Unraveling bacterial fingerprints of city subways from microbiome 16S gene profiles

Alejandro R. Walker, Tyler Gymes, Somnath Datta, Susmita Datta

Department of Biostatistics, University of Florida, FL 32610, USA

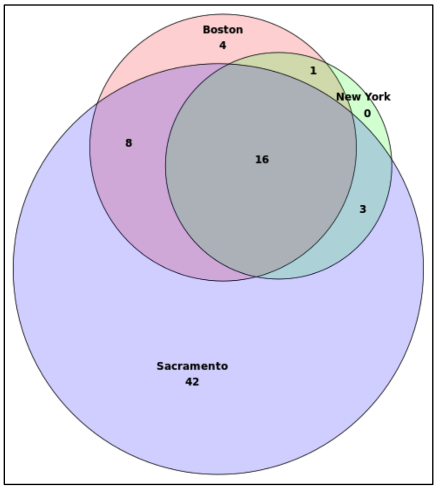
1 Introduction and a summary

The advent of NGS technologies had a tremendous effect on almost all imaginable scientific applications. Specially the reduction of costs over the years (Metzker 2010) has accelerated the use of this technology on metagenomics experiments (Simon and Daniel 2011, Thomas, Gilbert et al. 2012). Phylogenetic survey analyses based on 16S gene diversity has been fundamental on identification of bacterial varieties (Caporaso, Lauber et al. 2011, Kuczynski, Stombaugh et al. 2011, Clifford, Milillo et al. 2012). This sequencing revolution in conjunction with high performance computing and recently developed computing tools has had a tremendous impact on new 16S gene studies (Tringe and Hugenholtz 2008, Caporaso, Lauber et al. 2011). This analysis, as part of the 2017 CAMDA competition is focused on the MetaSUB Challenge dataset. We undertake a number of investigations for the OTU count data at the taxonomic level “Order” across the three cities. First, a PCA analysis showed a clear clustering of the data points for the three cities, where a large proportion of variability was explained by the first three principal components. Next we attempted to build a classifier using the OTU count data and were successful in achieving very high specificity and sensitivity. The relative abidance patterns of the OTUs varied significantly across the city, which was formally confirmed by an analysis of variance. Finally, we conduct a network analysis based on the co-abundance patterns the OTUs in a given city. Overall, we found finding different patterns in the three networks when inspected visually; the networks of close by cities showed similar bacterial co-abundance patterns compared to distant cities.

2 MetaSUB Organization Dataset

For the 2017 meeting, CAMDA has partnered with the MetaSUB (Metagenomics & Metadesign of Subways & Urban Biomes) International Consortium (http://metasub.org/), which has provided data from 3 cities across the United States as part of the MetaSUB Inter-City Challenge.

Figure 1: Summary of abundant OTUs found across all three cities and the relative contribution of each city reflected as the area of each circle.



Next generation sequencing data was generated from DNA samples taken on subway stations from Boston, New York and Sacramento in the form of FASTQ files for each sample from each city, plus a supplementary dataset with swab places, sequencing technology, DNA extraction and amplification, samples names, etc. A bioinformatics analysis with quality of the reads filtrations was conducted in order to improve OTU picking with QIIME (Caporaso, Kuczynski et al. 2010) and taxonomical classification and to shrink some large FASTQ files. The raw read counts, generated with QIIME, were grouped at the taxonomical level “Order” to generate a matrix of OTUs counts for the three cities. The number of distinct OTUs for each city including common OTUs is represented in a Venn diagram (Figure 1). The rest of the statistical analysis is carried out on the basis of 16 common OTUs finding additional patterns in the relative abundance distinguishing the cities that are not as obvious as the presence of city-specific OTUs. Other aspects of bio-diversity beyond what is immediate from Figure 1 (such that Sacramento samples exhibited most biodiversity) were not investigated further.

2.1 Boston Dataset

This dataset consisted of a total of 141 samples ranging from 1 Mbp to 11 Gbp single read Illumina data. The majority of the samples (117 Amplicon samples) were target sequenced after PCR amplification and the rest were whole genome shotgun (WGS) sequenced. Moreover, a small fraction of the amplicon samples did not effectively contribute to OTU counts, and hence they were removed from the analyses. Ultimately a total of 134 samples were included in further downstream analyses.

2.2 New York Dataset

A total 1,572 WGS samples were collected at New York ranging from 0 Mbp to 19 Gbp of Illumina sequence data. From the subset of samples, which contribute to the OTU counts, we randomly chose 280 samples in order to keep the computational burden in check.

2.3 Sacramento Dataset

Six locations were sampled three times each on different surfaces for a total of 18 sequenced samples ranging from 2.8 to 3.4 Gbp. All the samples contained enough sequencing data to positively contribute to OTU counts and therefore were included in all the analyses.

3 Statistical Analyses

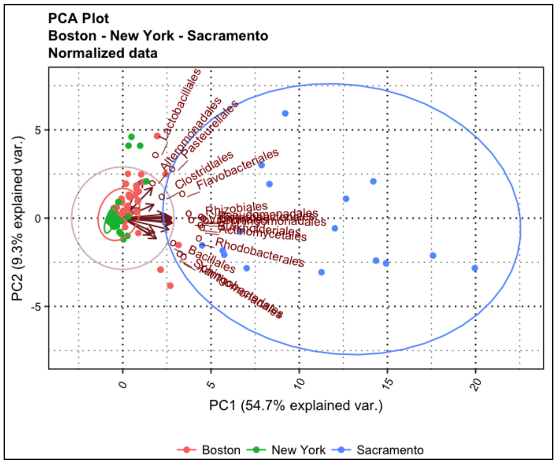
We used OTU counts for each city as starting point for the statistical analyses. The counts were then normalized to counts per million for each city before combining them in a single dataset (Formula 1) (Law, Chen et al. 2014).

Formula 1: OTU proportions calculated for each sample (p_{gi}), where r_{gi} is the gth OTU count for the ith sample. N is the number of OTU categories. R_i is the OTU mean of the ith sample

$$p_{gi} = log_2\left(\frac{r_{gi} + 0.5}{NR_i + 1}10^6\right)$$

The proceeding statistical analysis was done in multiple stages. The first was a PCA analysis, which also served as a proof that the normalized and transformed OTUs counts carry strong enough signals. The second was to build a statistical classifier, which can produce a well defined rule (e.g., a machine) in order to decode the city of origin from the OTU profiles of a sample. To this end, we used three well regarded classifiers, all within the R environment, and compared the findings. Finally, an association network analysis was conducted in order to assess how the OUT abundances vary jointly across the cities.

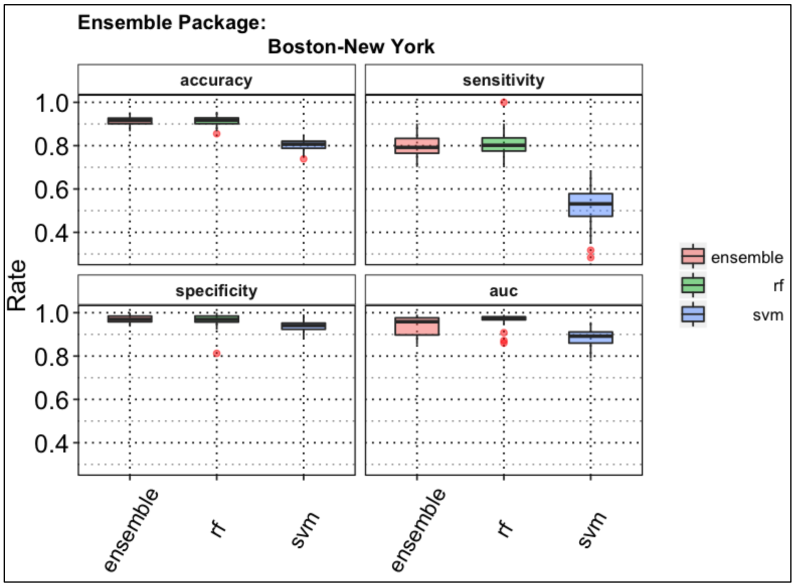
Figure 2: PCA Bi-plot of results showing clustering for Boston, New York and Sacramento normalized OTU counts per million and importance of taxonomical levels relative to the separation of the three cities.



3.1 PCA Analysis

Our PCA analysis shows that the first principal component is responsible for 54.7% of the total variation of the data and that the second component explains 9.3% of the remaining variation (see Figure 2). Additional review of the remaining eigenvalues revealed that PC3 is still highly relevant with 8.8% of the total variance explained (3D plot not shown). Figure 2 also shows a high correlation of 6 taxonomical orders with PC1 that have low to no correlation with PC2, which apparently makes them more important for Sacramento. With the exception of these orders, all other show similar contributions to both PC1 and PC2. As seen on Figure 2 the order Lactobacillales has the highest correlation with PC2 and an observable alignment with Boston and New York samples (both major ellipses axes), suggesting both cities share a common bacterial signature for the species of this "order". The relative importance of the order Lactobacillales signature for Boston and New York was also confirmed as a partial result from Random Forest classifier (result not shown in this extended abstract).

Figure 3: Ensemble Package results for Boston – New York pairwise classification. Accuracy, Sensitivity, Specificity and AUC for three classifiers; Ensemble Classifier, Random Forest and Support Vector Machine



3.2 Classification Analysis

Accurately predicting the origin of a sample on the basis of bacterial metagenomics in a robust fashion is our main objective of this work. We used three different classifiers to address this problem: *randomForest* (Breiman 2001), *ensemble* (Datta, Pihur et al. 2010), and the *support vector machine* (SVM) (Boser, Guyon et al. 1992).

The Random Forest (RF) classifier has improved classification accuracy as result of choosing vectors randomly and independently with a positive impact on the growth of each tree within the ensemble. This algorithm is robust to over-fitting, computationally efficient and calculates estimates for variable importance and internal error (Breiman 2001, Poona, van Niekerk et al. 2016). RF was implemented with 10 variables (OTUs), randomly chosen at each split, with 1000 threes. Results showed an estimated error rate of 6.93% on the classification of three cities which is rather remarkable given that we are not using any city specific OTUs (i.e., one that was not observed across all cities). The rates of partial classification errors for Boston, New York and Sacramento were 16.3%, 2.5% and 5.5%, respectively.

Next we describe the results of the *ensemble* classifier. As the name suggests, it is based on a number of individual (or component) classifiers. However, it is restricted to binary classifications, so we separated the dataset into three pairwise sets. For each pairwise comparison the analysis was conducted on a 2-fold training-test cross validation run with 100 iterations. The corresponding publicly available R-code internally compares results of the ensemble classifier (EC) with the component classification methods. In particular, both RF and SVM are amongst the component classifiers used within *ensemble.* For all three pairwise comparisons, EC and RF were highly effective in terms of overall accuracy, sensitivity, specificity and AUC (area under the curve); see Figure 3 for Boston versus New York. Results for Sacramento pairwise comparisons against Boston and New York showed that all these measures were even higher and close to the maximum for all three, especially, for RF and *ensemble*; the further details are not reported here.

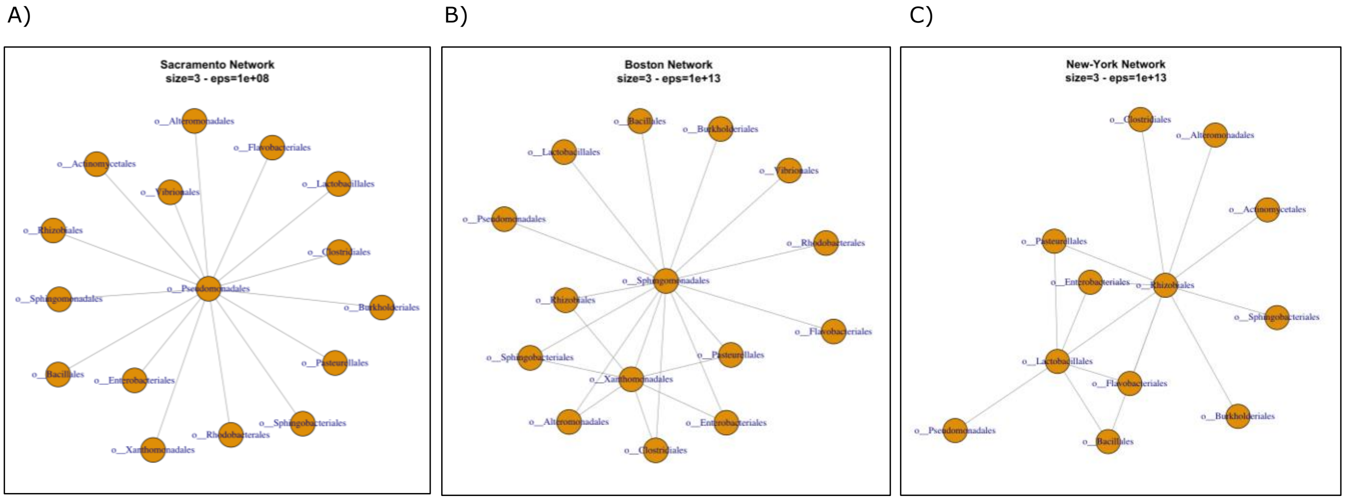
3.3 Differential abundance

Analysis of variance of normalized abundance for each city showed significant variation among cities (p-value<2.2e-16). Additionally, Tukey pairwise comparisons, showed significant differences for all three pairwise comparisons. Sacramento is the most significantly different city in terms of normalized OTUs as p-values of both comparisons with Boston and New York, respectively, were largely smaller than the p-value of Boston and New York test (p-value=0.0008). The ANOVA also showed a highly significant interaction between City and OTUs (p-value<2.2e-16).

3.4 Network Analysis

Network construction is often used in the context of gene-gene, gene-protein or protein-protein association/interaction networks (Gill, Datta et al. 2010). However, one may use the correlation of the transformed and normalized OTU counts to construct a "co-abundance" network. In this study, we applied Gill et al.'s strategy to identify differential structures and connectivity of bacterial fingerprints across three different cities. Networks are presented in Figure 4, purposely placed geographically from west on the left to the east on the right. Sacramento showed a centralized network centered on "order" Pseudomonadales, which morphs into a decentralized network for New York. In the middle Boston network as a transitional stage between both coastal cities. Further review of Boston and New York features revealed that the central "order" Sphingomonadales and Rhizobiales both belong to the same taxonomical "class" of Alphaproteobacteria, which might reflect that both cities share common bacterial fingerprint due to geographical proximity. On a close look at bottom-left nodes for the east coast cities networks there is a distinctive sub-cluster of orders that belong to the "class" Gammaproteobacteria, which is not shown on Sacramento since its centralized network doesn't allow these interconnected edges.

Figure 4: Visual networks discovered on three cities based on bacterial fingerprints from 16 common OTUs discovered across all cities. A) Sacramento, CA, B) Boston, MA and C) New York, NY



4 Discussion

It has been well established that WGS metagenomics can fail to detect rare species since DNA is not sequenced with high depth as result of its rarity (Kalyuzhnaya, Lapidus et al. 2008, Shah, Tang et al. 2011). Nevertheless, this was not an issue for the development of this work since our main objective was to determine the common bacterial composition of the three cities and use this data to predict the source of origin of a specific sample. Two of the three classifiers tested were highly effective in accurately predicting the source city from the samples in the two-class analysis.

It is fair to say that a variety of statistical and machine learning of methods showed distinct and consistent bacterial signatures amongst the three cities. It might be possible to develop these into an identification tool that may have applications in forensic science, among others.

This work also includes a novel application of network analysis in revealing city signatures which had explanations in terms of their geographical proximity.

References

Boser, B. E., I. M. Guyon and V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory. Pittsburgh, Pennsylvania, USA, ACM**:** 144-152.

Breiman, L. (2001). "Random forests." Machine Learning **45**(1): 5-32.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld and R. Knight (2010). "QIIME allows analysis of high-throughput community sequencing data." Nat Methods **7**(5): 335-336.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer and R. Knight (2011). "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample." Proc Natl Acad Sci U S A **108 Suppl 1**: 4516-4522.

Clifford, R. J., M. Milillo, J. Prestwood, R. Quintero, D. V. Zurawski, Y. I. Kwak, P. E. Waterman, E. P. Lesho and P. Mc Gann (2012). "Detection of Bacterial 16S rRNA and Identification of Four Clinically Important Bacteria by Real-Time PCR." Plos One **7**(11).

Datta, S., V. Pihur and S. Datta (2010). "An adaptive optimal ensemble classifier via bagging and rank aggregation with applications to high dimensional data." BMC Bioinformatics **11**: 427.

Gill, R., S. Datta and S. Datta (2010). "A statistical framework for differential network analysis from microarray data." Bmc Bioinformatics **11**.

Kalyuzhnaya, M. G., A. Lapidus, N. Ivanova, A. C. Copeland, A. C. McHardy, E. Szeto, A. Salamov, I. V. Grigoriev, D. Suciu, S. R. Levine, V. M. Markowitz, I. Rigoutsos, S. G. Tringe, D. C. Bruce, P. M. Richardson, M. E. Lidstrom and L. Chistoserdova (2008). "High-resolution metagenomics targets specific functional types in complex microbial communities." Nature Biotechnology **26**(9): 1029-1034.

Kuczynski, J., J. Stombaugh, W. A. Walters, A. Gonzalez, J. G. Caporaso and R. Knight (2011). "Using QIIME to analyze 16S rRNA gene sequences from microbial communities." Curr Protoc Bioinformatics **Chapter 10**: Unit 10 17.

Law, C. W., Y. S. Chen, W. Shi and G. K. Smyth (2014). "voom: precision weights unlock linear model analysis tools for RNA-seq read counts." Genome Biology **15**(2).

Metzker, M. L. (2010). "Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation." Nature Reviews Genetics **11**(1): 31-46.

Poona, N. K., A. van Niekerk, R. L. Nadel and R. Ismail (2016). "Random Forest (RF) Wrappers for Waveband Selection and Classification of Hyperspectral Data." Applied Spectroscopy **70**(2): 322-333.

Shah, N., H. Tang, T. G. Doak and Y. Ye (2011). "Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics." Pac Symp Biocomput: 165-176.

Simon, C. and R. Daniel (2011). "Metagenomic Analyses: Past and Future Trends." Applied and Environmental Microbiology **77**(4): 1153-1161.

Sokal, R. R. and P. H. A. Sneath (1963). Principles of numerical taxonomy, San Francisco and London, W. H. Freeman & Co.

Thomas, T., J. Gilbert and F. Meyer (2012). "Metagenomics - a guide from sampling to data analysis." Microb Inform Exp **2**(1): 3.

Tringe, S. G. and P. Hugenholtz (2008). "A renaissance for the pioneering 16S rRNA gene." Current Opinion in Microbiology **11**(5): 442-446.