


RESEARCH

Open Access

Selection of microbial biomarkers with genetic algorithm and principal component analysis



Ping Zhang^{1*} , Nicholas P. West^{1,2}, Pin-Yen Chen¹, Mike W. C. Thang³, Gareth Price³, Allan W. Cripps^{1,4} and Amanda J. Cox²

From 2nd International Workshop on Computational Methods for the Immune System Function
Madrid, Spain. 3-6 December 2018

Abstract

Background: 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 determine whether individuals can be clustered into two classification groups based on pre-determined criteria: control and disease group. However, a combination of other components may exist which better distinguish diseased individuals from healthy controls. Genetic algorithms (GAs) 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 a genetic algorithm (GA) for identifying sets of bacterial species associated with obesity and metabolic syndrome (Mets).

Results: The prediction models built using the combination of principal components (PCs) selected by GA were compared to the models built using the top PCs that explained the most variance in the sample and to models built with selected original variables. The advantages of combining PCA with GA were demonstrated.

Conclusions: The proposed algorithm overcomes the limitation of PCA for data analysis. It offers a new way to build prediction models that may improve the prediction accuracy. 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

Background

Association between the human gut microbiome and a diverse range of health issues has been reported in a number of studies [1, 2]. Knight and colleagues [3] reviewed the methodological approach 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, and metagenomic and metatranscriptomic sequencing. Next-step data analyses are needed to search for overall patterns in microbiome variation. The association between obesity and the gut microbiome from the phylum level to the species level has been studied and various results have been reported [4–6].

Several 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 sequence alignment, operational taxonomic unit (OTU) identification, taxonomy classification, and alpha and beta diversity

* Correspondence: p.zhang@griffith.edu.au

¹Menzies Health Institute QLD, Griffith University, Gold Coast, Australia
Full list of author information is available at the end of the article



calculation. They have been widely used for different biological and medical research purposes, such as associating gut microbiome diversity with diseases [10–13]. 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 produce different 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 produced comparable results for analysis of faecal samples, in terms of alpha diversity 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. D'Argenio et al. [16] also compared QIIME and MG-RAST, and observed a statistically significant difference between these two bioinformatics pipelines with regards to beta diversity measures.

Despite the effort from researchers to develop high quality analytical pipelines, it is recognised that the complexity and variability of the human microbiome can be sensitive to various environmental factors [17]. Improvement of analytical pipelines has been complicated by the limitation of available sample material and the relatively high cost of the sequence analysis necessary for microbiome profiling. As a result, most microbiome studies have used limited sample sizes, raising questions regarding the accuracy of their findings. In addition to efforts to improve the accuracy of OTU detection and taxonomic 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 started looking at defining microbial signatures for specific diseases. This includes the application of computational modelling and variable selection techniques. For example, Rivera-Pinto et al. [18] presented a greedy step-wise algorithm for selection of microbial signatures that preserves the principles of compositional data analysis. Sze and Schloss [19] 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 inter-personal variation and insufficient sample sizes. The same study also tested random forest models for classifying individuals as obese on the basis of microbiome composition and did not find obvious patterns that could separate the obese and healthy groups. Random forest models were also used by Peters et al. [20] to identify taxonomic signatures of obesity. These models were evaluated with Receiver Operator Characteristic (ROC) curves and the area under the curve (AUC) value produced by the optimal model, which included 49 OTUs, was 0.81. When the repeated cross-validation was performed, the AUC value decreased to 0.65. Other machine learning methods used for microbiome studies have been reviewed by Knights et al. [21].

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. The first two or three principle components account for the greatest proportion of the variance in the dataset. Usually, these components are then used to determine whether individuals can be clustered into one of two classification groups, control or diseased, based on pre-determined criteria. 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 our obesity study, to identify potential sets of bacterial species that may be associated with obesity with metabolic syndromes (MetS). The preliminary work has been presented in the 2018 IEEE International Conference on Bioinformatics and Biomedicine [22].

Methods

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 PCs, particularly in datasets that are noisy or have strongly correlated explanatory variables. The PCs can then be used to build predictive models. The PCs are the linear combinations

of the original variables that account for variance in the data. PCA can be performed using either 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 of 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 unit of measurement before PCA is used.

Genetic algorithms

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

GAs can be useful and efficient when searching for the best combination of variables to achieve the best outcome (e.g. accuracy of prediction). GAs have been developed and applied for biomarker profile identification in a range of settings such as Alzheimer's disease progression and breast cancer diagnosis [23–25]. The GAs have also been modified and improved to adapt to different computational environments and for different applications [26, 27]. Carter et al. [28] applied GA to their study to select vaginal microbiome features associated with bacterial vaginosis. 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 will be used to find the best subset of principal components produced from a PCA using gut microbial species data.

Proposed method

The method described 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" [29] in R

(<https://cran.r-project.org/>). GA is then applied for selection of the set of components created from the PCA that best predict individuals as obese or healthy weight. The scores of selected PCs calculated for each individual are used as the input for building a classification model. ROC curve analysis is used to evaluate 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 [24].

Experiments and results

In this study, faecal samples from 22 obese and 105 healthy-weight subjects were collected and sequenced using a 16S-based approach. The obese sample here was designed as those with body mass index (BMI) over 30 and with MetS [10]. The healthy-weight subjects included 39 recreational individuals and 66 athletes who were involved in rugby, football soccer, judo, rowing, triathlon or weightlifting. For sequencing analysis, paired-end reads were merged using the PEAR software (v0.9.6) [30]. Contaminant human reads were removed by mapping to the hg19 human genome using BWA software package (v0.7.12) [31] and the remaining reads were searched against the Greengenes 16S taxonomy database (GG v13.5) [32] using sequence analysis tool VSEARCH (v1.9.7) [33] to generate a single OTU raw count/abundance table for all 127 subjects. Amongst the 127 subjects 68,590 OTUs were identified (at all taxonomic levels), which mapped at the level of species to 163 observations, from Greengenes total reportable content of 3093 species. Species with low diversity across the cohort were filtered from future analysis, this was achieved by removing the species with zero abundance in 80% of both healthy and obese subjects. This excluded 126 species (77.3%) of the data leaving 37 species for further analysis. The abundance values of each of these species were normalized to the range of [0, 1] (highest abundance across the individuals as 1 and the lowest as 0) before applying the proposed method which combines PCA and GA for identifying obese from healthy subjects. The results were compared with those produced without GA and with those produced by using GA to select

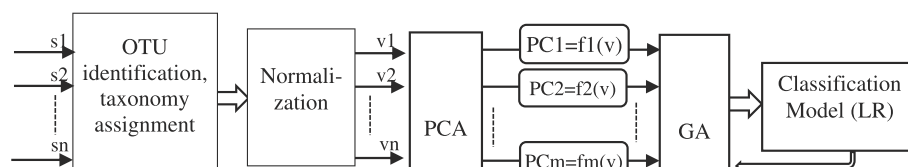


Fig. 1 A diagram of proposed method. s_1, s_2, \dots, s_n are the 16S rRNA sequences for this study (can be from other sequencing). v_1 to v_n are the abundance (normalized) of each species detected in each individual. m = number of PCs created by PCA, n = number of individuals included in the sample. 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

combinations of bacterial species for heathy and obese classification without PCA in the model.

GA models to select combination of PCs for classification

For experiments, we performed PCA across three circumstances using: the whole dataset, obese only sample, and healthy weight only sample. This approach is based on the possibility that for different populations, the correlation between species might differ. The function “prcomp” from the “stats” package in R [34] 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 identifying obese from healthy subjects. The algorithm used by “prcomp” for creating the PCs can be found in reference [35]. 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 the cross-validated AUC value created from the logistic regression model. More explanation about AUC can be found in Johnson et al. [24]. Constraints for GA were set to include 1 to 6 PCs in the classification model. Ten-times repeated five-fold cross-validation was used for testing the classification model with selected PCs. With each data set (all, healthy or obese), 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 results (Table 1) it can be seen that the selection from GA was quite consistent with slight variation from each run.

The PCA constructed from the whole data set and healthy-weight subjects both created 37 principal components (PC1 to PC37) while the PCA from obese subjects created 22 components (PC1 to PC22). Table 1

lists 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 in the corresponding classification model. Positive coefficient means that an 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 with MetS.

Table 2 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- within column Comp1 represents that Prausnitzii has negative correlation with Comp1 (PC1 for Whole, PC14 for obese, and PC1 for Healthy). That suggests that increased Prausnitzii abundance will decrease the Comp1 value. As Comp1 has a positive correlation with being overweight, it can be speculated that increased Prausnitzii abundance leads to decreased likelihood of being obese.

From the results presented in Table 1 and Table 2 each of the species were analysed and categorized into two groups; positive (indicated with an asterisk (*) in Table 2) or negative correlations with the 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 results from the first set of the experiments which ran PCA on the whole dataset, 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 a

Table 1 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+ (or PC2-)	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 2 Top species included in the GA selected 1, 2, 3, 4, 5 or 6 PCs produced with different data sets

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

Comp1, Comp2, Comp3, Comp4, Comp5 and Comp6 represent the 6 PCs selected by GA. For experiment with whole dataset they are PC1, PC7, PC2, PC27, PC11 and PC15 respectively; for experiment with obesity sample, they are PC14, PC18, PC2, PC4, PC19 and PC16; for experiment with healthy sample, they are PC1, PC34, PC23, PC28, PC3 and PC5

^aSpecies has a positive correlation with the probability of having healthy body mass

+ Positive correlation with the corresponding PC

- Negative correlation with the corresponding PC

potential benefit on health. On the other hand, high values for Gnavus, Catus, Moorei and Aerofaciens together are associated with high probability with of being obese.

A final classification model was built with each set of PCs selected by GA or first 1 to 6 PCs (which explain

the most variance of the data) from the PCA. Again, the PCs were calculated from the whole dataset, healthy-weight dataset or obese dataset. The AUCs produced from the GA-selected PCs were quite obviously higher than the ones from the top PCs of PCA. Figure 2 shows

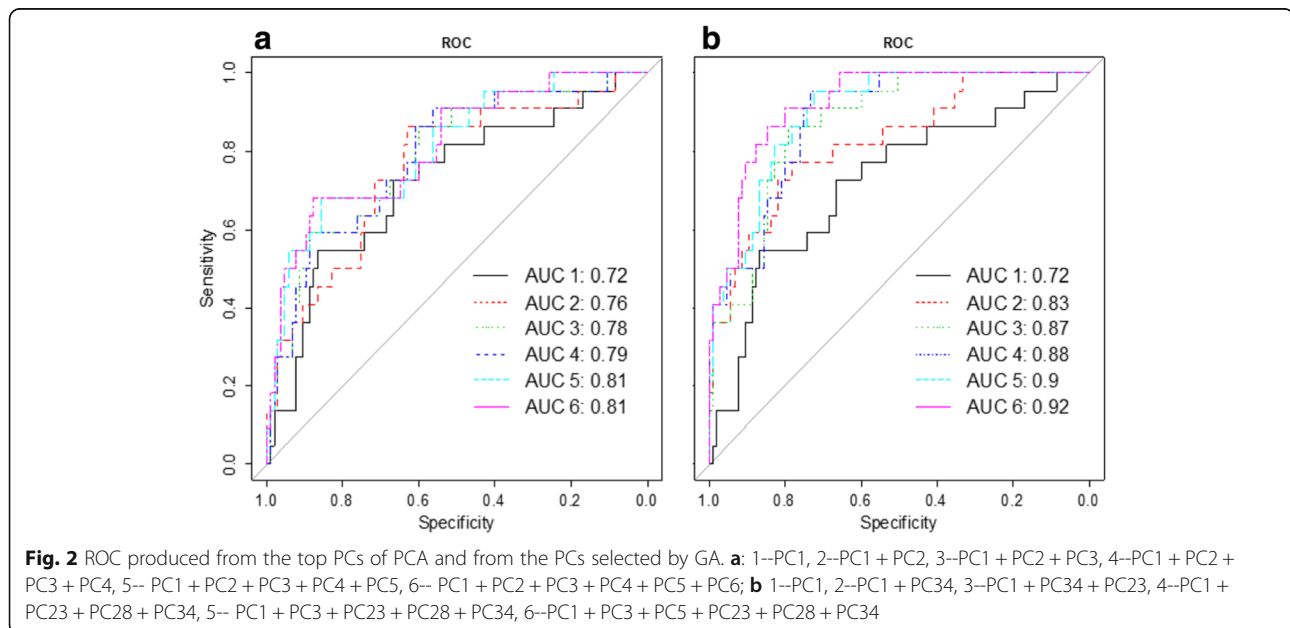


Fig. 2 ROC produced from the top PCs of PCA and from the PCs selected by GA. **a**: 1--PC1, 2--PC1 + PC2, 3--PC1 + PC2 + PC3, 4--PC1 + PC2 + PC3 + PC4, 5-- PC1 + PC2 + PC3 + PC4 + PC5, 6-- PC1 + PC2 + PC3 + PC4 + PC5 + PC6; **b** 1--PC1, 2--PC1 + PC34, 3--PC1 + PC34 + PC23, 4--PC1 + PC23 + PC28 + PC34, 5-- PC1 + PC3 + PC23 + PC28 + PC34, 6--PC1 + PC3 + PC5 + PC23 + PC28 + PC34

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-weight dataset (when compared with the result from the PCs calculated from whole dataset and obese dataset, the first PCs from the healthy data produced the highest AUC values).

GA models to select sets of species for classification

To compare the results from the model that combined PCA and GA and from the model where GA was applied directly for selection of the combination of the bacterial species, GA was implemented in conjunction with logistic regression using the species abundance directly as the input for classification.

For experiments, the number of species (number of input variables for logistic regression) was restricted to maximum six, which was the same as the maximum number of PCs used in the earlier experiments. Table 3 shows the combinations of the bacterial species selected by 100 repeated runs of GA, which achieved the highest AUC values. It can be seen that some of the species were commonly selected in different sets of the selections. The selection frequency of each species from the 100 repeated GA runs was calculated and a frequency chart showing the top 10 most selected species was drawn in Fig. 3. Eutactus and Gnavus appeared in the final selection of almost every run of the GA (96 out of 100 runs and 95 out of 100 runs). Muciniphila, Distasonis and Prausnitzii were also selected frequently (> 50% frequency) in the repeated GA runs. These highly selected bacterial species appeared to have relatively high contribution to the selected PCs shown in the previous section.

Discussion

In this study, a computational method that combines PCA and GA has been proposed to produce accurate prediction result and to find sets of features (variables) that contribute the most to the prediction models. The model was applied to identify sets of bacterial species associated with high body mass. Due to the high correlation between many species of the gut bacteria,

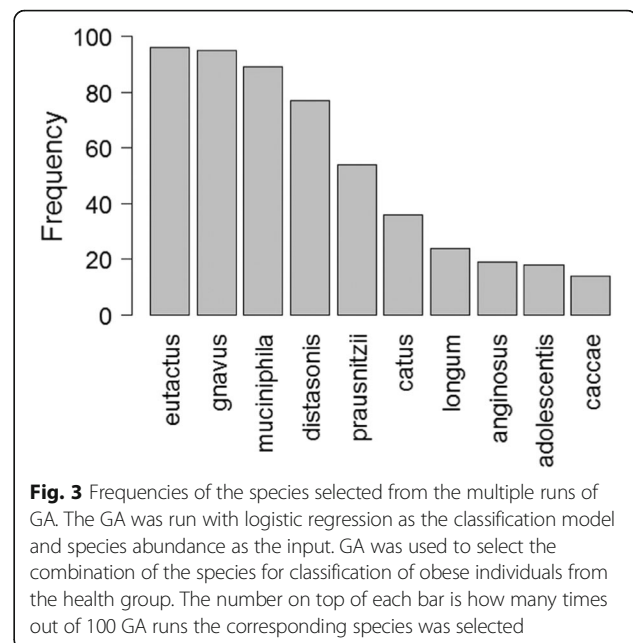


Fig. 3 Frequencies of the species selected from the multiple runs of GA. The GA was run with logistic regression as the classification model and species abundance as the input. GA was used to select the combination of the species for classification of obese individuals from the health group. The number on top of each bar is how many times out of 100 GA runs the corresponding species was selected

constructing PCA before the GA can improve the efficacy of GA for selecting multiple sets of microbial species associated with obesity and MetS. 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. The results were also compared with those produced from the GA selected models with bacterial species abundance values as the input directly, and it showed its advantages.

In the microbiome study, the results produced from the described method depends on the accuracy of the sequencing analysis. The microbial species identified here was based on the sample of 22 obese subjects and 105 healthy-weight subjects. Assuming this result was validated in multiple datasets with bigger sample sizes, the results from Table 1 and Table 2 can suggest a few combinations of microbial species groups that are

Table 3 Sets of species selected by GA using the species abundance as the input variables of logistic regression models

GA Selected Species						AUC
Adolescentis	Catus	Eutactus	Gnavus	Muciniphila	Prausnitzii	0.87
Adolescentis	Distasonis	Eutactus	Gnavus	Muciniphila	Prausnitzii	0.87
Aerofaciens	Distasonis	Eutactus	Gnavus	Longum	Muciniphila	0.88
Anginosus	Distasonis	Eutactus	Gnavus	Muciniphila	Prausnitzii	0.87
Catus	Distasonis	Eutactus	Gnavus	Muciniphila	Prausnitzii	0.88
Catus	Eutactus	Gnavus	Longum	Muciniphila	Prausnitzii	0.86
Distasonis	Eutactus	Gnavus	Longum	Muciniphila	Prausnitzii	0.88

GA Selected Species lists the set of species selected by GA, each row one set. AUC is the area under the ROC curve produced by the corresponding logistic regression model with the selected set of species. The result was cross validated with the same cross validation set up as the earlier experiments

beneficial to health. Some of the species in the combinations can be replaced by equivalent alternative species that are suggested by the algorithm, which gives flexibility for further intervention. 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 also be used for taxonomy classification of the OTUs identified.

Conclusion

This study demonstrated 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 options of combining the features included 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 a study comparing obese individuals with MetS and healthy-weight subjects. It can be applied for any other classification or biomarker identification study. The 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.

Abbreviations

AUC: Area under the ROC; BMI: Body mass index; GA: Genetic algorithm; MetS: Obesity with metabolic syndromes; OTU: Operational taxonomic unit; PC: Principal component; PCA: Principal component analysis; ROC: Receiver operator characteristic curve

Acknowledgements

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. Part of this work has previously been presented at conference (IEEE International Conference on Bioinformatics and Biomedicine (BIBM)).

About this supplement

This article has been published as part of BMC Bioinformatics Volume 20 Supplement 6, 2019: Towards computational modeling on immune system function. The full contents of the supplement are available online at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-6>.

Authors' contributions

PZ designed the methodology, performed the data analysis and computational experiments, and drafted the paper. MT and GP contributed to the 16S sequencing analysis and OTU identification. AWC, NPW and AJC designed the obesity study and contributed to participant recruitment and data collection. PY assisted for data interpretation and revised the draft of the manuscript. All authors read the paper, made comments, and agreed with the content. All authors read and approved the final manuscript.

Funding

This project was supported by Griffith Health Institute/Gold Coast Hospital Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of Griffith Health Institute/Gold Coast Hospital Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Publication costs are funded by Menzies Health Institute QLD, Griffith University, Australia.

Availability of data and materials

Data are available upon request from the Menzies Health Institute Queensland for researchers who meet the criteria for access to confidential data.

Ethics approval and consent to participate

This study was approved by Bond University Human Research Ethics Committee (0000015530) and Griffith University Human Research Ethics Committee (GU 2016/213 and GU 2015/229). Consent in writing was obtained from all participants who provided samples.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Menzies Health Institute QLD, Griffith University, Gold Coast, Australia. ²School of Medical Science, Griffith University, Gold Coast, Australia. ³QFAB Bioinformatics, Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia. ⁴School of Medicine, Griffith University, Gold Coast, Australia.

Received: 12 June 2019 Accepted: 18 July 2019

Published: 12 December 2019

References

- Jackson MA, Verdi S, Maxan ME, et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat Commun.* 2018;9(1):2655.
- Gilbert JA, Blaser MJ, Caporaso JG, et al. Current understanding of the human microbiome. *Nat Med.* 2018;24:392–400.
- Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. *Nat Rev Microbiol.* 2018;16(7):410–22.
- Ottosson F, Brunkwall L, Ericson U, et al. Connection between BMI-related plasma metabolite profile and gut microbiota. *J Clin Endocrinol Metab.* 2018;103(4):1491–501.
- Million M, Lagier JC, Yahav D, et al. Gut bacterial microbiota and obesity. *Clin Microbiol Infect.* 2013;19(4):305–13.
- Chakraborti CK. New-found link between microbiota and obesity. *World J Gastrointest Pathophysiol.* 2015;6(4):110–9.
- Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods.* 2010. <https://doi.org/10.1038/nmeth.f.303>.
- Keegan KP, Glass EM, Meyer F. MG-RAST, a metagenomics Service for Analysis of microbial community structure and function. *Methods Mol Biol.* 2016;1399:207–33. https://doi.org/10.1007/978-1-4939-3369-3_13.
- Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol.* 2009;75(23):7537–41.

10. Han GG, Lee JY, Jin JD, et al. Evaluating the association between body weight and the intestinal microbiota of weaned piglets via 16S rRNA sequencing. *Vet Microbiol.* 2016;196:55–62.
11. Clemente J, Ursell L, Parfrey L, et al. The impact of the gut microbiota on human health: an integrative view. *Cell.* 2012;148(6):1258–70.
12. Spencer M, Hamp T, Reid R, et al. Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology.* 2011;140(3):976–86. <https://doi.org/10.1053/j.gastro.2010.11.049>.
13. Zhong L, Shanahan ER, Raj A, et al. Dyspepsia and the microbiome: time to focus on the small intestine. *Gut.* 2016. <https://doi.org/10.1136/gutjnl-2016-312574>.
14. Brooks JP, Edwards DJ, Harwich MD, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* 2015;15:66. <https://doi.org/10.1186/s12866-015-0351-6>.
15. Plummer E, Twin J, Bulach DM, et al. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *J Proteomics Bioinformatics.* 2015;8:283–91. <https://doi.org/10.4172/jpb.1000381>.
16. D'Argenio V, Casaburi G, Precone V, et al. Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. *Biomed Res Int.* 2014;2014:325340. <https://doi.org/10.1155/2014/325340>.
17. Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012;486:207–14.
18. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, et al. Balances: a new perspective for microbiome analysis. *mSystems.* 2018;3(4). <https://doi.org/10.1128/mSystems.00053-18>.
19. Sze M, Schloss P. Looking for a signal in the noise: revisiting obesity and the microbiome. *mBio.* 2016;7(4):e01018-16. <https://doi.org/10.1128/mBio.01018-16>.
20. Peters BA, Shapiro JA, Church TR, et al. A taxonomic signature of obesity in a large study of American adults. *Sci Rep.* 2018;8:9749. <https://doi.org/10.1038/s41598-018-28126-1>.
21. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. *FEMS Microbiol Rev.* 2011;35:343–59.
22. Zhang P, West N, Chen P, Cripps A, Cox A. Combination of principal component analysis and genetic algorithm for microbial biomarker identification in obesity. Madrid: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018.
23. Zhang P, Verma B, Kumar K. Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection. *Pattern Recogn Lett.* 2003; 26(7):909–19.
24. Johnson P, Vandewater L, Wilson L, et al. Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. *BMC Bioinformatics.* 2015;15:S11.
25. Zhang P, Kumar K, Verma B. A hybrid classifier for mass classification with different kinds of features in mammography. *LNCS.* 2005;3614:316–9.
26. Khan M, Mendes A, Zhang P, et al. Evolving multi-dimensional wavelet neural networks for classification using Cartesian genetic programming. *Neurocomputing.* 2017;247:39–58.
27. Vandewater L, Brusci V, Wilson W, et al. An adaptive genetic algorithm for selection of blood-based biomarkers for prediction of Alzheimer's disease progression. *BMC Bioinformatics.* 2015;16(18):S1.
28. Carter J, Beck D, Williams H, et al. GA-based selection of vaginal microbiome features associated with bacterial vaginosis. *Genet Evol Comput Conf.* 2014; 2014:265–8.
29. Callahan B, McMurdie P, Rosen M, et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016;13:581–3. <https://doi.org/10.1038/nmeth.3869>.
30. Zhang J, Kobert K, Flouri T, et al. PEAR: a fast and accurate Illumina paired-end reAdmerger. *Bioinformatics.* 2014;30:614–20.
31. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics.* 2009;25:1754–60.
32. DeSantis T, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol.* 2006;72(7):5069–72.
33. Rognes T, Flouri T, Nichols B, et al. VSEARCH: a versatile open source tool for metagenomics. *Peer J.* 2016;4:e2584.
34. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for statistical computing; 2018. URL <https://www.R-project.org/> (Accessed on 20 Jul 2018)
35. Mardia KV, Kent JT, Bibby JM. *Multivariate analysis.* London: Academic; 1979.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

