COMPARING BACTERIAL COMMUNITIES INFERRED FROM 16S rRNA GENE SEQUENCING AND SHOTGUN METAGENOMICS

Neethu Shah [1], Haixu Tang [1], Thomas G. Doak [2] and Yuzhen Ye [1]

[1]: *School Of Informatics and Computing,* [2]*: Biology Department, Indiana University*
Bloomington, IN 47408, U.S.A
E-mails: neetshah{hatang, tdoak, yye}@indiana.edu

16S rRNA gene sequencing has been widely used for probing the species structure of a variety of environmental bacterial communities. Alternatively, 16S rRNA gene fragments can be retrieved from shotgun metagenomic sequences (metagenomes) and used for species profiling. Both approaches have their limitations—16S rRNA sequencing may be biased because of unequal amplification of species' 16S rRNA genes, whereas shotgun metagenomic sequencing may not be deep enough to detect the 16S rRNA genes of rare species in a community. However, previous studies showed that these two approaches give largely similar species profiles for a few bacterial communities. To investigate this problem in greater detail, we conducted a systematic comparison of these two approaches. We developed PHYLOSHOP, a pipeline that predicts 16S rRNA gene fragments in metagenomes, reports the taxonomic assignment of these fragments, and visualizes their taxonomy distribution. Using PHYLOSHOP, we analyzed 33 metagenomic datasets of human-associated bacterial communities, and compared the bacterial community structures derived from these metagenomic datasets with the community structure derived from 16S rRNA gene sequencing (71 datasets). Based on several statistical tests (including a statistical test proposed here that takes into consideration differences in sample size), we observed that these two approaches give significantly different community structures for nearly all the bacterial communities collected from different locations on and in human body, and that these differences cannot be be explained by differences in sample size and are likely to be attributed by experimental method.

Keywords: Bacterial community; 16S rRNA gene sequencing; shotgun metagenomics.

1. Introduction

Metagenomics is the study of microbial communities sampled directly from their natural environment, without prior culturing.[1] There has been remarkable progress in this field of research due to the recent advances of Next Generation Sequencing (NGS) technologies.[2] Since over 99.8% of the microbes in some environments cannot be cultured,[3] metagenomics offers a path to the study of their community structures, phylogenetic composition, species diversity, metabolic capacity, and functional diversity. A motivation for the field is medical: human microbial flora have long been recognized to be important to human disease and health, and the human gastrointenstinal tract is one of the most thoroughly surveyed bacterial ecosystems in nature,[4] although this ecosystem remains incompletely characterized and its diversity poorly defined.[5] It is essential to evaluate not only the species diversity of microbial communities but also to analyze how the species structures of those communities change over time and space.[6] The National Institute of Health has initiated the Human Microbiome Project (HMP) with the mission of generating resources enabling comprehensive characterization of the human microbiota and the analysis of its role in human health and disease (http://nihroadmap.nih.gov/hmp/).[7]

16S rRNA gene profiling has been applied to the analysis of the genetic diversity of com-

plex bacterial populations since the middle 1990s,[8] and is one of the primary steps in any metagenomics project. The application of 16S rRNA profiling has recently been boosted by advances in DNA sequencing techniques and the application of barcoded pyrosequencing.[9] NGS technologies—including 454 and Illumina sequencers—use 16S rRNA amplification primers targetting hypervariable regions, although it is still arguable which regions are best for species profiling: 16S rRNA genes contain nine hypervariable regions (V1–V9) that demonstrate considerable and differential sequence diversity among different bacteria. Although no single hypervariable region is able to distinguish among all the bacteria,[10] hypervariable regions V2 (nuceotides 137–242), V3 (nucleotides 433-497) and V6 (nucleotides 986–1043) contain the maximum heterogeneity and provide the maximum discriminating power for analyzing bacterial groups[10] . Barcoded pyrosequencing can produce large 16S rRNA datasets that contain hundreds of thousands of 16S RNAs fragments,[11] enabling deep views into hundreds of bacterial communities simultaneously, and have revealed much greater species diversity in many environments (e.g., soil, ocean water, and human bodies) than previously anticipated.

16S rRNA based analysis of metagenomic samples is complicated by several artifacts, including chimeric sequences caused by PCR amplification and sequencing errors.[12] According to a study by Ashelford K.E *et al*, at least 1 in 20 16S rRNA sequences currently in public repositories contains substantial anomalies,[13] and it was shown in one study[12] that some metagenomics projects may overestimate the species diversity because of the presence of sequencing errors and chimeric sequences.

Whole genome shotgun (WGS) sequencing of environmental DNA can also be used to study the species composition and diversity of natural bacterial communities,[14–16] and an increasing numbers of shotgun metagenomic sequencing datasets have been produced for various bacterial communities. Although shotgun metagenomic sequencing does not involve the biased amplification of 16S rRNA genes, the relative organism abundances inferred from metagenomic sequences vary significantly depending on the DNA extraction and sequencing protocol utilized.[17] Furthermore, shotgun metagenomic sequencing is generally not deep enough to detect rare species in complex communities.[18] Still, previous studies have shown that these two approaches give largely similar (although not identical in detail) pictures of the species structure for bacterial communities; for instance, Kalyuzhnaya et al[18] reported that the taxonomic distribution of 16S rRNA gene sequences derived from metagenomes is similar to distributions inferred from PCR-amplified libraries.[19]

Here we carry out a systematic comparison of these two approaches. We developed PHYLOSHOP, a pipeline that extracts 16S rRNA gene fragments from metagenomic sequences, reports the taxonomic assignment of the identified 16S rRNA fragments, and visualizes the taxonomy distribution. The bacterial community of a sample inferred from the identified 16S rRNA gene fragments can then be compared to the community derived from 16S rRNA gene sequencing, using the UniFrac metric,[20] which measures the phylogenetic distance between two sets of taxa, one for each community, on a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from one environment or the other. For a group of communities, a matrix of pairwise UniFrac measures can be prepared, and further subjected to Principal Coordinates Analysis (PCoA, a multivariate method that represents distance, or

similarity measures, in the space of principal coordinates)[21] to study the relationship between communities. We used a P test, a commonly used phylogenetic approach to assess community differentiation.[20,22] For a given set of sequences sampled from multiple communitites, the P test estimates the minimum number of changes (switch from one community to another) required to explain the observed species distribution, and computes the significance of the difference by determining the expected number of changes, under the null hypothesis that the communities from which the sequences are sampled do not covary with phylogeny.[22] Since a significantly smaller number of 16S rRNA gene fragments can be extracted from metagenomic datasets, as compared to a 16S rRNA gene sequencing project, we also propose a new statistical test for comparing the community diversities that are inferred from collections of 16S rRNA gene fragments with vastly different numbers.

2. Methods

2.1. *PHYLOSHOP: extracting and annotating 16S rRNA gene fragments from metagenomes*

The PHYLOSHOP pipeline (Figure 1) includes the following steps.

(a) 16S rRNA sequence calling. If the given sequences are metagenomic sequences, 16S rRNA gene fragments are predicted by a HMMER search (see 2.1.1).[23,24]
(b) Chimeric sequence checking. 16S rRNA gene fragments are examined for chimeras using ChimeraSlayer and putative chimeras are removed (see 2.1.2)
(c) Mapping of 16S rRNA gene fragments. Filtered 16Sr RNA gene fragments are mapped to a phylogenetic tree of the Greengenes[25] core set of 4,938 16S rRNA genes (as of May 2009) using MEGABLAST (with a default E-valule cutoff of 1e-30). The tree and the sequences of the core set were downloaded from the Fast UniFrac website (`http://128.138.212.43/fastunifrac`). The taxonomic assignments of the core set sequences were obtained from the Greengenes website (`http://greengenes.lbl.gov`).
(d) Taxonomic assignment of 16S rRNA gene fragments. PHYLOSHOP classifies the 16S rRNA gene fragments based on their mapping to the phylogenetic tree of 16S rRNA genes.

2.1.1. *16S rRNA gene fragment prediction*

We used the bacterial 16S rRNA Hidden Markov Model (HMM) of Huang et al[23] (downloaded from `http://tools.camera.calit2.net/camera/meta_rna/`), which was constructed from 16S rRNA sequences in the European rRNA database. 16S rRNA gene fragments can then be predicted using HMM scanner (HMMER 3.0 package[26]) against a dataset of metagenomic sequences.

2.1.2. *Checking chimeric sequences*

ChimeraSlayer (`http://microbiomeutil.sourceforge.net/`) is included in PHYLOSHOP for detecting chimeric sequences in the samples used for this analysis. As chimeric sequence

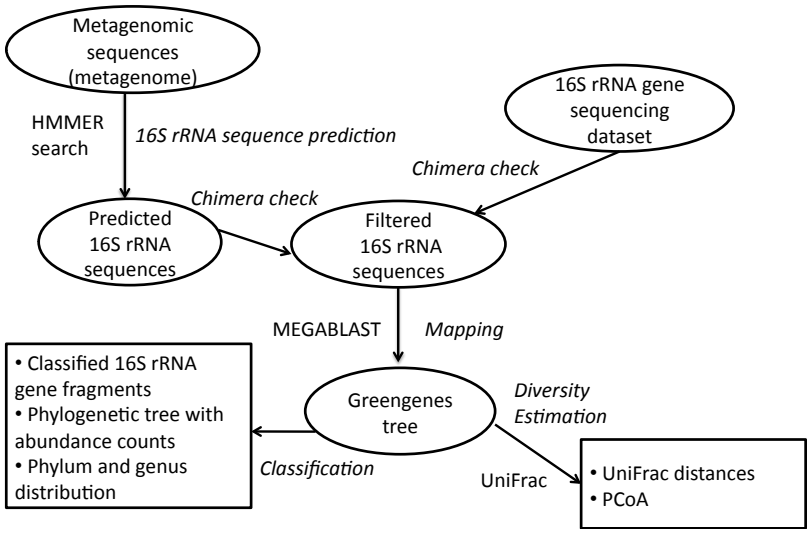


Fig. 1. Schematic representation of the PHYLOSHOP pipeline.

checkers do not work with very short reads (e.g., 100 bps), this option is only available for relatively long 16S rRNA gene fragments.

2.1.3. *PHYLOSHOP output*

PHYLOSHOP reports the following results, summarizing the taxonomic assignments of 16S rRNA sequences at different phylogenetic levels.

(a) Extracted 16S rRNA gene fragments, if the input is a metagenome in FASTA format.
(b) Classified 16S rRNA sequences, with an option for the user to choose the taxonomy systems—RDP,[27] NCBI or Hugenholtz.[28]
(c) Length distribution of the 16S rRNA sequences classified/extracted in a png figure.
(d) Phylum and genus disribution of the sequences mapped to the Greengenes tree.
(e) Rooted and unrooted trees in png format, showing the number of reads mapped to each identified species.

2.2. *Comparison of bacterial communities*

We used Fast UniFrac[6] to compare the structure and composition of bacterial communities.

2.3. *Statistical test of community structure differences by sampling*

A typical 16S rRNA gene sequencing dataset contains many more 16S rRNA gene fragments than those retrieved from a metagenome, so it is necessary to devise a measure that can be used to test if the observed difference in species structure between bacterial communities is statistically different, or if the difference is more likely to be caused by the dramatic difference in the numbers of 16S rRNA fragments used for inferring the bacterial communities. We propose a significance test based on multiple random sampling of subsets of 16S rRNA sequences

from the larger 16S rRNA dataset. Assume there is a sample that has both a metagenomic and a 16S rRNA sequencing dataset. From the shotgun metagenomic dataset, we extract 16S rRNA gene fragments and infer the bacterial community (denote as community-m). Denote the community inferred from the 16S rRNA sequencing dataset as community-s0. From the 16S rRNA sequencing dataset, we generate n subsets of 16S rRNA sequences by random sampling and the inferred bacterial communities are denoted as community-s1, community-s2, and so on. We use the UniFrac metric to define the distance between two communities; denote the UniFrac distance between community-m and community-s0 as d_0, and the distance between the community-m and simulated community-s1, ..., community-sn as d_1, d_2, ..., and d_n. We define the P-value of the difference between the communities inferred from metagenomic sequences and from 16S rRNA sequencing dataset as the fraction of random sampling experiments that result in distance larger than d_0; this value can then be used to evaluate the significance of observed community differences, when comparing communities that have been characterized by separate methods.

2.4. *Data sets*

We analyzed 104 datasets, including 33 (32 gut and 1 oral) shotgun metagenomic datasets and 71 (42 gut and 29 oral) 16S rRNA sequencing datasets of human-associated bacterial communities; see Supplementary Tables 1–4 for the details of the datasets. The sequences were downloaded from CAMERA (`http://camera.calit2.net/`),[29] NIH Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra`), and MG–RAST (`http://metagenomics.nmpdr.org/`).[30] Among these datasets, the twin study[16] has sequence datasets from both techniques—shotgun and 16S rRNA sequencing—for 18 individuals (see Table 1).

3. Results

Using PHYLOSHOP, we analyzed 33 metagenomic datasets of human-associated bacterial communities. We further compared the bacterial community structures derived from these metagenomic datasets to community structures inferred from 16S rRNA sequencing datasets, and observed clear differences in the inferred species structures associated with the different approaches (shotgun metagenomics versus 16S rRNA gene sequencing), in addition to the differences due to the different human body locations from which the samples were collected.

3.1. *Evaluation of 16S rRNA gene fragment prediction*

We first need to predict 16S rRNA gene fragments from metagneomic datasets. We compared the performance of 16S rRNA gene prediction by HMMER search[23] (implemented in the PHYLOSHOP pipeline) to predictions from the MG-RAST server, which uses BLAST searches against the Greengenes sequences. The comparison shows that HMMER searchs predicted slightly more 16S rRNA gene fragments in 11 out of the 17 metagenomic datasets shown in Figure 2. The difference is not significant, but considering that the HMMER search method is efficient and has shown high specificity and sensitivity in predicting 16S rRNA gene fragments,[23] we chose to use this method in the PHYLOSHOP pipeline. We then used 16S rRNA gene predictions from the PHYLOSHOP pipeline for the following analysis.

Table 1. Summary of the 18 gut samples that have both metagenomic datasets and 16S rRNA sequencing datasets.

Individuals	Metagenomic datasets			16S rRNA datasets		Significantly different?[d]	
	Reads[a]	Length[b]	16S rRNA[c]	Reads	Length	P test[e]	Our method[f]
TS1	217,386	238	464	25,140	126	Yes	Yes
TS2	443,526	178	658	42,186	126	Yes	Yes
TS3	510,972	201	871	17,726	126	Yes	Yes
TS4	414,754	229	731	25,705	126	Yes	Yes
TS5	490,776	205	1,108	26,608	126	Yes	Yes
TS6	535,763	221	1,207	27,007	126	Yes	Yes
TS7	555,853	243	1,310	17,469	126	Yes	Yes
TS8	414,497	243	1,036	17,170	126	Yes	Yes
TS9	499,499	250	1,024	14,787	126	Yes	Yes
TS19	498,880	165	767	43,639	126	Yes	Yes
TS20	495,039	198	1045	13,476	126	Yes	Yes
TS21	413,772	215	905	23,714	126	Yes	Yes
TS28	302,772	335	734	20,905	126	Yes	Yes
TS29	502,399	345	1,301	15,698	126	Yes	Yes
TS30	495,865	190	961	32,083	126	Yes	Yes
TS49	519,072	177	1,028	22,201	126	Yes	Yes
TS50	549,700	204	1,446	30,498	126	Yes	Yes
TS51	434,187	187	756	22,691	126	Yes	Yes

[a]: the total number of reads. [b]: the average length of reads. [c]: the total number of 16S rRNA gene fragments extracted from the metagenomic datasets. [d]: statistical significance of the difference between two communities, one inferred from the 16S rRNA sequencing dataset, and the other from the metagenomic dataset for the same individual. [e]: P-values for the P test[22] (computed using the Fast UniFrac website) are 0 for all the 18 individuals. [f]: P-values (computed using our method; see section 2.3) are $<$ 1e-4 for all the 18 individuals, based on 10,000 sampling experiments for each.

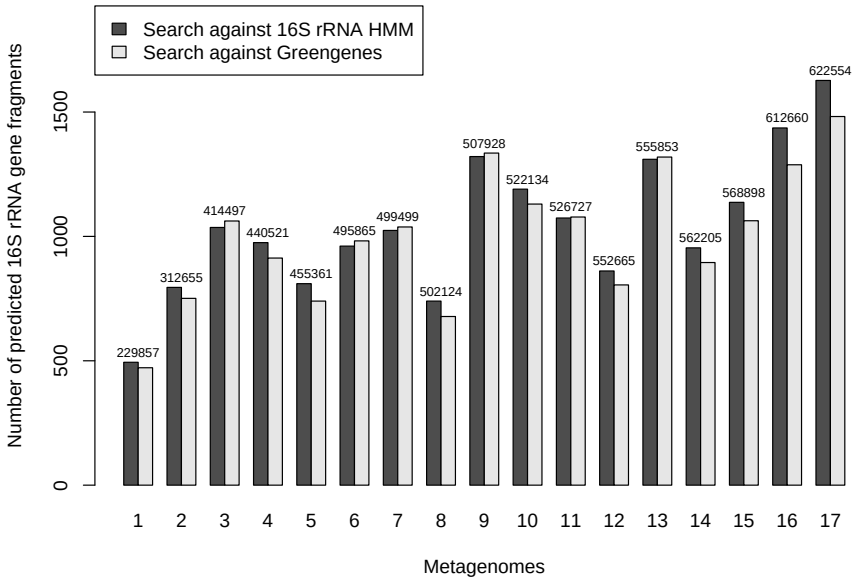
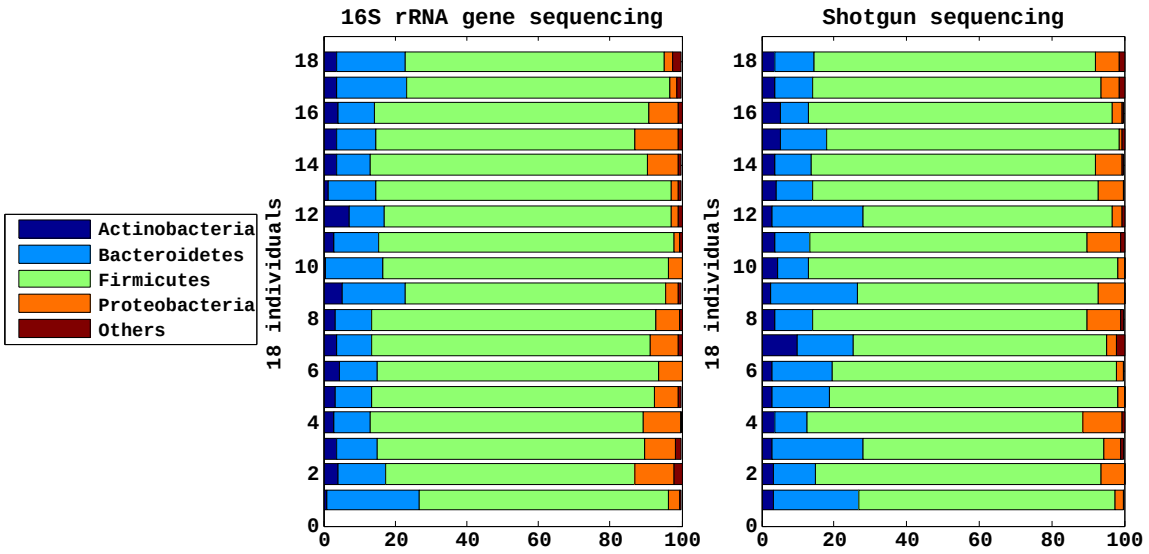
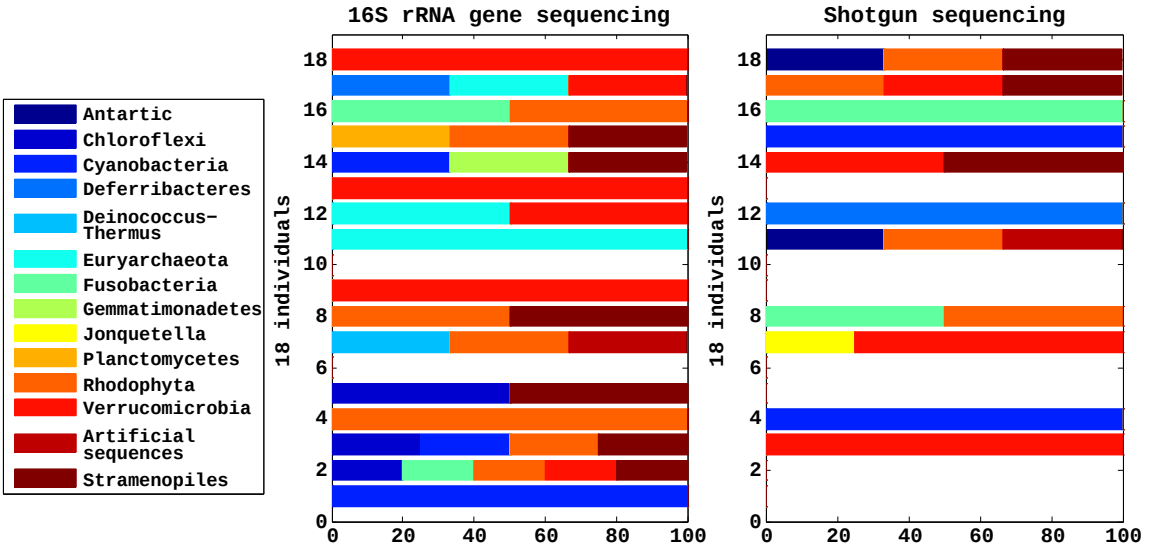


Fig. 2. Comparison of 16S rRNA prediction methods. The number of reads in each metagenome is shown above the corresponding bars.



(a)



(b)

Fig. 3. Phylum distributions of 18 gut-associated bacterial communities, inferred from 16S rRNA gene sequencing and shotgun metagenomics, in the four major (a) and other phyla (b). X-axis shows the percentage, and the phylum distribution for each individual is shown as a horizontal bar in each plot. Note that some communities (e.g., the communities in individual 6) have no reads assigned to the minor phya. The NCBI taxonomy was used, and reads assigned to "Unclassified" taxa were excluded in this analysis.

3.2. *16S rRNA gene sequencing reveals more species*

We analyzed the bacterial communities inferred from the 18 gut-associated individuals (see Table 1) that have both shotgun metagenomic and 16S rRNA gene sequencing datasets. Phylogenetic distributions of these samples show that there are clear differences in the relative abun-

dance of the four major phylum (Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria) (Figure 3a); e.g., for individual 12 (TS21), the 16S rRNA gene sequencing dataset contains more reads from Firmicutes as compared to the metagenomic dataset. Figure 3b shows that, for most of the individuals, 16S rRNA sequencing also reveals more diverse phyla than whole genome shotgun sequencing. 16S rRNA sequencing data also found a greater diversity within genera; e.g., 35 Firmicutes genera were identified by 16S rRNA sequencing reads, whereas only 22 genera were identified by metagenomics for individual TS1 (see Supplementary Figure 1).

3.3. *Bacterial communities inferred from metagenomes are different from those inferred from 16S rRNA sequencing reads*

P tests for the 18 gut-associated samples show that, for each of these samples, the bacterial communities inferred from the metagenome and from the corresponding 16S rRNA sequencing dataset are significantly different (see Table 1). Our sampling-based tests showed similar results—the difference between the inferred communities can not be explained by the different numbers of 16S rRNA sequences. Here we use individual TS50 as an example. The TS50 metagenome includes 549,700 reads with 1,446 16S rRNA gene fragments, while its 16S rRNA gene sequencing dataset contains 30,498 16S rRNA gene fragments. The UniFrac distance (weighted) between the communities inferred from the two methods is 0.164. We simulated 10,000 subsets of 16S rRNA gene fragments from the 16S rRNA gene sequencing dataset, each containing the same number of 16S rRNA gene sequences as in the metagenome, and computed the community distances between the sampled subsets and the complete 16S rRNA gene sequencing dataset. The species structures inferred from these sampled subsets are all significantly more similar to the structure inferred from the complete 16S rRNA gene sequencing dataset (with an average UniFrac distance of 0.021; see Figure 4 for the distribution of the distances) than the complete data set is to the metagenomic dataset (0.164).

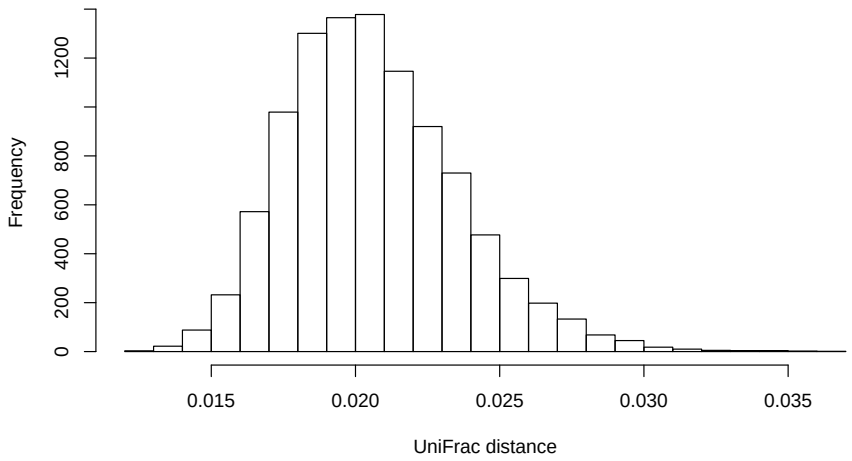


Fig. 4. Distribution of the UniFrac distance between a subset and the complete set of 16S rRNA sequencing data for the TS50 sample, based on 10,000 sampling experiments.

3.4. *Both body location and experimental technique matter*

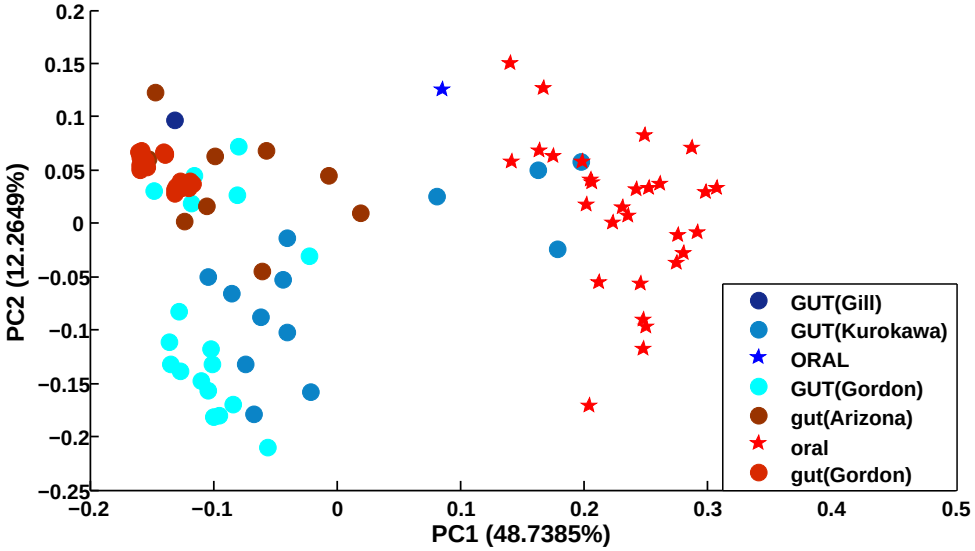
We further analyzed and compared 104 bacterial communities for different body sites inferred from metagenomes and 16S rRNA sequences, using PCoA. All of the 16S rRNA sequences (from the 16S rRNA gene sequencing, and extracted from metagenomes) were mapped to the phylogenetic tree of the core gene set of Greengenes to derive phylogenetic distributions of 16S rRNA sequences, from which UniFrac distances between any two communities can be computed. We used both weighted and unweighted UniFrac distances (weighted UniFrac weights the branches based on abundance information)[20] to derive UniFrac distance matrices. The PCoA results of the two matrices (Figure 5) show that there are at least two factors that affect community clustering: the body location, and the experimental method. The separation of the communities by experimental technique is more prominent when unweighted UniFrac distances were used (Figure 5b). For example, gut samples derived from 16S rRNA gene sequencing and whole genome shotgun sequencing (note there are 18 gut samples that have both, see Table 1) are far away from each other in the two-dimensional projection of the communities.

4. Discussion

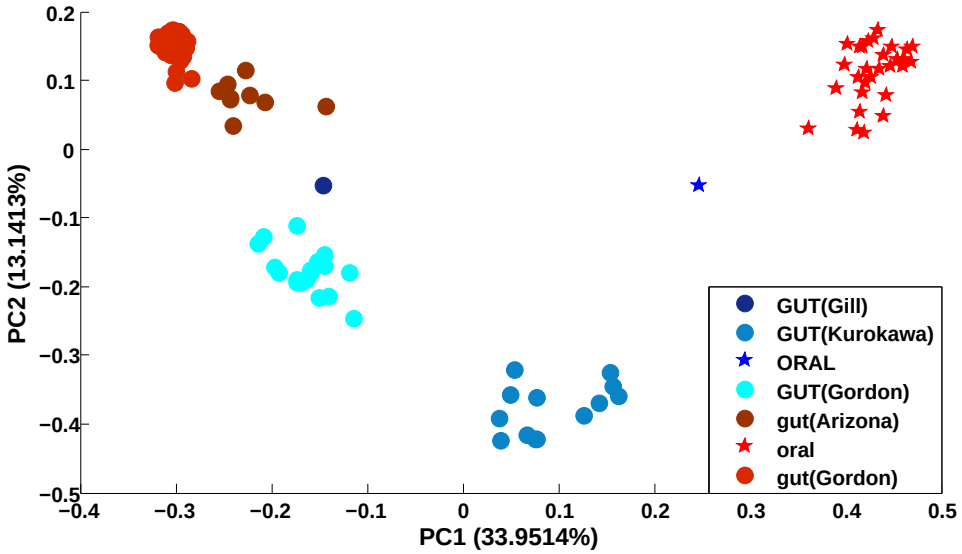
Our comparative studies revealed significant differences in the bacterial diversities derived from 16S rRNA gene sequencing and whole genome shotgun sequencing (metagenomics) of the same sample. These differences are not due simply to the different depths of sampling in the two methods, and indicate that 16S rRNA gene sequencing can profile the bacterial communities in a greater detail than can metagenomics. Our results indicate that even when corrected for depth, conclusions derived from 16S rRNA gene sequencing and shotgun metagenome sequencing cannot be directly compared. In addition, low abundance species are best identified through 16S rRNA gene sequencing.

There can be other factors that cause the differences observed between bacterial communities inferred from 16S rRNA gene sequencing and metagenomics. For example, the 16S rRNA gene fragments derived from metagenomic datasets may cover different regions as compared to the 16S rRNA gene fragments from PCA-based pyrosequencing (which often targets a certain region of 16S rRNA gene). And it has been shown that different regions of 16S rRNA gene have different sequence diversity,[10] and therefore a certain region may serve well for profiling a certain spectrum of bacteria but not all. Ideally we could do the comparison using only the 16S rRNA gene fragments that cover the same region, but we were only able to extract a small number of such 16S rRNA gene fragments from the metagenomic datasets we tested. When bigger metagenomic datasets become available, it will be interesting and necessary to do such a comparison, using the fragments spanning the same region of 16S rRNA gene derived from different experimental techniques.

We focused on bacterial communities in this paper, but the PHYLOSHOP pipeline can easily be extended by incorporating HMMs of other phyla's RNA genes, such as archarea or fungi. Notably, the reference tree in this analysis contains only the core set of Greengenes 16sRNA genes, and thus can be further refined. Finally, the rapidly growing numbers of metagenomic samples in the public domain will provide a more comprehensive resource to



(a)



(b)

Fig. 5. Two-dimensional projection of metagenomic samples by using PCoA of the weighted (a) and unweighted (b) UniFrac distance matrices of their bacterial communities. The labels of the samples indicate the source (gut or oral), the research group involved (Gordon,[16] Gill,[31] and Kurokawa[32]), and the technique that was used (shotgun metagenomics in capital letters, and 16S rRNA gene sequencing in lower case letters). For instance, GUT (Gordon) and gut (Gordon) represent gut-associated metagenome and 16S rRNA datasets, respectively, which were produced from the same research lab. The gut (Arizona) datasets were downloaded from the NIH SRA website (accession number: SRP001377).

conduct our analysis more thoroughly and elaborately, but we suggest that for the foreseeable future metagenomic projects should be paired with 16S rRNA gene sequencing.

5. Availability

PHYLOSHOP is implemented in Python and the source codes are available for download at http://omics.informatics.indiana.edu/mg/phyloshop. The supplementary tables and figure also available in the same website.

6. Acknowledgments

The authors thank Dr. Sun Kim for useful suggestions, and thank the anonymous reviewers for valuable comments. This work was supported by National Institutes of Health grant 1R01HG004908-02.

References

1. J. Wooley and Y. Ye, *Journal of Computer Science and Technology* **25**, 71 (2010).
2. S. C. Schuster, *Nat. Methods* **5**, 16 (2008).
3. W. Streit and A. Schmitz, *Current Opinion in Microbiology* **7**, 492 (2004).
4. J. Xu, M. Mahowald, R. Ley, C. Lozupone, M. Hamady, E. Martens, B. Henrissat, P. Coutinho, P. Minx, P. Latreille, H. Cordum, A. Brunt, K. Kim, R. Fulton, L. Fulton, A. Clifton, R. Wilson, R. Knight and J. Gordon, *PLoS Biology* **5**, 1574 (2007).
5. L. Hooper and J. Gordon, *Science* **292**, 1115 (2001).
6. M. Hamady, C. Lozupone and R. Knight, *The ISME Journal* **4**, 17 (2010).
7. J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, C. C. Baker, V. Di Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A. R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson and M. Guyer, *Genome Res.* **19**, 2317 (2009).
8. G. Muyzer, E. C. de Waal and A. G. Uitterlinden, *Appl. Environ. Microbiol.* **59**, 695 (1993).
9. M. Hamady, J. J. Walker, J. K. Harris, N. J. Gold and R. Knight, *Nat. Methods* **5**, 235 (2008).
10. S. Chakravorty, D. Helb, M. Burday, N. Connel and D. Alland, *Journal of Microbiology Methods* **69**, 330 (2007).
11. E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon and R. Knight, *Science* **326**, 1694 (2009).
12. C. Quince, A. Lanzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read and W. T. Sloan, *Nat. Methods* **6**, 639 (2009).
13. K. Ashelford, N. Chuzhanova, J. Fry, A. Jones and A. Weightman, *Applied and Environmental Microbiology* **71**, 7724 (2005).
14. G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar and J. F. Banfield, *Nature* **428**, 37 (2004).
15. S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz and E. M. Rubin, *Science* **308**, 554 (2005).
16. P. Turnbaugh, M. Hamady, T. Yatsunenko, B. Cantarel, A. Duncan, R. Ley, M. Sogin, W. Jones, B. Roe, J. Affourtit, M. Egholm, B. Henrissat, A. Heath, R. Knight and J. Gordon, *Nature* **457**, 480 (2009).

17. J. L. Morgan, A. E. Darling and J. A. Eisen, *PLoS ONE* **5**, p. e10209 (2010).
18. M. G. Kalyuzhnaya, 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, *Nat. Biotechnol.* **26**, 1029 (2008).
19. M. G. Kalyuzhnaya, M. E. Lidstrom and L. Chistoserdova, *ISME J* **2**, 696 (2008).
20. C. Lozupone and R. Knight, *Applied and Environmental Microbiology* **71**, 8228 (2005).
21. W. Krzanowski, *Principles of multivariate analysis. A user's perspective.* (Oxford University Press, Oxford, United Kingdom, 2000).
22. A. Martin, *Applied and Environmental Microbiology* **68**, 3673 (2002).
23. Y. Huang, P. Gilna and W. Li, *Bioinformatics* **25**, 1338 (2009).
24. R. Durbin, S. Eddy and A. Krogh, *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* (Cambridge University Press, 1999).
25. T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. Andersen, *Applied and Environmental Microbiology* **75**, 5069 (2006).
26. S. Eddy, *Genome informatics. International Conference on Genome Informatics* **23**, 205 (2009).
27. Q. Wang, G. M. Garrity, J. M. Tiedje and J. R. Cole, *Appl. Environ. Microbiol.* **73**, 5261 (2007).
28. A. C. McHardy, H. G. Martin, A. Tsirigos, P. Hugenholtz and I. Rigoutsos, *Nat. Methods* **4**, 63 (2007).
29. R. Seshadri, S. Kravitz, L. Smarr, P. Gilna and M. Frazier, *PLoS Biology* **5** (2007).
30. F. Meyer, D. Paarmann, M. DŚouza, R. Olson, E. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening and R. Edwards, *BMC Bioinformatics* **9** (2008).
31. S. Gill, M. Pop, R. DeBoy, P. Eckburg, P. Turnbaugh, B. Samuel, J. Gordon, D. Relman, C. Fraser-Liggett and K. Nelson, *Science* **312**, 1355 (2006).
32. K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. Sharma, T. Srivastava, T. Taylor, H. Noguchi, H. Mori, Y. Ogura, D. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi and M. Hattori, *DNA Research* **14**, 169 (2006).