pp. 1258-1270, 2012.

[12] M. Spencer, T. Hamp, R. Reid, L. Fischer, S. Zeisel and A. Fodor, "Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency," Gastroenterology, vol. 140, no. 3, pp. 976-86. doi: 10.1053/j.gastro.2010.11.049. Epub 2010 Dec 1., 2011.

[13] L. Zhong, E. R. Shanahan, A. Raj, N. A. Koloski, L. Fletcher, M. Morrison, M. M. Walker, N. J. Talley and G. Holtmann, "Dyspepsia and the microbiome: time to focus on the small intestine," Gut, pp. DOI:10.1136/gutjnl-2016-312574, 2016.

[14] J. P. Brooks, D. J. Edwards, M. D. Harwich Jr, M. C. Rivera, J. M. Fettweis, M. G. Serrano, R. A. Reris, N. U. Sheth, B. Huang, P. Girerd, J. F. Strauss III, K. K. Jefferson and G. A. Buck, "The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies," BMC Microbiology, p. 15:66, 2015.

[15] E. Plummer, J. Twin, D. M. Bulach, S. M. Garland and S. N. Tabrizi, "A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data," Journal of Proteomics & Bioinformatics, vol. 8, pp. 283-291. doi:10.4172/jpb.1000381, 2015.

[16] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla and e. al, "Structure, function and diversity of the healthy human microbiome," Nature , vol. 486, pp. 207-214, 2012.

[17] J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian and M. L. Calle, "Balances: a New Perspective for Microbiome Analysis," mSystems, vol. 3, no. 4, pp. 10.1128/mSystems.00053-18, 2018.

[18] M. Sze and P. Schloss, "Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. mBio. 2016;7(4):e01018-16. doi:10.1128/mBio.01018-16.," mBio, vol. 8, no. 6, pp. e01995-17, 2016.

[19] B. A. Peters, J. A. Shapiro, T. R. Church, G. Miller, C. Trinh-Shevrin, E. Yuen, C. Friedlander, R. B. Hayes and J. Ahn, "A taxonomic signature of obesity in a large study of American adults," Scientific Reports, vol. 8:9749, pp. doi:10.1038/s41598-018-28126-1., 2018.

[20] D. Knights, E. K. Costello and R. Knight, "Supervised classification of human microbiota," FEMS Microbiology Reviews, vol. 35, p. 343–359, 2011.

[21] P. Johnson, L. Vandewater, W. Wilson, P. Maruff, G. Savage, G. Petra, L. Macaulay, K. Ellis, C. Szoeke, R. Martins, C. Rowe, C. Masters, D. Ames and P. Zhang, "Genetic Algorithm with Logistic Regression for Prediction of Progression to Alzheimer's Disease," BMC Bioinformatics , vol. 15, p. S11, 2015.

[22] P. Zhang, B. Verma and K. Kumar, "Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection," Pattern Recognition Letters, vol. 26, no. 7, pp. 909-919, 2003.

[23] M. Khan, A. Mendes, P. Zhang and S. Chalup, "Evolving multi-dimensional wavelet neural networks for classification using Cartesian Genetic Programming," Neurocomputing , vol. 247, pp. 39-58, 2017.

[24] L. Vandewater, V. Brusic, W. Wilson, L. Macaulay and P. Zhang, "An adaptive genetic algorithm for selection of blood-based biomarkers for prediction of Alzheimer's disease progression," BMC Bioinformatics, vol. 16, no. 18, p. S1, 2015.

[25] J. Carter, D. Beck, H. Williams, J. Foster and G. Dozier, "GA-Based Selection of Vaginal Microbiome Features Associated with Bacterial Vaginosis," in Genet Evol Comput Conf., 2014.

[26] B. Callahan, P. McMurdie, M. Rosen, A. Han, A. Johnson and S. Holmes, "DADA2: High-resolution sample inference from Illumina amplicon data," Nature Methods, vol. 13, pp. 581-583, doi: 10.1038/nmeth.3869, 2016.

[27] T. Rognes, T. Flouri, B. Nichols, C. Quince and F. Mahé, "VSEARCH: a versatile open source tool for metagenomics," PeerJ, vol. 4, p. e2584, 2016.

[28] R Core Team , "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/., 2017.

[29] K. V. MARDIA, J. T. KENT and J. M. BIBBY, MULTIVARIATE, Academic Press LImited , 1995.

[30] V. D'Argenio, G. Casaburi, V. Precone and F. Salvatore, "Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines," BioMed research international, p. 2014: 325340, 2014.

[31] P. Zhang, K. Kumar and B. Verma, "A Hybrid Classifier for Mass Classification with Different," Lecture Notes in Aritificial Intelligence, vol. 3614, p. 316 – 319, 2005.

[32] A. Mosca, M. Leclerc and J. P. Hugot, "Gut Microbiota Diversity and Human Diseases: Should We Reintroduce Key Predators in Our Ecosystem?," Frontiers in Microbiology, vol. 7, no. 455, p. doi: 10.3389/fmicb.2016.00455, 2016.

[33] L. Moles, M. Gómez, H. G. H. J. Heilig, G. Bustos, S. Fuentes, W. M. D. Vos, L. Fernández, J. M. Rodríguez and E. Jiménez, "Bacterial Diversity in Meconium of Preterm Neonates and Evolution of Their Fecal Microbiota during the First Month of Life," PLoS ONE, vol. 8, no. 6, p. e66986, 2013.