Systematic evaluation of supervised machine learning for sample origin prediction using metagenomic sequencing data

Chih-yu Chen (chih-yu.chen@canada.ca)
Public Health Agency of Canada - National Microbiology Laboratory https://orcid.org/0000-0003-4790-0025

Andrea D Tyler
Public Health Agency of Canada

Research

Keywords: Machine learning, Metagenomics, Microbiome, Multivariate regression, Lasso regularization, Multiclass Classification, CAMDA, MetaSUB

Posted Date: March 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-15502/v1

Version of Record: A version of this preprint was published on December 10th, 2020. See the published version at https://doi.org/10.1186/s13062-020-00287-y.

Abstract

Background:

The advent of metagenomic sequencing provides microbial abundance patterns that can be leveraged for sample origin prediction. Supervised machine learning classification approaches have been reported to predict sample origin accurately when the origin has been previously sampled. Using metagenomic datasets provided by the 2019 CAMDA challenge, we evaluated the influence of technical, analytical and machine learning approaches for result interpretation and source prediction of new origins.

Results:

Comparison between 16S rRNA amplicon and shotgun sequencing approaches as well as metagenomic analytical tools showed differences in measured microbial abundance of the same samples, especially for organisms present at low abundance. Shotgun sequence data analyzed using Kraken2 and Bracken taxonomic annotation, had higher detection sensitivity than did other methods. As classification models are limited to labeling previously trained origins, we proposed an alternative approach using Lasso-regularized multivariate regression to predict geographic coordinates for comparison. In both models, the prediction errors were much higher in Leave-1-city-out than in 10-fold cross validation, the former of which realistically forecasted the difficulty in accurately predicting samples from new origins than pre-trained origins. The challenge was further confirmed using mystery samples obtained from new origins. Overall, prediction performances between regression and classification models, as measured by mean squared error, were comparable on mystery samples. Due to higher prediction errors for samples from new origins, we provided an additional strategy based on prediction ambiguity to infer whether a sample is from a new origin for practical applications. Lastly, we showed increased prediction error when data from a different sequencing protocol were included as training data.

Conclusions:

Here we highlighted the capacity of predicting sample origin accurately with pre-trained origins and the challenge of predicting new origins through both regression and classification models. Overall, the work provided a summary evaluation of sequencing techniques, protocol, taxonomic analytical approaches, and machine learning approaches to inform future designs in metagenomic prediction of sample origin.

Background

Studies of microbiome have demonstrated successes in detecting microbial compositional patterns in health and environmental contexts. Large scale studies such as the Human Microbiome Project [1], the Metagenomics & Metadesign of Subways & Urban Biomes (MetaSUB) [2] and the Earth Microbiome Project [3] exemplify global efforts to facilitate the understanding of microbial presence and abundance in relation to diseases or environmental factors. Recent technological advances have enhanced the ability to both detect varied species and estimate their abundance in collected samples. The 16S ribosomal RNA

(rRNA) amplicon sequencing approach targets and specifically sequences the 16S rRNA gene of bacteria and archaea; whereas the shotgun whole genome sequencing approach sequences all genetic material present in a sample, potentially allowing identification to the level of species, detection of the presence of functional units, and concurrent identification of eukaryotes, fungi and DNA viruses. Comparisons between 16S amplicon and shotgun sequencing data have been examined previously, but discrepancy exists among studies regarding which technique provides better robustness and higher biodiversity [4–7]. Despite the pros and cons of each technique, successes in extracting meaningful biological information have been found for disease and environmental studies [2,8–12].

Leveraging metagenomics sequencing data, the majority of analytical approaches for sample source prediction used to date have focused on supervised classification methods such as support vector machines and random forest, in order to assign trained source labels to unknown samples [9,10,13,14]. Delgado-Baquerizo *et al*. found high variability in relative abundance across various geographical locations through examining soil microbiome, and used random forest modeling to predict habitat preference for dominant phylotypes [9]. In the Earth Microbiome Project, random forest models were built to distinguish samples from various environmental factors including association with plants or animals as well as saline presence [10]. From the perspective of identifying potentially mixed sources, SourceTracker [15] uses a Bayesian approach to estimate the proportions of source environments in a sample without the assumption of one source label. In the 2018 Critical Assessment of Massive Data Analysis (CAMDA) challenge, supervised classification approaches have been applied to predict sample source using urban microbiome with high accuracies up to 0.91, where the unknown samples were of the same origins as samples previously trained [13,16–18].

The objective of the 2019 CAMDA metagenomics forensic challenge was to use urban microbiome data to predict locations of samples from new origins that have not been sampled previously (Figure S1). Classification models are limited to predicting trained origins from which the training samples were already collected and trained; hence can never predict a new origin. For the purpose of predicting new origins, one alternative approach is to model geographic coordinates, as inspired from a previous report on association between human genetics and geographical locations [19]. There have been existing literature reports on the association between latitude and microbial composition in various contexts [20–23]. Richness/diversity in planktonic marine bacteria and microbiome from ambulances in USA were found to be inversely correlated with latitude, a pattern called the “latitudinal diversity gradient” [21,22]. Using human gut microbe data from 23 populations, Suzuki *et al*. found significant positive and negative correlations to latitude with Firmicutes and Bacteriodetes, respectively [23]. Fisman *et al*. reported correlation between bloodstream infection from gram negative bacteria and proximity to the equator measured by latitude-squared [20].

Given the availability of 16S rRNA amplicon and shotgun data, we first set out to compare and contrast organism abundance from datasets generated using 16S amplicon versus shotgun sequencing technologies, and evaluated different analytical approaches on a subset of samples. In order to perform attribution of samples to a new geographic origin, we herein model the longitude and latitude as the

outcome variables, and evaluate the use of multivariate regression for predictions of new sample origins with Lasso regularization to avoid model overfitting. Subsequently, we compare prediction performance between multivariate regression and multiclass classification models for the mystery data from new origin. Lastly, we report a computational approach to identify whether a sample is from a new or trained origin through the Simpson's diversity index on classification probabilities.

Methods

Analyses were conducted in R 3.5.3 version unless otherwise stated.

Dataset description

The Boston Urban [11] and MetaSUB [2] datasets used for this study were provided by the CAMDA organizers: 1) Boston pilot 16S and shotgun data of 23 samples, 2) shotgun data of 294 samples from 16 cities for training and 3) shotgun data of 60 mystery samples from 8 new city origins as the test set. For the Boston pilot dataset, the 16S and shotgun abundance tables were provided through the CAMDA challenge and had been previously generated by Hsu et al [11] using QIIME v1.8 [24] and MetaPhlAn2 [25], respectively. We then further analyzed the latter using the Kraken2-Bracken [26, 27] assignment approach described below. The shotgun training dataset was generated via different sequencing protocols, and included sequencing data obtained from 16 cities: Auckland (AKL), Berlin (BER), Bogota (BOG), Hamilton (HAM), Hong Kong (HGK), Ilorin (ILR), London (LON), Marseille (MAR), New York (NYC), Offa (OFA), Porto (PXO), Sacramento (SAC), SAO Paulo (SAO), Sofia (SOF), Stockholm (STO) and Tokyo (TOK). Data from fifteen of the training cities (276 samples) were generated using paired-end sequencing at 150-basepair (bp) reads, whereas SAC data were single-ended with 125 bp reads. More details on samples regarding sequence lengths, single-/paired-end and percentage of unclassified reads were reported in Table S1.

Taxonomic abundance estimation and data processing

Taxonomic sequence classification and organism abundance estimation for shotgun datasets were conducted using Kraken2 [26] and Bracken [27], respectively, using a customized reference database which included reference sequence representatives of the bacterial, viral, archaeal groups as well as the human genome, all obtained from NCBI Refseq December, 2018. The Bracken database was constructed using 150 bp reads, as these were the most commonly identified in the CAMDA dataset. The median percentage of unclassified reads for all CAMDA files was 44.92%. Reads assigned to the human genome were filtered out. Normalization of the total abundance table for each taxonomic scale was done using the cumulative sum scaling approach at the 50th percentile (metagenomeSeq [28] R package). To avoid spurious results from sparse features, taxa that did not contain at least one read in more than eight samples were filtered out. The Bray Curtis dissimilarity and principal coordinate analysis with the Cailliez correction for negative eigenvalues were powered by the vegan [29] and ape [30] R packages.

Machine learning models for prediction of sample source

Instead of estimating latitude and longitude in separate models, we chose to take into account dependencies between the two coordinates, and modeled them together using multivariate regression with Lasso regularization (glmnet R package) [31]. For each taxonomic model, the normalized and $\log_2$-transformed microbial abundance was standardized as input features, and latitudes and longitudes were the response variables. Ten-fold cross validation was conducted with mean squared error (MSE) evaluation to choose the hyperparameter λ such that the error from the model is within one standard error of the minimum. Performance was reported using 10-fold nested cross validation (CV) to evaluate prediction accuracy of samples from previously sampled cities or leave-1-city-out (L1CO) CV to evaluate accuracy of samples from new cities. For comparison, a classification (Lasso-regularized multinomial logistic regression) approach was also performed using the glmnet R package [31]. While accuracy is a common performance measure for classification models, MSE was also reported for prediction of mystery samples using the classification model in order to compare to the regression model.

Binary machine learning classifier on prediction ambiguity

For the purpose of predicting whether a sample is from a previously trained city, a Simpson's diversity value was obtained using class prediction probabilities of each sample from the sample source classification model (vegan R package [29]). The index was used to reflect the prediction ambiguity of each sample from the model, as was previously proposed in a genomic study [32]. Overall, two diversity values were obtained for each sample in L1CO and 10-fold CV settings, mimicking the prediction of a new origin or a previously trained origin, respectively. Subsequently, a Naïve Bayes classifier using kernel density estimation was built to learn the distributions of Simpson's diversity values from new (L1CO CV) versus previously-trained (10-fold CV) cities. Classifier performance was evaluated by leaving each Simpson's value out to predict the CV setting it belongs to. The classifier was then used to predict whether a mystery sample is from a new or trained origin given its class prediction probabilities.

Results

Abundance differences between technological and analytical approaches

Given that both 16S and shotgun approaches were used to sequence the 23 Boston samples collected from several surfaces and two sources [11], we first investigated counts of detected organisms at varying taxonomic scales and their abundance using both datasets. For the shotgun data, we evaluated the abundance table extracted from Kraken2 [26] and Bracken [27] (referred to as SG-KB) as well as the table provided by CAMDA using MetaPhlAn2 [25] (referred to as SG-MP). There is an overall higher sensitivity in the SG-KB data, with at least twice as many distinct bacterial genera, families and orders compared to the other methods. SG-MP data analysis reported the least distinct taxa at all taxonomic levels except for species (Table 1). As expected, both shotgun datasets identified more distinct species than 16S data. While there were 90 overlapping species identified between SG-KB and 16S data, there were no overlaps between SG-MP and 16S datasets. We next examined reported abundance of commonly detected

organisms between technologies in all 23 Boston samples. Pearson correlation coefficients between 16S and SG-KB abundance at species, genus, family and order levels were 0.48, 0.70, 0.75 and 0.85, respectively (Figs. 1A-D). The correlation coefficients between 16S and SG-KB were slightly, but significantly higher than those between 16S and SG-MP data (One-sided paired t-test p = 0.0093; Figs. 1E-H). Better correlation was found at higher taxonomic levels. At any given taxonomic rank, variation was greater for organisms detected at lower abundance. The histograms further revealed an overall trend of inflated zero counts in 16S datasets for many taxa that were detected at low abundance by SG-KB (Figs. 1A-D). On the other hand, such zero inflation was revealed in SG-MP instead when compared with the 16S data (Figs. 1E-H). The discrepancy in zero inflation between Kraken2 + Bracken and MetaPhlAn2 processed shotgun data with respect to 16S data highlights the impact on overall model interpretation due to the taxonomic identification tools and databases used. Given the higher sensitivity of SG-KB data, we proceeded using the taxonomic analysis of shotgun data using Kraken2 and Bracken for the rest of the manuscript. Examination of Bray-Curtis dissimilarities between samples from both 16S and SG-KB data showed clustering according to sequencing method (Fig. 2). The principal coordinate analysis (PCoA) highlighted the differences in the two sets in the first dimension, which explains 45% of the total variance, whereas samples collected from different surfaces were observed to cluster in later dimensions irrespective of technologies (Fig. 2B).

Table 1
Counts of unique taxa identified using Boston pilot 16S amplicon and shotgun metagenomics datasets in at least one samples

		Overall Taxa Counts					
Technology (tool)	Taxa	species	genus	family	order	class	phylum
Amplicon 16S	Bacteria	143	328	173	83	52	18
Shotgun (Kraken2 + Bracken)	All	4200	1339	399	176	82	43
	Bacteria Only	3392	1098	342	154	72	38
	Overlap	90	226	136	54	27	15
Shotgun (MetaPhlAn2)	All	493	239	116	50	28	16
	Bacteria Only	459	211	102	44	23	12
	Overlap	0	128	84	36	17	9

Modelling geographic coordinates using microbiome data

To tackle the challenge of predicting sample source in general, we first examined the training dataset of MetaSUB paired-end shotgun data from 15 cities. Principal coordinate analysis at species abundance

level with Bray-Curtis dissimilarity and heatmap visualization showed clustering of samples from Oceania and South Africa, whereas other cities or continents overlapped one another in the first two dimensions (Fig. 3). Better separation of continents was observed with the third dimension. In order to predict locations of mystery samples from new origins, we modeled geographic coordinates using multivariate regression with Lasso regularization and using microbial abundance as features/predictors. We constructed a Lasso-regularized regression model at each taxonomic scale using the organism abundance table of the corresponding taxa scale as input data. To first confirm the potential of predicting a city where some samples have been analyzed and the model trained on, we evaluate model performance using nested 10-fold CV. The model performance was the highest at the species level compared to that of genus and family (Table 2). Scatterplots of true coordinates and the 10-fold CV predictions from species regression model showed a linear trend between predicted and true coordinates in Figs. 4A and B (r^2 = 0.8891 for latitude and 0.8860 for longitude). In comparison, 10-fold nested CV performance of the Lasso regularized species classification model achieved a high accuracy level of 0.942 (Fig. 4C).

Table 2
Nested 10-fold and leave-one-city-out cross validation performance for geographic coordinate prediction using multivariate regression with LASSO regularization

LASSO	EMP	lambda 1se	df	MSE 10-fold	MSE L1CO
MetaSUB Shotgun	species	3.0609	108	837	5962
	genus	2.7028	96	1081	5686
	family	2.4684	81	1480	6202

The challenge in prediction of new sources without previous samples

To evaluate the prediction performance on samples from novel origins, we conducted leave-1-city-out CV (Figure S2). The r^2 values of the regression model were 0.5129 and 0.1299 for latitude and longitude respectively using the prudent setting. The L1CO MSE of the regression model is 7-fold the MSE for the nested 10-fold CV, which highlights the challenge in predicting new origins that have not been included in the model training. When predicting sources for the mystery data, the r^2 values were 0.4697 and 0.4670 for latitude and longitude with the regression model (Fig. 5AB). Contrary to our expectation, prediction performances on the mystery samples by Lasso-regularized regression and classification models were comparable as indicated by MSE values (Table 3; Table S2). As a side note, we also conducted random forest classification, a nonlinear approach, for comparison, but the classifier resulted in higher MSE values on mystery samples compared to Lasso-regularized models (Figure S3). Overall, mystery samples

from Brisbane had the highest prediction error in both Lasso-regularized models, despite having 'nearby' samples from New Zealand as training data (Figs. 5D-G). The squared errors for samples from Kiev and Rio de Janerio were significantly lower in the regression model than the classification model, indicating better performance by regression in these cases (one-sided Wilcoxon test, Benjamini-Hochberg adjusted p = 0.0263, 0.0210, respectively). On the other hand, predictions for Oslo, Paris and Santiago de Chile samples were better in classification (p = 0.0000, 0.0104, 0.0450). Despite overall comparable MSE values between the regression and classification models, we observed better prediction for some cites using either model.

Table 3
Prediction performance of mystery (new) cities in the test dataset as measured by MSE

Lasso-regularized	Paired and single end	Longitude MSE	Latitude MSE	M of total SE
Regression	No	2285.70	1103.41	3389.11
Classification	No	2554.01	1091.56	3645.57
Regression	Yes	8468.87	1435.02	9903.88
Classification	Yes	6823.99	1826.64	8650.63

As the classification model cannot accurately predict new origins at the level of longitude and latitude, we next evaluated the classification predictions in the context of continents. In L1CO CV setting, all samples from AKL, HAM, SOF and MAR were predicted to be on the same continent, while the rest of cities varied (Figure S2C). Prediction results from the classification model showed that all mystery samples from Vienna, Santiago de Chile, France and Oslo were predicted to be cities within the same continent (Fig. 5C). Although Doha is considered to be in Asia, it is geographically closer to Marseille than to Hong Kong, and two of the three Doha samples were predicted to be from Marseille. Conversely, Kiev is within Europe and closer to European cities, but the five out of seven Kiev samples were predicted to be in Hong Kong.

Using prediction ambiguity to evaluate if a sample is from a new origin

Given the much lower prediction accuracy for samples from new origins, we rationalized that it may be beneficial to tag whether a new sample is from a new origin in real life application when using a classification approach. We hypothesized that samples from new origins will have higher prediction ambiguities. Specifically, we investigated the prediction ambiguity from the classification model on each leftout sample in both 10-fold and L1CO CV settings. Here we defined the ambiguity of each predicted sample to be the Simpson's diversity index of its class probabilities. Samples with higher prediction ambiguity are expected to have prediction probabilities distributed across multiple classes, as reflected by higher diversity in class probabilities. As expected, leftout samples from new origins (in L1CO CV) have significantly higher ambiguity compared to leftout samples from previously-trained origins (in 10-fold CV; Wilcoxon test p-value = 4.6×10^{-54}; Fig. 6A). Based on Simpson's values from both CV settings, we built a

Naïve Bayes classifier with kernel density estimation on the Simpson's values to classify whether a sample is from a new origin. The evaluation using leave-one-out CV reported an accuracy of 0.83 (sensitivity = 0.79; specificity = 0.87; Fig. 6B). Forty-five out of sixty mystery samples were correctly predicted to be from new origins using this binary classifier (sensitivity = 0.75; Fig. 6C). While eight Oslo, four Paris and three Vienna samples were wrongly predicted to be from pre-trained origins, we note that samples from these cities were predicted to be cites within the correct continent and with low squared errors in the sample source classification model (Figs. 5 CFG). Hence, our new-origin classification model based on prediction ambiguity of the sample-source classification model indicated its potential to inform whether a mystery sample is from a new origin, which can serve as a flag given the lower prediction accuracy of samples from new origins.

Inclusion of training data with different experimental protocols impacts model performance

Due to protocol differences in sequence lengths and single versus paired-end sequencing, we excluded the training data from the 16th city, SAC, provided by CAMDA in the main analyses above. To evaluate the effect of mixing data from different experimental protocols, we also examined the model performance with the inclusion of the SAC along with 15 other cities as training data (Table 3 and Figure S4). The MSEs of regression and classification models on the mystery samples increased 2.9- and 2.4-folds compared to models trained on 15 cities, respectively. Overall, the regression predictions resulted in a longitudinal shift away from the diagonal (Figure S4A). Notably, the squared errors of Oslo samples on longitude increased substantially after incorporating SAC data for training in both regression and classification models (Figures S4D and S4F). Specifically, the classification model predicted Oslo samples to be in Sacramento exclusively (Figure S4C). These results indicated the impact of incorporating datasets with very different sequencing protocols.

Discussion

Here we utilized MetaSUB and Boston Urban datasets provided by the CAMDA organizers to extract knowledge and elucidate factors that impact the prediction of sample sources. Through the comparison between 16S amplicon and shotgun metagenomic sequencing data on the same samples, we highlighted differences in detection sensitivity and abundance between sequencing technologies as well as analytical tools and databases used. On predicting sample sources, we demonstrate the importance of using 10-fold CV to evaluate prediction performance on samples from trained origins versus using leave-1-city-out CV to evaluate prediction performance on samples from new origins. The substantially higher prediction errors of the L1CO CV highlighted that prediction of new cities are more challenging than prediction of samples for which the origins were included in the training. Our comparison of Lasso-regularized classification (prediction of cities) and regression (prediction of geographic coordinates) approaches reported comparable MSE, and showed the benefits of either approach for predicting trained and new origins. As an extension of the source classification model, our use of prediction ambiguity

based on the Simpson's index allowed flagging of samples from a new origin. Lastly, we demonstrated reduced model performance when incorporating a dataset from a different experimental protocol as training data, highlighting the impact of heterogeneous experimental protocol.

Differences in abundance detection sensitivity between technologies, tools and databases

Consistent with recent work on stool and colon biopsy samples comparing between 16S versus pair-end shotgun data [5], we reported variation in abundance results between taxonomic classification tools and databases in addition to sequencing technologies. Our results further showed higher variation between technologies for taxa with lower abundance, and demonstrated that zero inflation in abundance is dependent on both experimental and analytical approaches. The zero inflation in 16S data compared to shotgun data by Kraken2 and Bracken is potentially due to high conservation of 16 s rRNA gene making finer taxonomic levels more difficult to identify [33]. Amplification biases attributed to use of the 16S marker gene have been previously described, and may have also contributed to variability between methods [34, 35]. Importantly, it is also possible that differential multiplexing in combination with other variation in sequencing methodologies employed by shotgun and 16S amplicon analysis, contributed to these differences. Unfortunately, further information on sample handling and sequencing preparation was not made available for this analysis. On the other hand, the zero inflation observed in shotgun data by MetaPhlAn2 and its lower detection sensitivity is likely due to the database used. To our knowledge, the default database used for the provided MetaPhlAn2 data was geared towards gut microbes, which may have resulted in information loss for the study of urban microbiome. Lastly, alternative normalization and filtering approaches may be further evaluated and refined before downstream analyses [33, 34], especially in the case of analyzing data with mixed experimental protocols.

Prediction of new locations & supervised regression and classification approaches

As reported in the 2018 CAMDA challenge [13, 16–18], our 10-fold CV results showed a strong potential to predict the source of samples from previously-trained origins using shotgun metagenomic data. Abundance at the species level resulted in the highest predictive power for geographic coordinates compared to other taxonomic scales. On the other hand, the L1CO results highlighted the difficulty of predicting origins without prior training samples, which was further confirmed in the source prediction of mystery samples from new cities. Despite the attempt to utilize a regression approach to address this challenge, the prediction performance as measured by MSE on mystery samples did not show much improvement compared to the classification approach. We note however that some cities were predicted more accurately by either approach. Overall, the regression approach may be limited by its assumption of linearity, increased sensitivity to outlying coordinates, and the difficulty in interpreting feature importance along geographic coordinates. The classification approach can elucidate signatures in trained cities, which may be easier to interpret, but it can only predict new origins to the closest cities at best. This limitation may gradually be alleviated as samples from more origins are collected and trained. Lastly, the

diversity of the classification prediction can be used to inform prediction uncertainty, indicating whether the sample is from a new origin.

Covariates and metadata

The machine learning approaches we took for prediction of sample source did not include the metadata as they were unavailable. Further investigation to include metadata into the analyses will be beneficial to account for variation from other covariates, such as micro-environment, seasons and city connections. Hidden covariates can introduce noise in the model, and may confound the analysis via causing spurious association as a confounder. Previous work from the global atlas of soil bacteria has shown habitat preferences for different soil bacterial phylotypes depending on environmental conditions, such as pH levels, plant productivity and aridity [9]. A recent study on tropical air ecosystem reported higher inter-day taxonomic diversity than day-to-day and month-to-month variation [36]. We also note that the sampling date of MetaSUB for cities in the northern and southern hemispheres corresponds to summer and winter solstices, respectively, which may have introduced seasonal influence. On the other hand, studies in other contexts such as bloodstream infection [20] and topsoil [37] have reported associations between microbial pattern and the distance from equator. The non-linearity would not have been captured in our regression model. Furthermore, despite having training samples from nearby New Zealand cities, the high prediction errors of mystery samples from Brisbane Australia in both models indicated other factors are likely at play.

Heterogeneous experimental protocols used for sample collection and sequencing between cities can have a substantial influence on the prediction performance of new sources. In our experience, the inclusion of single-end data from Sacramento as training data reduced the performance of the model prediction on mystery samples by two- to three-fold as measured by MSE. This is likely because the SAC data was generated single-ended with read lengths of 125bps, instead of paired-ended with 150 bp reads like all other training data. Short sequence reads are more challenging to taxonomically assign to a group, given the reduced amount of information available in each read. When SAC data was included for training, mystery samples originating from Oslo had an increase in longitudinal squared errors and were predicted to be in Sacramento exclusively in the classification model. We note that within mystery samples, those from Oslo were the only ones which were single-ended and had shorter read lengths (125bps and 100bps); hence it is possible that read length differences in experimental protocol was substantially impacting model performance. This finding highlighted the need to evaluate data from different protocols on the same samples first, before deciding to analyze them together, as was done for the Boston 16S and shotgun data.

Other features and future prospects

Here we used only microbial taxonomic abundance as the features from shotgun metagenomic data to predict the sample source. Given the advantage of shotgun data, further investigations on other types of features [38, 39] can be extended from our approach. Biologically-driven features include functional pathway and antimicrobial resistance profiles [11, 18], whereas data-driven features include k-mer counts

and genomic bins without annotation requirements [38, 40, 41]. This information can be incorporated into multifaceted analyses for prediction of the sample source through comparison of models with different feature types and/or integration of multi-layered feature sets. One important observation is the high proportion of unclassified sequences across many of the samples included in this analysis. Annotation independent information extraction such as k-mer counts may be beneficial, if there exist taxa that are not yet annotated and differ in abundance between cities. As our knowledge of microbes increases, and databases become more complete, the number of microbes remaining unclassified will be reduced, enhancing the capacity to delineate signals between sources. The continual updates in databases with newly identified taxa will ultimately enhance the capacity to delineate signals between sources.

Conclusions

In this work, we have highlighted the impact of sequencing approaches, taxonomic annotation tools, databases and heterogeneous protocols on result interpretation and model performance. We demonstrate the practical purpose of performance evaluation using 10-fold and leave-one-city-out cross validation for predicting pre-trained and new origins, respectively. The proposed Lasso-regularized multivariate regression provided a novel and alternative approach to source prediction with comparable performance to the classification approach. Due to the demonstrated challenge in predicting new origins without any metadata, we further provided the strategy of using classification prediction diversity to flag whether a sample is from a new origin for real life applications. Overall, our work informs future metagenomics studies on the potential and challenges for source prediction using machine learning methods.

Abbreviations

bp:Base Pair; CAMDA:Critical Assessment of Massive Data Analysis; CV:Cross Validation; L1CO:Leave-1-City-Out; MetaSUB:Metagenomics & Metadesign of Subways & Urban Biomes; MSE:Mean Squared Error; PCoA:Principal coordinate analysis; SG-KB:Organism abundance extracted by the Kraken2 and Bracken tools; SG-MP:Organism abundance provided by CAMDA using the MetaPhlAn2 tool

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

All authors have given their consent to publish the findings in this paper.

Availability of data and materials

No experimental datasets were generated for the study. The datasets used are available through the CAMDA 2019 website

(http://camda2019.bioinf.jku.at/doku.php/contest_dataset#metagenomic_forensics_challenge).

Competing interests

The authors declare they have no competing interests.

Funding

Not applicable

Authors' contributions

CYC and ADT conceived the project. ADT designed and implemented data preprocessing and QC of the dataset. CYC designed and implemented the statistical and machine learning analyses and generated the figures. CYC and ADT interpreted the results. CYC drafted and ADT revised the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors thank the CAMDA organizing committee and MetaSUB for organizing the challenge and providing the metagenomics data. We also thank Drs. Gary Van Domselaar, Natalie Knox, and Morag Graham as well as Mr. Eric Marinier for the feedback and discussion on the CAMDA presentation.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007;449:804–10.
2. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. Microbiome [Internet]. 2016 [cited 2019 May 27];4:24. Available from: www.metasub.org
3. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biol [Internet]. BioMed Central; 2014 [cited 2019 Oct 16];12:69. Available from: http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-014-0069-1
4. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, et al. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci Rep [Internet]. Nature Publishing Group; 2017 [cited 2019 May 30];7:6589. Available from: http://www.nature.com/articles/s41598-017-06665-3
5. Mas-Lloret J, Obón-Santacana M, Ibáñez-Sanz G, Guinó E, Pato ML, Rodriguez-Moranta F, et al. Gut microbiome diversity detected by high-coverage 16S and shotgun sequencing of matched stool and colon biopsy samples. bioRxiv [Internet]. 2019;742635. Available from: https://www.biorxiv.org/content/10.1101/742635v1

6. 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 [Internet]. Academic Press; 2016 [cited 2019 May 30];469:967–77. Available from: https://www-sciencedirect-com.ezproxy.cscscience.ca/science/article/pii/S0006291X15310883?via%3Dihub
7. Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C. Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. Omi A J Integr Biol. 2018;22:248–54.
8. Forbes JD, Chen C, Knox NC, Marrie R, El-gabalawy H, Kievit T De, et al. A comparative study of the gut microbiota in immune-mediated inflammatory diseases — does a common dysbiosis exist? Microbiome [Internet]. BioMed Central; 2018 [cited 2019 Jan 17];6:1–15. Available from: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0603-4
9. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, et al. A global atlas of the dominant bacteria found in soil. Science (80-). 2018;359:320–5.
10. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature [Internet]. 2017 [cited 2019 Apr 18];551:457–63. Available from: http://www.earthmicrobiome.
11. Hsu T, Joice R, Vallarino J, Abu-Ali G, Hartmann EM, Shafquat A, et al. Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment. mSystems. 2016;1:1–18.
12. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: Successes and aspirations. BMC Biol. 2014.
13. Ryan FJ. Application of machine learning techniques for creating urban microbial fingerprints. Biol Direct [Internet]. BioMed Central; 2019 [cited 2019 Sep 24];14:13. Available from: https://biologydirect.biomedcentral.com/articles/10.1186/s13062-019-0245-x
14. Pasolli E, Truong DT, Malik F, Waldron L, Segata N, Grisel O. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. Eisen JA, editor. PLOS Comput Biol [Internet]. Springer; 2016 [cited 2017 May 23];12:e1004977. Available from: http://dx.plos.org/10.1371/journal.pcbi.1004977
15. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods. 2011;8:761–5.
16. Harris ZN, Dhungel E, Mosior M, Ahn T-H. Massive metagenomic data analysis using abundance-based machine learning. Biol Direct [Internet]. BioMed Central; 2019 [cited 2019 Sep 24];14:12. Available from: https://biologydirect.biomedcentral.com/articles/10.1186/s13062-019-0242-0
17. Walker AR, Datta S. Identification of city specific important bacterial signature for the MetaSUB CAMDA challenge microbiome data. Biol Direct [Internet]. BioMed Central; 2019 [cited 2019 Sep 24];14:11. Available from: https://biologydirect.biomedcentral.com/articles/10.1186/s13062-019-0243-z

18. Casimiro-Soriguer CS, Loucera C, Perez Florido J, López-López D, Dopazo J. Antibiotic resistance and metabolic profiles as functional biomarkers that accurately predict the geographic origin of city metagenomics samples. Biol Direct [Internet]. BioMed Central; 2019 [cited 2019 Sep 24];14:15. Available from: https://biologydirect.biomedcentral.com/articles/10.1186/s13062-019-0246-9
19. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature [Internet]. 2008 [cited 2018 Nov 28];456:98–101. Available from: https://www-nature-com.ezproxy.cscscience.ca/articles/nature07331.pdf
20. Fisman D, Patrozou E, Carmeli Y, Perencevich E, Tuite AR, Mermel LA, et al. Geographical Variability in the Likelihood of Bloodstream Infections Due to Gram-Negative Bacteria: Correlation with Proximity to the Equator and Health Care Expenditure. Chuang J-H, editor. PLoS One [Internet]. Public Library of Science; 2014 [cited 2019 Jun 20];9:e114548. Available from: https://dx.plos.org/10.1371/journal.pone.0114548
21. Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown M V, Green JL, et al. A latitudinal diversity gradient in planktonic marine bacteria. Proc Natl Acad Sci U S A [Internet]. National Academy of Sciences; 2008 [cited 2019 Jul 11];105:7774–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18509059
22. O'Hara NB, Reed HJ, Afshinnekoo E, Harvin D, Caplan N, Rosen G, et al. Metagenomic characterization of ambulances across the USA. Microbiome. 2017;5:125.
23. Suzuki TA, Worobey M. Geographical variation of human gut microbial composition. Biol Lett [Internet]. 2014 [cited 2019 Jun 20];10:20131037. Available from: http://dx.http//rsbl.royalsocietypublishing.org.
24. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods. 2010.
25. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods [Internet]. Nature Publishing Group; 2012 [cited 2019 Oct 9];9:811–4. Available from: http://www.nature.com/articles/nmeth.2066
26. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol [Internet]. 2014 [cited 2017 Jul 21];15:R46. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46
27. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci [Internet]. PeerJ Inc.; 2017 [cited 2017 Oct 3];3:e104. Available from: https://peerj.com/articles/cs-104
28. Paulson JN, Colin Stine O, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods [Internet]. Nature Publishing Group; 2013 [cited 2019 May 30];10:1200–2. Available from: http://www.nature.com/articles/nmeth.2658
29. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, et al. vegan: Community Ecology Package. R package version 2.5-4. Community Ecol Packag [Internet]. 2019; Available from:

https://cran.r-project.org/package=vegan

30. Paradis E, Schliep K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;
31. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw [Internet]. NIH Public Access; 2010 [cited 2019 May 30];33:1–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20808728
32. Zhang S, Li S, Gu W, Den Bakker H, Boxrud D, Taylor A, et al. Zoonotic source attribution of Salmonella enterica serotype typhimurium using genomic surveillance data, United States. Emerg Infect Dis. 2019;25:82–91.
33. Lan Y, Rosen G, Hershberg R. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. Microbiome [Internet]. BioMed Central; 2016 [cited 2019 May 30];4:18. Available from: http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-016-0162-5
34. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. 2015;
35. Laursen MF, Dalgaard MD, Bahl MI. Genomic GC-content affects the accuracy of 16S rRNA gene sequencing bsed microbial profiling due to PCR bias. Front Microbiol. 2017;
36. Gusareva ES, Acerbi E, Lau KJX, Luhung I, Premkrishnan BNV, Kolundzija S, et al. Microbial communities in the tropical air ecosystem follow a precise diel cycle. Proc Natl Acad Sci U S A. National Academy of Sciences; 2019;116:23299–308.
37. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. Nature [Internet]. Springer US; 2018;560:233–7. Available from: http://dx.doi.org/10.1038/s41586-018-0386-6
38. Bai Y, Rizk G, Klingenberg H, Quince C, Chia BKH, Fiedler J, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nat Methods [Internet]. 2017 [cited 2019 Feb 7];14:1063–71. Available from: https://www.nature.com/articles/nmeth.4458.pdf
39. Danko DC, Bezdan D, Afshinnekoo E, Ahsanuddin S, Alicea J, Bhattacharya C, et al. Global Genetic Cartography of Urban Metagenomes and Anti-Microbial Resistance. bioRxiv [Internet]. 2019 [cited 2019 Oct 31];724526. Available from: https://www.biorxiv.org/content/10.1101/724526v1
40. Choi I, Ponsero AJ, Bomhoff M, Youens-Clark K, Hartman JH, Hurwitz BL. Libra: Scalable k-mer-based tool for massive all-vs-All metagenome comparisons. Gigascience [Internet]. 2018 [cited 2020 Feb 6];8. Available from: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giy165/5266304
41. Vervier K, Mahé P, Vert JP. MetaVW: Large-scale machine learning for metagenomics sequence classification. Methods Mol Biol. Humana Press Inc.; 2018. p. 9–20.

Supplementary Files Legend

Figure S1. The world map labeled with training origins and mystery new origins (question mark)

Figure S2: Leave-one-city-out cross validation predictions from Lasso-regularized regression and classification compared to true locations.
Latitude (A) and longitude (B) predictions from the multivariate regression model on species abundance data are plotted against the true geographic coordinates on the x-axis. Each data point represents a sample from the corresponding city, as indicated in the legend. The dashed line shows where predictions would be exactly correct. (C) Predictions from the classification model are illustrated in comparison to true sources. Each entry shows the number of samples predicted to be the corresponding city (row) and originally from the corresponding source (column). As classification models can only assign new samples to pre-trained sources, diagonal counts are zero. Cities within the same continent are boxed in blue.

Figure S3. Out-of-bag and mystery sample prediction performance using random forest classification algorithm.

(A) Model performance as assessed from out-of-bag prediction. (B) Source prediction of mystery samples versus the true sources. Cities within the same continent w.r.t. reference are boxed in blue. Squared errors for latitude (C) and longitude (D) are shown in boxplots for each city.

Figure S4. Inclusion of training data from a heterogeneous sequencing protocol affects model performance.
This figure is analogous to Figure 5, with the distinction of including single-end data from Sacramento into training. Latitude (A) and longitude (B) predictions from the multivariate regression model on species abundance data are plotted against the true geographic coordinates on the x-axis. Each data point represents a sample. The dashed line shows where predictions would be exactly correct. (C) Predictions from the classification model are illustrated in comparison to true sources. Each entry shows the number of samples predicted to be the corresponding city (row) and originally from the corresponding source (column). Cities within the same continent w.r.t. reference are boxed in blue. Squared errors for latitude and longitude from the regression (D,E) and classification (F,G) are shown in boxplots for each city.

Table S1. Sequencing data information

Table S2. Mystery sample predictions

Figures

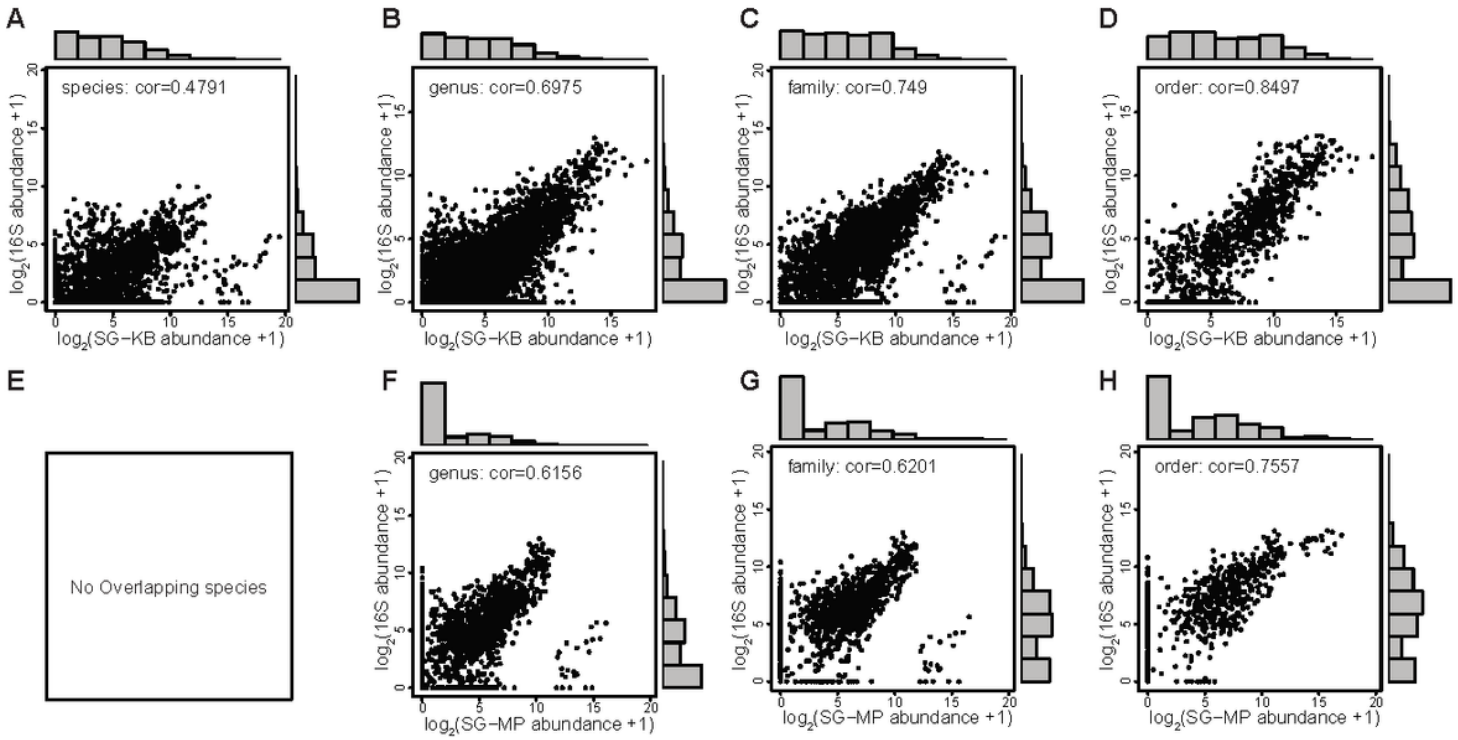

Figure 1

Log2-transformed abundance of taxa between Boston Pilot's 16S and shotgun data at different taxonomic scales. The y-axes represent the abundance from 16S amplicon data. (A-D) The x-axes represent the abundance extracted using Kraken2 and Bracken from shotgun data (SG-KB). (E-H) The x-axes represent the abundance extracted using MetaPhlAn2 from shotgun data (SG-MP). Histograms on top and right-hand panels show the distributions of shotgun data and 16S data, respectively. Correlations were reported using non-transformed abundance.

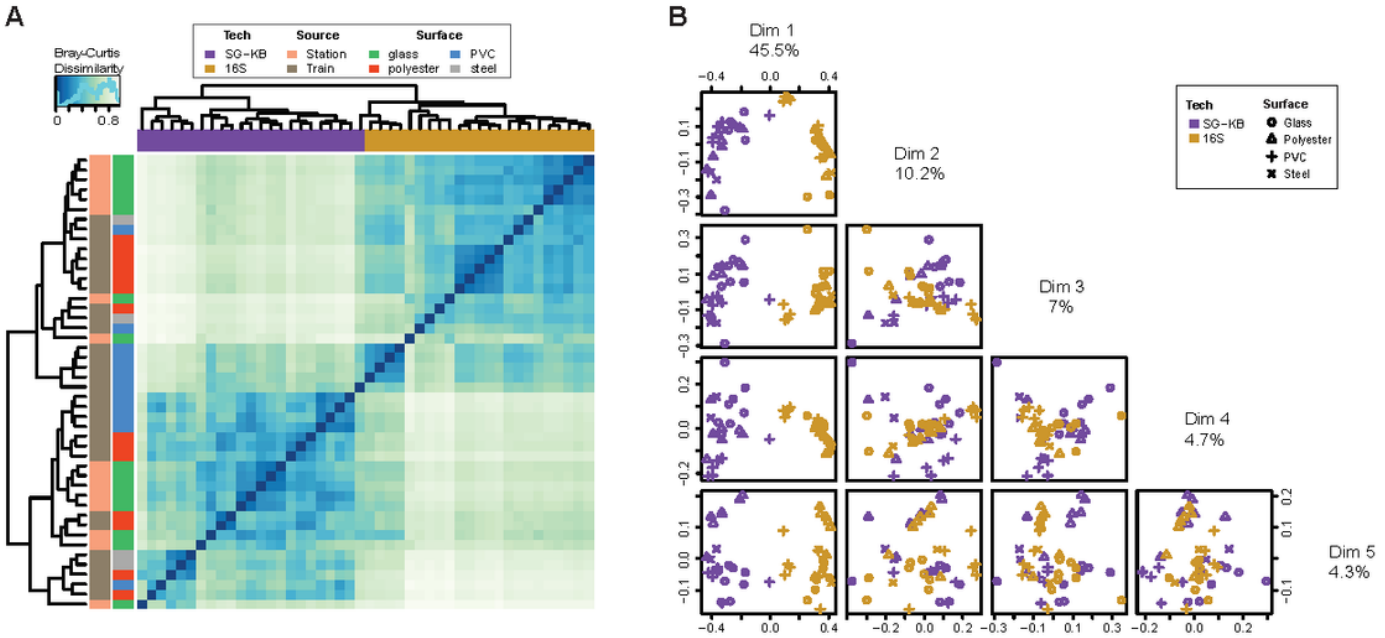

Figure 2

Relations between 16S rRNA amplicon and shotgun sequencing data generated from the same set of Boston pilot samples. (A) Heatmap visualization of the Bray-Curtis dissimilarity matrix computed from Genus abundance of 16S and shotgun (SG-KB) data (B) The first five dimensions of PCoA computed using Bray-Curtis dissimilarity

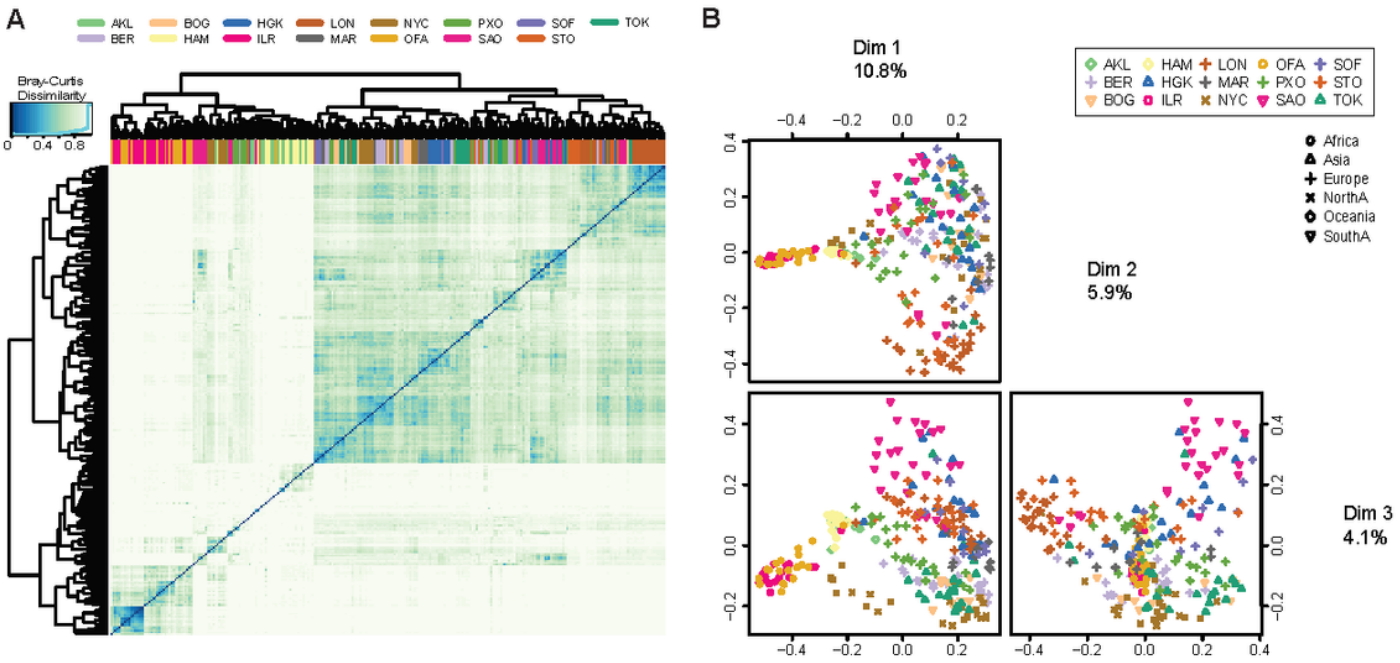

Figure 3

Relations between MetaSUB samples according to species abundance data. (A) Heatmap visualization of the Bray-Curtis dissimilarity matrix and and (B) The first three dimension of PCoA using Bray-Curtis dissimilarity

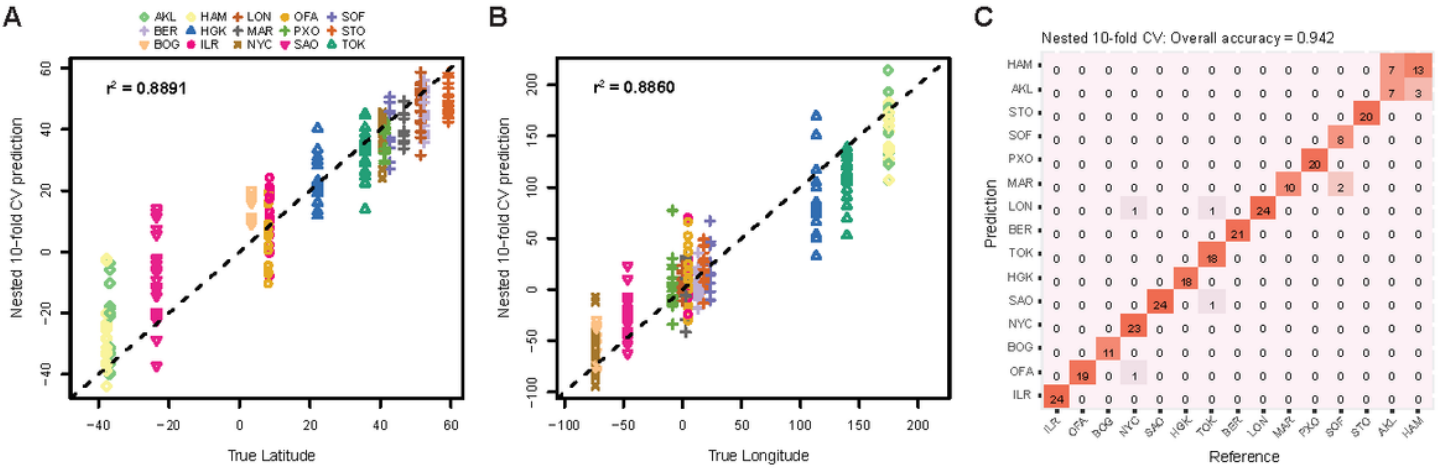

Figure 4

Nested 10-fold cross validation predictions from Lasso-regularized regression and classification compared to true locations. Latitude (A) and longitude (B) predictions from the multivariate regression model on species abundance data are plotted against the true geographic coordinates on the x-axis. Each data point represents a sample from the corresponding city, as indicated in the legend. The dashed line shows where predictions would be exactly correct. (C) Predictions from the classification model are illustrated in comparison to true sources. Each entry shows the number of samples predicted to be the corresponding city (row) and originally from the corresponding source (column).

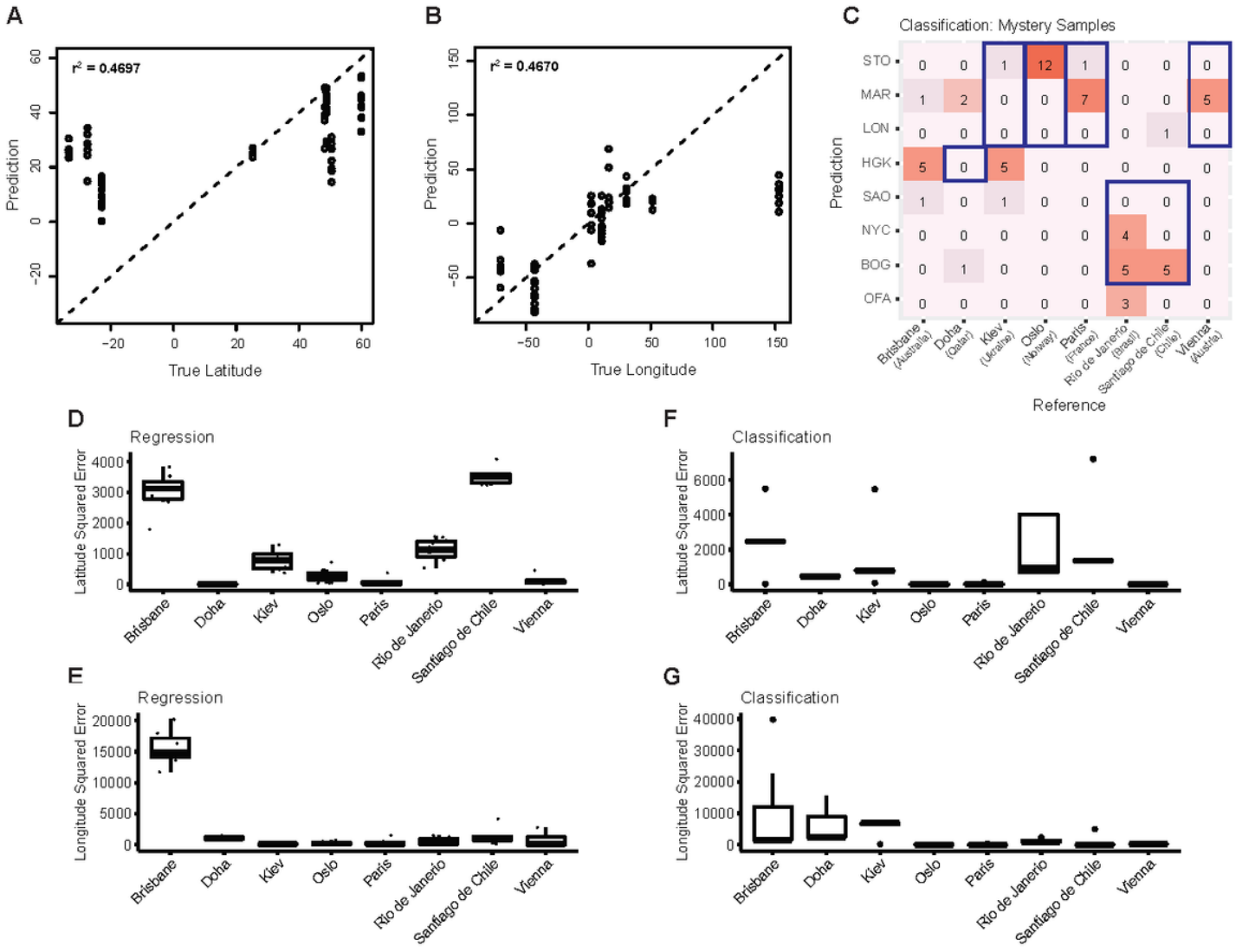

Figure 5

Source prediction of mystery samples originated from new cities using the species models. Latitude (A) and longitude (B) predictions from the multivariate regression model on species abundance data are plotted against the true geographic coordinates on the x-axis. Each data point represents a sample. The dashed line shows where predictions would be exactly correct. (C) Predictions from the classification model are illustrated in comparison to true sources. Each entry shows the number of samples predicted to be the corresponding city (row) and originally from the corresponding source (column). Cities within

the same continent w.r.t. reference are boxed in blue. Squared errors for latitude and longitude from the regression (D,E) and classification (F,G) are shown in boxplots for each city.

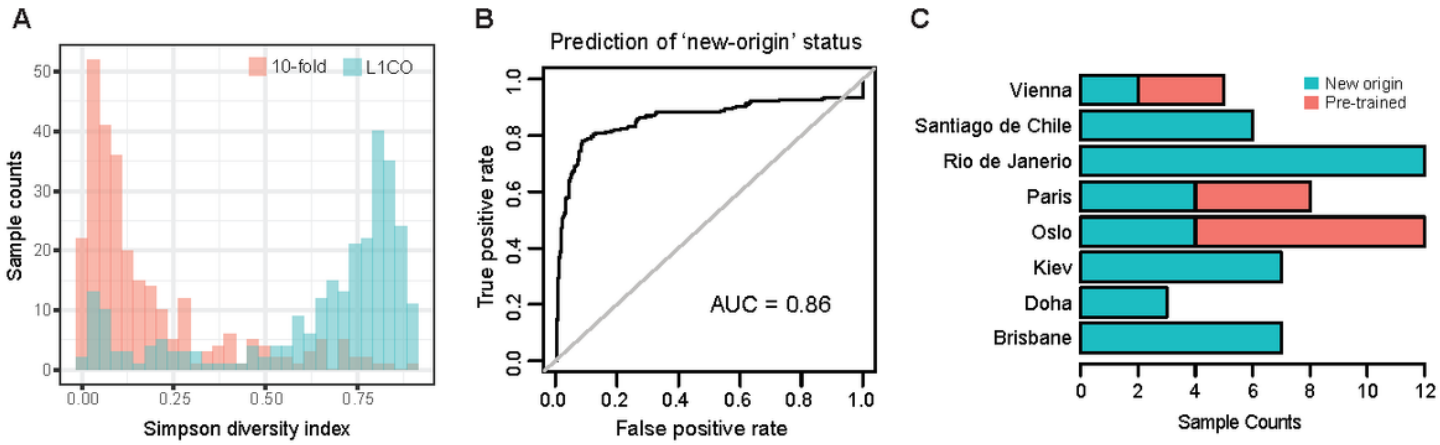

Figure 6

Ambiguity in classification prediction probabilities informs whether a sample is from a new origin. (A) The distributions of Simpson index on the class prediction probabilities of each sample based on 10-fold (red) and leave-one-city-out (blue) cross validation settings, which indicate the diversity pattern for samples from pre-trained or new origins, respectively. (B) The receiver operating characteristics curve of the Naïve Bayes model on predicting new-origin status using the Simpson index values computed through a leave-one-out design. (C) Prediction of new-origin status on mystery samples from cities that have not been pre-trained.

Supplementary Files

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

- TableS2.xlsx
- FigureS4new.pdf
- FigureS2new.pdf
- FigureS3newRF.pdf
- FigureS1new.pdf
- TableS1.xlsx