Unraveling city-specific signature and identifying sample origin locations for the data from CAMDA MetaSUB challenge

Runzhi Zhang
University of Florida https://orcid.org/0000-0002-0763-5928
Alejandro R. Walker
University of Florida
Susmita Datta (susmita.datta@ufl.edu)
University of Florida https://orcid.org/0000-0002-7408-699X

Research

Keywords: Microbiome, OTU, WGS, feature selection, machine learning, Random Forest, Support Vector Machine, Linear Discriminant Analysis, PCoA, ANCOM

Posted Date: January 13th, 2020

DOI: https://doi.org/10.21203/rs.2.20675/v1

Version of Record: A version of this preprint was published on January 4th, 2021. See the published version at https://doi.org/10.1186/s13062-020-00284-1.

Abstract

Background

Composition of microbial communities can be location specific, and the different abundance of taxon within location could help us to unravel city-specific signature and predict the sample origin locations accurately. In this study, the whole genome shotgun (WGS) metagenomics data from samples across 16 cities around the world and samples from another 8 cities were provided as the main and mystery datasets respectively as the part of the CAMDA 2019 MetaSUB “Forensic Challenge”. The feature selection, normalization, three methods of machine learning, PCoA (Principal Coordinates Analysis) and ANCOM (Analysis of composition of microbiomes) were conducted for both the main and mystery datasets.

Results

Feature selection, combined with the machines learning methods, revealed that the combination of the common features was effective for predicting the origin of the samples. The average error rates of 11.6% and 30.0% of three machine learning methods were obtained for main and mystery datasets respectively. Using the samples from main dataset to predict the labels of samples from mystery dataset, nearly 89.98% of the test samples could be correctly labeled as “mystery” samples. PCoA showed that nearly 60% of the total variability of the data could be explained by the first two PCoA axes. Although many cities overlapped, the separation of some cities was found in PCoA. The results of ANCOM, combined with importance score from the Random Forest, indicated that the common “family”, “order” of the main-dataset and the common “order” of the mystery dataset provided the most efficient information for prediction respectively.

Conclusions

The results of the classification suggested that the composition of the microbiomes was distinctive across the cities, which was also supported by the results from ANCOM and importance score from the RF. The analysis utilized in this study can be of great help in field of forensic science to efficiently predict the origin of the samples. And the accurate of the prediction could be improved by more samples and better sequencing depth.

Background

The advent of next generation sequencing (NGS) technologies for metagenomics has experienced a tremendous improvement, which allows the generation of large sequence datasets derived from diverse ecosystems, such as the human body, soil, and ocean water [1]. The use of whole genome sequencing (WGS) has been reported to have multiple advantages when compared with the 16S rRNA amplicon data [2]. As the composition of microbial communities can be location specific [3], studying the microbiome

from different cities improves our understanding of city-specific microbes and their contributions to ecosystem composition and diversity.

In this work, we aimed to unravel city-specific signature and find the appropriate features for identifying and predicting the origin location of samples from different areas. The dataset was provided by MetaSUB (http://camda2019.bioinf.jku.at/doku.php/contest_dataset), which aimed to build an international metagenomic map of urban spaces, based on extensive sampling of mass-transit system and other public areas around the world. They partnered with CAMDA for an early release of microbiome data obtained from global City Sampling Days, comprising the WGS metagenomics data. The main dataset covered 16 cities across the globe, with tens of samples per city. Moreover, one more dataset with 8 cities was provided as mystery set from the CAMDA 2019 MetaSUB challenge to serve as testing samples. And the true city-information of the mystery data was provided much later in the process. **Table 1** presented a tabulated insight of the data for all the cities.

Results

Feature selection

First, we selected the common “species”, “family”, and “order” existing across all the 16 cities and the number of variables was 7, 9 and 9 respectively. The number of features was limited and more information about the microbes was needed. In order to increase the number of features, we selected the features based on additional rules: a) The features were selected based on the ubiquity of the “species”, “family” and “order” across all the cities. b) The features were selected based on the ubiquity of the “species”, “family” and “order” across all the samples. c) The combinations of the common “species”, “family”, and “order” were regarded as the combined features. **Table 2** presented the details of the features selected based on additional rules. For simplicity, the mystery dataset was analyzed based on the common features and combined features. After feature selection, the aggregated raw counts of each dataset were normalized to generate log2-cpm for further analyses.

Machine learning analysis

For the main and mystery datasets respectively, results from Random forest (RF) [4], Support Vector Machine (SVM) [5] and Linear Discriminant Analysis (LDA) [6] were obtained with the leave-one-out cross validation (CV), one test sample was randomly selected in each run with 1000 runs repeated. **Table 3** presented the details of the classification error rate based on different rules using the main and mystery datasets respectively.

As seen in the table, when the features existing in at least N cities were selected for the analysis, the changing trends of the error rate of RF-species (qualitied “species” was used for analysis using RF), SVM-species, LDA-species, LDA-family and LDA-order shared a similar pattern, a decreased CV error rate was obtained when increasing the number of features, and then the lowest error rate was achieved. For RF-family and SVM-family, the error rate hasn’t changed considerably when we increased the number of the

variables at the family rank. Therefore, using the common “family” was better, as we obtained the low error rate without including too many features. And for RF-order and SVM-order, the lowest error rate was obtained using common “order”. Additionally, when top M features with the highest ubiquity across all the samples were selected for analysis, error rates decreased with the increasing number of features used and then the lowest error rate was achieved, no matter which machine learning methods or which kinds of feature we used for analysis. In addition, for the combined features, the best performance was achieved using the features combined with common “species”, “family” and “order” (7 species, 9 families, 9 orders), the error rates obtained from RF, SVM and LDA were 11.6%, 11.5% and 11.8% respectively. And for the mystery dataset, we obtained the lowest average error rate when we used the combined features with 8 species and 15 orders. The error rates were 29.7%, 32.1% and 28.2% for RF, SVM and LDA respectively.

The error rates of predictions for each city was presented in **Figure 1**. From **Figure 1.A**, the low error rates were obtained from the cities with better sequencing depth (**Table S1**) including Bogota, Ilorin, New York, Offa, Sacramento and Tokyo. The average error rates of the three methods for these four cities were 7.84%, 5.49%, 2.03%, 6.31%, 3.08% and 5.80% respectively. However, the error rates of Auckland and Hamilton, two cities also with good sequencing depth, were 26.39% and 18.44% respectively. By looking into the details of the results, we found that the samples AKL_1, AKL_7, AKL_14 and HAM_7, HAM_12 cannot be predicted correctly by all the three methods. In other words, the microbial composition of these samples was different from the other samples in Auckland and Hamilton (**Table S2**), making them difficult to be identified. And we found that the samples from London with the poor sequencing depth showed low error rate. Upon finding excessive zeros in samples from London, the samples of London could be easily identified from all the samples, which resulted in the low error rate of London. In addition, a possible explanation for low error rate could be the insufficient number of samples, as the cities with the lowest number of samples including Sofia and Marseille showed high error rates. And from **Figure 1.B**, the cities with deep sequencing (**Table S3**), i.e. Oslo and Rio de Janeiro, were the two of the first three cities with the lowest average error rate (6.64% and 16.81% respectively). Similarly, the cities with poor sequencing depth such as Doha and Brisbane, showed the high average error rate (85.71% and 62.61% respectively). And Doha, also with the limited number of samples, achieved the highest error rate among all the cities. The evident from the **Figure 1** that most cities with limited samples and poor sequencing depth had high error rates, indicating that sufficient samples and deep sequencing were necessary for successfully predicting the provenance of samples.

In addition to the analyses based on the main and mystery datasets respectively, we have also used the prediction models built based on the main dataset to predict the samples from mystery dataset. As the main dataset and mystery dataset had no cities in common, the information of cities from mystery dataset was lacked in the main dataset. Therefore, 50% of the samples of each city from mystery dataset were randomly sampled and added to the main dataset to serve as the part of the training dataset, and the rest of the samples from mystery dataset were served as the test samples. In the training dataset, the samples from mystery dataset were labeled as the “mystery”. The prediction models were built based on the common “family” (5 families) and “order” (6 orders), as there was no common “species” between the

main and mystery datasets. Three different classifications were used and the random samplings were conducted for 1000 times. For each run, the predicted labels were checked with the real labels. The average error rates for RF, SVM and LDA were 10.48%, 9.21% and 10.36% respectively, indicating that the prediction models could effectively identify the mystery samples from all the samples.

The following analyses were based on the features which achieved the lowest average error rate.

Principal Coordinates Analysis

The results of PCoA [7] in **Figure 2** presented the bi-plots for both datasets. **Figure 2.A** illustrated the main dataset and the 58.4% of total variability of the data could be explained by the first two PCoA axes. A separation of the cities could be referred from the plot. For example, London was separated from most cities and on the rightmost site, which was corresponding to the result from machine learning methods, indicating that the excessive zeros in samples from London made the prediction easier. However, many cities overlapped together. Specifically, Ilorin and Offa, both the cities of Nigeria, whose ellipses showed a massive overlap. Also, Auckland and Hamilton, both being in New Zealand, overlapped with each other. The result of mystery dataset was given in **Figure 2.B**. The first two PCoA axes explained 65.4% of total variability in the data, which was comparable with the percentage explained in the main dataset. Although many cities overlapped, samples of Oslo were clustered together and distributed at the top of the plot, separating from the most samples. Also, some of the samples from Rio de Janeiro were far away from the most cities, which made these samples easier to be identified. These results were corresponding to the low error rates of Oslo and Rio de Janeiro in the previous section.

Analysis of composition of microbiomes

The results from the analysis of composition of microbiomes (ANCOM) [8] were presented in **Figure 3**. The relative abundances of the features were used to conduct the pair-wise comparisons among all the cities. Upon the significance of the features, the differentially abundant features were found. The features, which served as the predictors, on the right were ordered by the number of times the relative abundance was significantly different in the pair-wise comparisons. As presented in **Figure 3.A**, the top 10 features, i.e. *Bacillaceae*, *Bacillales*, *Actinomycetales*, *Sphingomonadaceae*, *Pseudomonadaceae*, *Pseudomonas.spp*, *Sphingomonadales*, *Lactobacillales*, *streptococcaceae* and *Enterobacteriaceae* showed the highest counts. For the top feature, the count of *Bacillaceae* is 43 out of 120, which meant that in the 120 pair-wise comparisons among the 16 cities, *Bacillaceae* was found to be significantly different in 43 comparisons. Similarly, in **Figure 3.B**, the top 10 features were *Bacillales*, *Clostridiales*, *Pseudomonadales*, *Staphylococcus.epidermidis*, *Lactobacillales*, *Rhodospirillales*, *Flavobacteriales*, *Streptophyta*, *Burkholderiales* and *Enterobacteriales*. Additionally, little change was seen from the importance score (**Figure 4**) derived from the RF. It could be inferred from **Figure 4.A** that *Bacillales, Actinomycetales*, *Sphingomonadales*, *Sphingomonadaceae*, *Enterobacteriaceae* and *streptococcaceae* were also in top 10 features. Also, for the mystery dataset, *Bacillales*, *Streptophyta, Rhodospirillales*, *Enterobacteriales*, *Lactobacillales*, *Clostridiales* and *Pseudomonadales* were in the top 10 of both lists. In summary, for the main dataset, the common “family” and “order” of the OTU provided the most

informative data for predicting the origins of the samples, which was also corresponding to the low error rate of three methods when using the common “family” and “order” only compared with using the common “species”. For the mystery dataset, the results of the ANCOM and feature importance combined with the error rates from three methods indicated that the common “order” dominated the prediction.

Discussion And Conclusions

For the CAMDA challenge MetaSUB data of this year, 16 cities were included and 10 or more samples were collected for each city in the main dataset. The feature selection, normalization, three methods of machine learning algorithms, PCoA and ANCOM were conducted for both the main and mystery datasets. Combining the feature selection and the results of machine learning methods, we found that the machine learning analysis was effective in predicting the origin of the samples when the combined common features were used, indicating that the combination of the common features could be the effective microbial fingerprint for unraveling city-specific signature and identifying sample origin locations, which proved to be with great potential for forensic science. Therefore, more taxonomic ranks of microbiomes such as class and genus could be added to the combination to investigate the performance of the prediction in the future works. As in this study, by combining the common features, we obtained the low error rate of prediction without including too many features for both datasets. Additionally, we have used the main dataset and half of mystery samples as our training dataset to predict the labels of another half of mystery samples. As the error rates were 10.48%, 9.21% and 10.36% respectively for RF, SVM and LDA, most of the samples from mystery dataset could be identified correctly by the classification. PCoA analysis for both datasets showed that nearly 60% of the total variability of the data could be explained by the first two PCoA axes, most cities overlapped with each other. However, some cities such as London and Oslo were separated from most cities, indicating the unique composition of microbiomes in these cities, which was also validated by ANCOM analysis and was corresponding to the low error rates in machine learning analysis. The heatmaps of ANCOM revealed that some of the features, such as some of the common “family” and “order” in the main dataset and the common “order” in mystery dataset, were significantly different in pair-wise comparisons, and these features were also given high importance score in RF, indicating the effectiveness of these features during the classification. Additionally, this was also supported by the results of the machine learning analysis, we obtained a lower error rate using common “family” and “order” and using common “order” only for the main and mystery datasets respectively.

However, due to the limited resources such as the insufficient number of samples from some cities and the poor sequencing depth of the samples, the error rate of Sofia was higher compared with other cities. Therefore, sufficient samples and better sequencing depth were necessary in order to improve the accuracy of the prediction. A similar problem was found in the mystery dataset, the number of samples was limited for most of the cities in the mystery dataset, which resulted in an overall higher error rate in the mystery dataset compared with the main dataset. And the poor sequencing of Brisbane and Doha also resulted in higher error rates compared to other cities in the mystery dataset. Additionally, the microbial composition could vary across samples within the same city, which also limited the ability of classification to correctly predict the origins of the samples. For this study, only the combined common

features were selected for further analysis, some city-specific microbiomes might be ignored in this study, as they only existed in a specific city. These city-specific microbiomes could be combined with the common features to provide more powerful prediction in the future work.

Methods

The design of the analysis was motivated by the experience from the CAMDA 2017 and CAMDA 2018 MetaSUB Challenges [9, 10]. Compared to date from previous MeteSUB challenges, the data this year was with higher quality and deeper sequencing depth. As we have more cities included this year, the common features shared by all the cities were further limited. Therefore, the features selection was implemented this year to help us obtain the appropriate features for prediction. Finally, unsupervised and supervised techniques were used for the analyses. A more detailed description of the implementations was supplemented in the following sections.

Bioinformatics and data preparation

The Bioinformatics and data preparation was based on our previous paper [10]. Samples were Illumina-sequenced at different depths and delivered as FASTQ format for further analysis. Subsequent bioinformatics processing and data preparation were conducted in the “HyperGator2” high-performance cluster at the University of Florida. After quality control, the data was picked for OTUs in open-reference mode with QIIME [11]. After quality control, samples with the poor sequencing depth were removed. OTUs were aggregated as counts, and all further data processing and analysis in this study were conducted in R [12].

Feature selection

Feature selection was conducted based on different rules. For the main dataset, the common “species”, “family” and “order” existing across all the 16 cities were selected at first. And three additional rules were conducted for feature selection: a) “species”, “family” and “order” existing in at least N cities were selected. N was set to 15, 14, 13, 12, 11, 10, 9 and 8 respectively, the count of the feature that did not exist in the city was marked as zero. b) “species”, “family” and “order” were reordered based on their ubiquity across all the samples respectively, the top M features with the highest ubiquity were selected. M was set to 10, 20, 30, 50, 100, 150 respectively. c) The combinations of the common “species”, “family”, and “order” were regarded as the combined features. The mystery dataset was analyzed based on the common features and combined features. The data were then normalized to generate log2-cpm to make the data meaningful using the function “voom" [13] in R package “limma” [14] for further analysis.

Machine learning analysis

Three classification algorithms were implemented at this stage for both the main and mystery datasets: Random Forest (RF) [4], Support Vector Machine (SVM) [5] and Linear Discriminant Analysis (LDA) [15]. When the analysis was conducted for the main and mystery datasets respectively, each method was

implemented 1000 times with the leave-one-out cross validation based on the selected features. For each run, one sample was randomly selected as the test data with other samples served as the training set. RF was conducted using the R package “randomForest”, 1000 trees were used and the count of variables chosen at each split was equivalent to the square root of the number of features in the dataset. The model was fitted based on the training set and then used to predict the origin of the test sample. The result of each prediction was obtained. The overall error rate and the error rate for each city were calculated based on the result of the 1000 runs. The variable importance score [16] computed by the RF was also recorded. The SVM classifier was implemented in a similar manner. The two important parameters in SVM, i.e. gamma and c-value, affected the fitting of the models. Using the R function “best.svm” in package “e1071” [17], the SVM model with the appropriate parameters was obtained by testing the performance of models with different parameters. The LDA was conducted in a similar manner using the R package “MASS” [18].

When the models based on the main dataset were used to predict the labels of samples from mystery dataset. 50% of the samples of each city were randomly sampled from mystery dataset and added to the main dataset with the labels of “mystery” to serve as the part of the training dataset, the rest of the samples were served as the test samples. The sampling and the prediction were implemented for 1000 times. For each run, the predicted labels of test samples were checked with the real labels, a good prediction meant all test samples could be labeled as “mystery” by the classification.

Principal Coordinates analysis

Principal coordinates analysis (PCoA) [7] of normalized data was conducted using the R package “vegan” [19] and “ape” [20]. Firstly, the dissimilarity matrix was constructed based on Bray-Cruits distance. Then, the dissimilarity matrix was used for PCoA and a set of uncorrelated axes was generated to summarise the variability in the dataset. The two-dimensional plot was generated for assessing the separation of the cities.

Analysis of composition of microbiomes

Analysis of composition of microbiomes for the normalized data was conducted using the R package “ANCOM” [8], which accounted for the underlying structure in the microbial data and was used to comparing the composition of microbiomes in two or more populations. For the main and mystery datasets, ANCOM across all pair-wise comparisons in each dataset was implemented. The 120 and 28 pair-wise comparisons were made for the combination of all cities in the main and mystery datasets respectively. And the output of the ANCOM was a set of differentially abundant features between the two cities for each comparison at the significance level of 0.05. The heatmap was made based on the output for investigating the difference of microbial composition between different cities. Additionally, the results of ANCOM were compared with the importance score derived by the RF method.

Declarations

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Availability of data and materials

The datasets supporting the conclusions of this article can be obtained from the CAMDA 2019 website http://camda2019.bioinf.jku.at/doku.php/contest_dataset.

Competing interests

The authors declare that they have no competing interests.

Funding

Datta, S. was partially supported by NIH grant 1UL1TR000064 from the National Center for Advancing Translational Sciences.

Author's contributions

SD reviewed the manuscript and provided theoretical support when required, RZ and ARW designed, run the analyses, RZ wrote the manuscript. All the authors have read and approved the final manuscript.

Acknowledgements

The samples were provided to the CAMDA 2019 competition by the MetaSUB Consortium.

Authors' information

[1]Department of Biostatistics, University of Florida, 2004 Mowry Rd, Gainesville, FL, 32610, USA

[2]Department of Oral Biology, University of Florida, 1395 Center Drive, Gainesville, FL, 32610, USA

Abbreviations

NGS: Next Generation Sequencing

WGS: whole genome sequencing

OTU: Operational Taxonomic Unit

RF: Random Forest

SVM: Support Vector Machine

LDA: Linear Discriminant Analysis

CV: Cross Validation

PCoA: Principal Coordinates Analysis

ANCOM: Analysis of composition of microbiomes

References

1. Simon C, Daniel R: Metagenomic analyses: past and future trends. Appl Environ Microbiol 2011, 77(4):1153-61. doi:10.1128/aem.02345-10.
2. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL: Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun 2016, 469(4):967-77. doi:10.1016/j.bbrc.2015.12.083.
3. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N: A global atlas of the dominant bacteria found in soil. Science 2018, 359(6373):320-5. doi:10.1126/science.aap9516.
4. Breiman L: Random Forests. Machine Learning 2001, 45(1):5-32. doi:10.1023/a:1010933404324.
5. Cortes C, Vapnik V: Support-vector networks. Machine Learning 1995, 20(3):273-97. doi:10.1007/bf00994018.
6. Balakrishnama S, Ganapathiraju A: Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing 1998, 18:1-8.
7. Borg I, Groenen P: Modern Multidimensional Scaling: Theory and Applications. Journal of Educational Measurement 2003, 40(3):277-80. 10.1111/j.1745-3984.2003.tb01108.x.
8. Mandal S, Van Treuren W, White RA, Eggesbo M, Knight R, Peddada SD: Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis 2015, 26:27663. doi:10.3402/mehd.v26.27663.
9. Walker AR, Grimes TL, Datta S, Datta S: Unraveling bacterial fingerprints of city subways from microbiome 16S gene profiles. Biology Direct 2018, 13(1):10. 10.1186/s13062-018-0215-8.
10. Walker AR, Datta S: Identification of city specific important bacterial signature for the MetaSUB CAMDA challenge microbiome data. Biology Direct 2019, 14(1):11. 10.1186/s13062-019-0243-z.
11. Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R: Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Bioinformatics 2011, Chapter 10:Unit 10.7. doi:10.1002/0471250953.bi1007s36.
12. Team RC: R: A language and environment for statistical computing. R foundation for Statistical Computing 2018.

13. Law CW, Chen Y, Shi W, Smyth GK: voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 2014, 15(2):R29. doi:10.1186/gb-2014-15-2-r29.
14. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 2015, 43(7):e47-e. doi:10.1093/nar/gkv007.
15. Balakrishnama S, Ganapathiraju A: Linear Discriminant Analysis—A Brief Tutorial, vol. 11; 1998.
16. Strobl C, Boulesteix AL, Zeileis A, Hothorn T: Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007, 8:25. doi:10.1186/1471-2105-8-25.
17. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A: Misc functions of the Department of Statistics (e1071), TU Wien. R package 2008, 1:5-24.
18. Ripley B: MASS: support functions and datasets for Venables and Ripley's MASS. R package version 2011:7.3-29.
19. Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, Suggests M: The vegan package. Community ecology package 2007, 10:631-7.
20. Paradis E, Blomberg S, Bolker B, Brown J, Claude J, Cuong HS, Desper R: Package 'ape'. Analyses of phylogenetics and evolution, version 2019, 2(4).

Tables

Table 1. Number of samples included in the analyses and their corresponding city and country of provenance. Table also showed the number of "species", "family", and "order" existing in each city.

The main dataset						
City	Country	Number of samples	Number of samples used	Species	Family	Order
Auckland (AKL)	New Zealand	14	14	133	43	21
Berlin (BER)	Germany	21	21	235	101	51
Bogota (BOG)	Colombia	15	15	703	205	138
Hamilton (HAM)	New Zealand	16	16	141	39	22
Hong Kong (HGK)	China	18	17	289	112	60
Ilorin (ILR)	Nigeria	24	24	230	54	29
London (LON)	U.K.	24	22	48	29	15
Marseille (MAR)	France	10	10	212	85	44
New York (NYC)	U.S.A.	26	26	562	159	83
Offa (OFA)	Nigeria	20	20	349	85	42
Porto (PXO)	Portugal	20	20	286	122	66
Sacramento (SAC)	U.S.A.	18	18	554	208	126
Sao Paulo (SAO)	Brazil	24	24	227	99	54
Sofia (SOF)	Bulgaria	10	10	275	102	49
Stockholm (STO)	Sweden	20	20	186	76	42
Tokyo (TOK)	Japan	25	25	573	173	103
All cities	-	305	302	1047	276	180
The mystery dataset						
City	Country	Number of samples	Number of samples used	Species	Family	Order
Brisbane (Bri)	Australia	7	6	74	53	32
Doha (Doh)	Qatar	3	3	59	38	24
Kiev (Kie)	Ukraine	8	7	144	80	45
Oslo (Osl)	Norway	12	12	505	148	83
Paris (Par)	France	8	6	136	73	41
Rio de Janeiro (Rio)	Brasil	12	12	343	109	53
Santiago (San)	Chile	6	6	309	113	58
Vienna (Vie)	Austria	5	5	167	72	35
All cities	-	61	57	700	188	109

Table 2. The number of the features selected based on additional rules.

The main dataset				
Rules		Number of Features		
		Species	Family	Order
a) Features existing in at least N cities	N=15	13	23	17
	N=14	26	31	19
	N=13	52	43	23
	N=12	75	54	29
	N=11	110	64	33
	N=10	150	73	36
	N=9	188	86	43
	N=8	234	97	48
b) Top M features with the highest ubiquity across all the samples	M=10	10	10	10
	M=20	20	20	20
	M=30	30	30	30
	M=50	50	50	50
	M=100	100	100	100
	M=150	150	150	150
c) Combination of the common features	“species”, “family” and “order”	25 (7 species, 9 families, 9 orders)		
	“species” and “family”	16 (7 species, 9 families)		
	“species” and “order”	16 (7 species, 9 orders)		
	“family” and “order”	18 (9 families, 9 orders)		
The mystery dataset				
c) Combination of the common features	“species”, “family” and “order”	41 (8 species, 18 families, 15 orders)		
	“species” and “family”	26 (8 species, 18 families)		
	“species” and “order”	23 (8 species, 15 orders)		
	“family” and “order”	33 (18 families, 15 orders)		

Table 3. The error rate with the leave-one-out cross-validation based on different rules. The number of features selected was retained in brackets.

Methods / Rules	Random Forest			Support Vector Machine			Linear Discriminant Analysis		
	Species	Family	Order	Species	Family	Order	Species	Family	Order
The main dataset									
Common features	0.583 (7)	0.299 (9)	0.218 (9)	0.562 (7)	0.276 (9)	0.239 (9)	0.613 (7)	0.303 (9)	0.328 (9)
Features existing in at least N cities									
N=15	0.464 (13)	0.357 (23)	0.369 (17)	0.459 (13)	0.327 (23)	0.344 (17)	0.517 (13)	0.361 (23)	0.381 (17)
N=14	0.370 (26)	0.335 (31)	0.349 (19)	0.340 (26)	0.298 (31)	0.339 (19)	0.352 (26)	0.286 (31)	0.384 (19)
N=13	0.345 (52)	0.287 (43)	0.313 (23)	0.325 (52)	0.289 (43)	0.291 (23)	0.336 (52)	0.263 (43)	0.289 (23)
N=12	0.351 (75)	0.297 (54)	0.299 (29)	0.324 (75)	0.273 (54)	0.292 (29)	0.296 (75)	0.235 (54)	0.243 (29)
N=11	0.329 (110)	0.299 (64)	0.291 (33)	0.313 (110)	0.270 (64)	0.281 (33)	0.291 (110)	0.240 (64)	0.216 (33)
N=10	0.318 (150)	0.296 (73)	0.290 (36)	0.321 (150)	0.276 (73)	0.266 (36)	0.345 (150)	0.257 (73)	0.218 (36)
N=9	0.291 (188)	0.292 (86)	0.299 (43)	0.299 (188)	0.287 (86)	0.283 (43)	0.379 (188)	0.249 (86)	0.210 (43)
N=8	0.306 (234)	0.303 (97)	0.280 (48)	0.308 (234)	0.298 (97)	0.268 (48)	0.506 (234)	0.274 (97)	0.203 (48)
Top M features with the highest ubiquity across all the samples									
M=10	0.456 (10)	0.411 (10)	0.425 (10)	0.474 (10)	0.416 (10)	0.433 (10)	0.502 (10)	0.474 (10)	0.447 (10)
M=20	0.350 (20)	0.318 (20)	0.333 (20)	0.355 (20)	0.322 (20)	0.328 (20)	0.366 (20)	0.309 (20)	0.325 (20)
M=30	0.351 (30)	0.274 (30)	0.289 (30)	0.317 (30)	0.286 (30)	0.275 (30)	0.336 (30)	0.270 (30)	0.239 (30)
M=50	0.290 (50)	0.290 (50)	0.269 (50)	0.272 (50)	0.276 (50)	0.250 (50)	0.251 (50)	0.231 (50)	0.198 (50)
M=100	0.264 (100)	0.298 (100)	0.282 (100)	0.272 (100)	0.305 (100)	0.301 (100)	0.231 (100)	0.233 (100)	0.272 (100)
M=150	0.273 (150)	0.313 (150)	0.296 (150)	0.276 (150)	0.308 (150)	0.398 (150)	0.282 (150)	0.275 (150)	0.398 (150)
Combination of the common features									
7 species, 9 families, 9 orders	0.116 (25)			0.115 (25)			0.118 (25)		
7 species, 9 families	0.248 (16)			0.215 (16)			0.236 (16)		
7 species, 9 orders	0.189 (16)			0.189 (16)			0.249 (16)		
9 families, 9 orders	0.133 (18)			0.118 (18)			0.133 (18)		
The mystery dataset									
Common features	0.598 (8)	0.370 (18)	0.307 (15)	0.615 (8)	0.420 (18)	0.332 (15)	0.659 (8)	0.320 (18)	0.314 (15)
Combination of the common features									
8 species, 18 families, 15 orders	0.272 (41)			0.327 (41)			0.445 (41)		
8 species, 18 families	0.369 (26)			0.47 (26)			0.415 (26)		
8 species, 15 orders	0.297 (23)			0.321 (23)			0.282 (23)		
18 families, 15 orders	0.255 (33)			0.332 (33)			0.343 (33)		

Figures

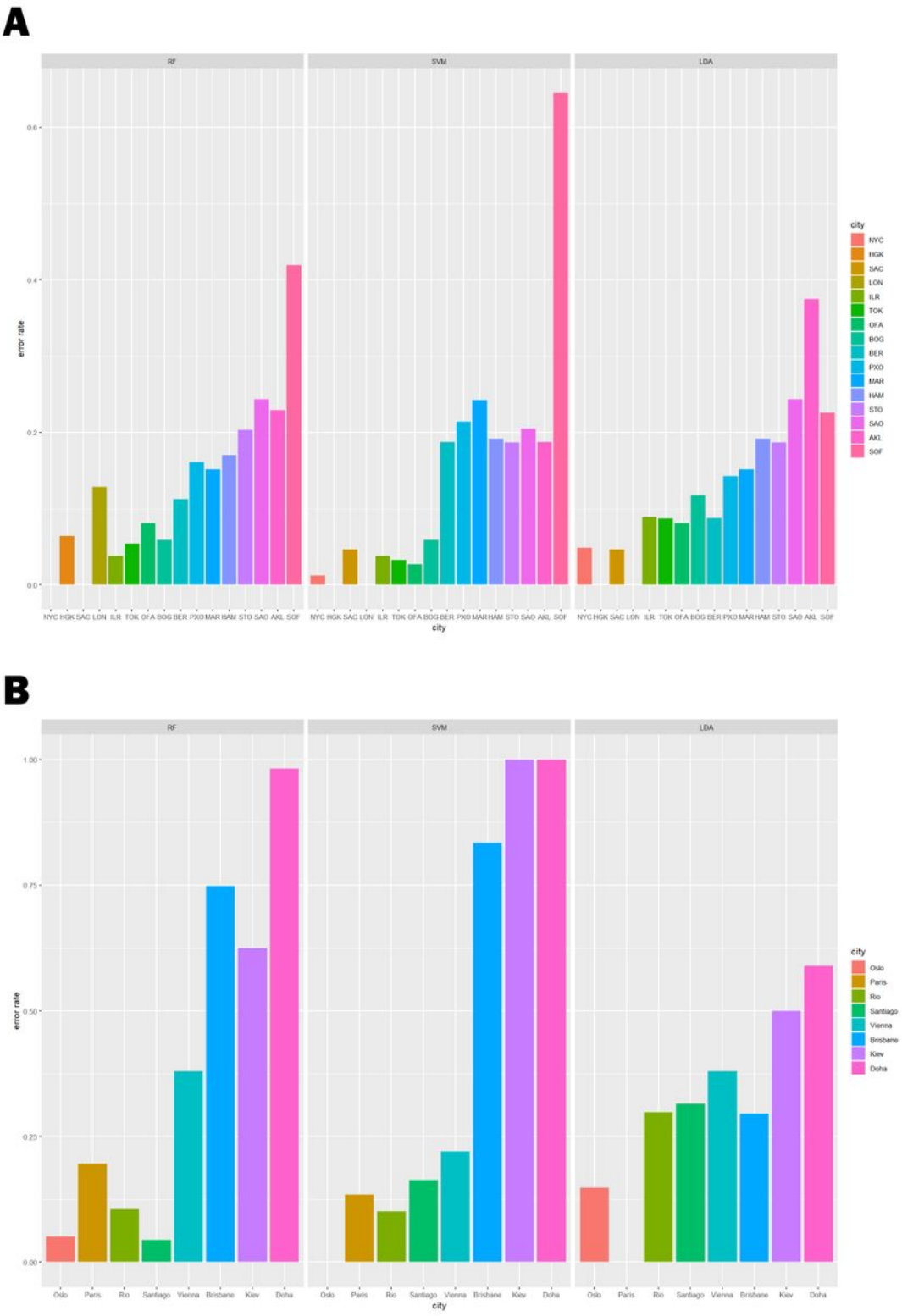

Figure 1

Error rate of each city based on combined features for both training datasets: main data dataset in A and mystery dataset in B. The cities are ordered by average error rate of three methods. The features were combined with common “species”, “family”, “order” and common “species”, “order” for main and mystery dataset, respectively.

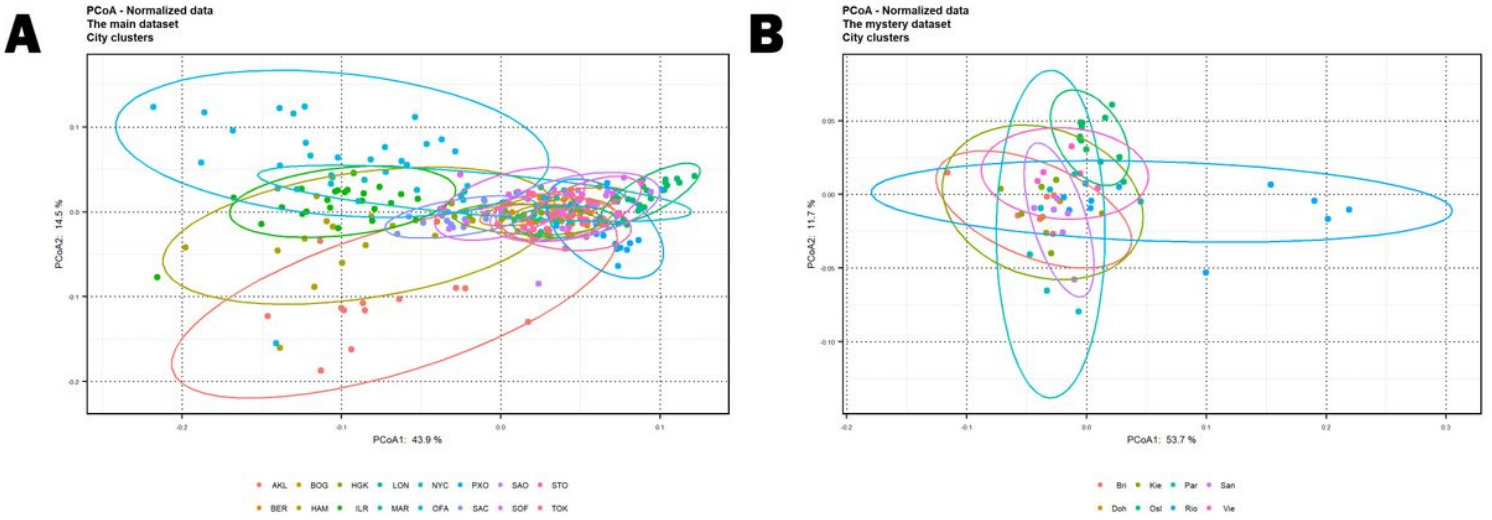

Figure 2

The bi-plots of first and second PCoA axes: main dataset in A and mystery dataset in B.

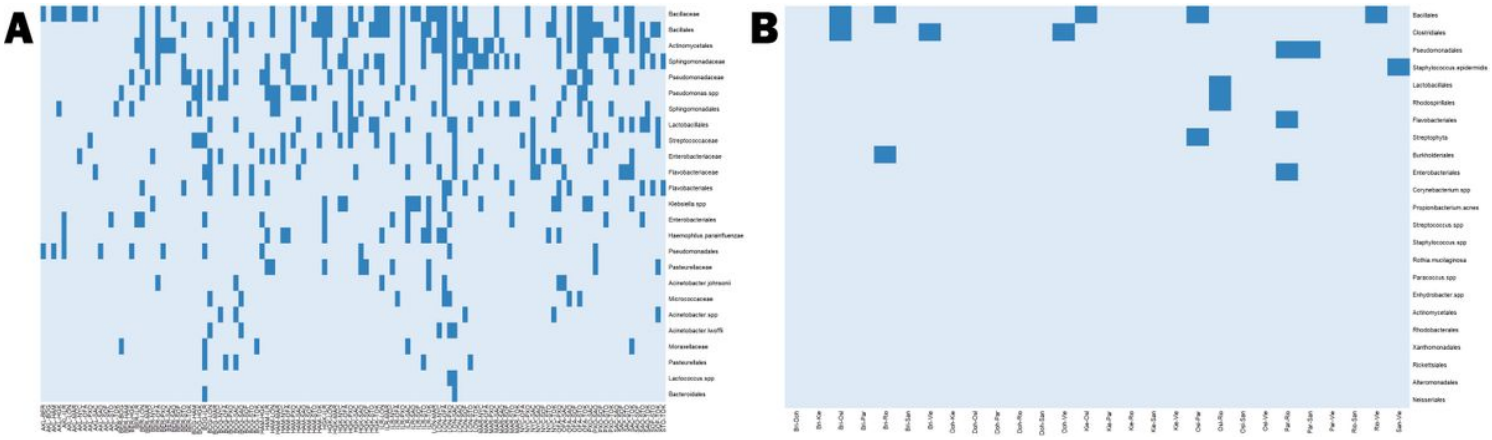

Figure 3

The analysis of composition of microbiomes across all pair-wise comparisons of cities: main dataset in A and mystery dataset in B. The significant features are denoted by deep blue, the features that are not significantly different in two cities are denoted by light blue.

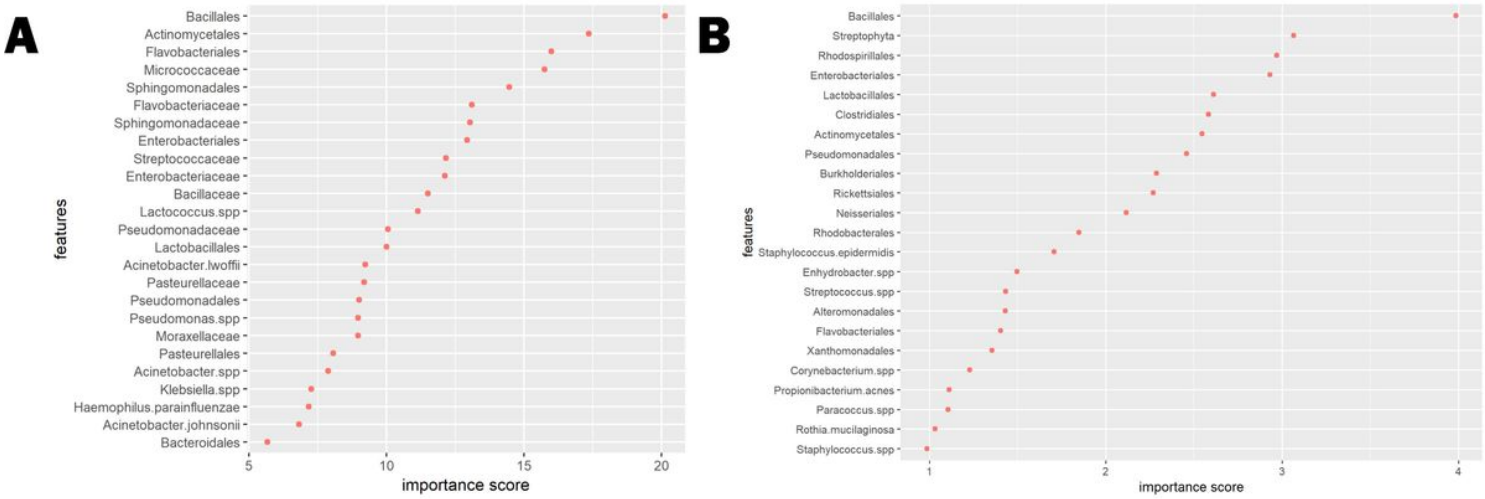

Figure 4

The importance of features obtained from the RF. The features are ordered by the importance.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS2.xlsx
- TableS1.xlsx
- TableS3.xlsx