



Object data analysis of taxonomic trees from human microbiome data

Journal:	<i>Statistics in Medicine</i>
Manuscript ID:	SIM-11-0395.R1
Wiley - Manuscript type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	La Rosa, Patricio; Washington University School of Medicine, Medicine Shands, Berkley; Washington University School of Medicine, Medicine Deych, Elena; Washington University School of Medicine, Medicine Zhou, Yanjiao; Washington University School of Medicien, Medicine; Washington University School of Medicien, Genome Institute Sodergren, Erica; Washington University School of Medicine, Genome Institute Weinstock, George; Washington University School of Medicine, Genome Institute Shannon, William; Washington University School of Medicine, Medicine
Keywords:	Human microbiome, Object data analysis, ODA of taxonomic trees

View

Object data analysis of taxonomic trees from human microbiome data

Patricio S. La Rosa¹, Berkley Shands¹, Elena Deych¹, Yanjiao Zhou², Erica Sodergren², George Weinstock², and William D. Shannon¹

¹ *Division of General Medical Sciences, Washington University in St. Louis Medical School*

² *Genome Institute, Washington University in St. Louis Medical School*

Abstract

Human microbiome project (HMP) data from next generation sequencing of microbial populations are naturally represented as phylogenetic or taxonomic trees. In this paper, we introduce Object Oriented Data Analysis (OODA) methods to analyze taxonomic tree objects directly using parametric methods. In particular, using a probability measure to model a set of taxonomic trees, we introduce an approximate MLE procedure for estimating model parameters and derive likelihood-ratio test statistics for comparing the distributions of two metagenomic populations. For illustration purposes, we apply the proposed methodology to analyze data collected through the NIH's Human Microbiome Project Roadmap Initiative. Developing formal statistical methods of estimation, hypothesis testing and experimental design based on fully parametric models is essential to allow clinicians, geneticists, and ecologists to begin looking at the microbiome in clinical research. This work is a major step in this direction and should help speed up the transition from technical development to clinical application of HMP data.

1. INTRODUCTION

The Human Microbiome Project (HMP) [1] was initiated by the NIH to identify and characterize the microbes and their communities found in or on human body, focusing on the nasal cavity, oral cavity, vagina, skin, and gastrointestinal tract. The goal of the HMP is “determining whether individuals share a core human microbiome, understanding whether changes in the human microbiome can be correlated with changes in human health, and developing the new technological and bioinformatics tools needed to support these goals” [2].

Microbiome samples are collected from patient Body sites by swabbing (e.g., skin, nasal, oral) or bulk collection (e.g., saliva, stool). These samples contain within them the entire bacterial community (i.e., the microbiome) as well as other organisms (e.g., human cells, viruses, fungi). Samples are processed to isolate the genomic content (i.e., all DNA from the entire bacterial microbiome, all patient DNA, all viral DNA, etc.) within that sample, and prepared for state-of-the-art ‘next generation’ sequencing. To characterize the microbial community structure, 16S rRNA genes are sequenced using the high throughput 454 FLX Titanium sequencing platform (Roche). The sequences are analyzed using either a phylogenetic or taxonomic approach [3, 4]. The phylogenetic approach studies communities’ evolutionary relationships between sequences within the sample, and generally represents the microbiome by a phylogenetic tree. The taxonomic approach assigns sequences to taxonomic units using unsupervised and supervised methods: The unsupervised taxonomic method computes pairwise nucleotide distances between 16S rRNA gene sequences and an operational taxonomic unit (OTU) is assigned, by alignment-based clustering, to sequences that are at least 97% identical; and the supervised taxonomic method matches a sequence to a hierarchical taxa or taxonomy bins defined in a bacterial-taxonomy library such as, for example, the Ribosomal Database Project (RDP) [5], Greengenes [6], SILVA [7], and GAST [8]. The supervised taxonomic analysis allows us to represent each sample (set of sequences) by a rooted taxonomic tree where the root corresponds to taxon at the Kingdom level, i.e., bacteria, and the leaves correspond to the taxa at the Genus level, and the width of the edges (paths) between taxonomic levels correspond to the abundances of the descending taxon. A number of reviews on the phylogenetic and taxonomic analysis of sequences have appeared recently (e.g. see [4, 9, 10]). In [4], the authors showed that both the

1
2
3 supervised and unsupervised taxonomic methods arrive at similar ecological/biological
4 conclusions. However, the supervised taxonomic analysis is more tolerant to sequencing errors,
5 and it requires significantly less computational power than the taxonomy unsupervised analysis
6 [4].
7
8
9

10
11 With the goal to enumerate the content and abundances in the microbial communities of 18/15
12 body habitats of 300 healthy female/male adults, 7,000 16S rRNA sequences were produced
13 from an individual on average per body site sample. These sequence data sets provide the
14 opportunity to estimate the microbial diversity with high resolution, but statistical tools and
15 strategies to analyze the microbial communities are needed to take full advantage of the data
16 density. In recent years several tools have been developed to compare Human microbiome
17 communities using either phylogenetic or taxonomical classification of metagenomic sequences.
18 Current strategies are based primarily on exploratory cluster analysis, phylogenetic inferences,
19 biological diversity indices, bootstrap or resampling methods, and application of univariate and
20 non-parametric statistics to different subsets of the data [11-23].
21
22
23
24
25
26
27
28
29

30 Tools currently being used to analyze HMP data for limited numbers of sequence reads include
31 UniFrac, SONS, and DOTUR [11-14]. UniFrac, for example, uses phylogenetic distances and
32 permutation testing to compare samples, which does not require a large number of sequence
33 reads to detect significant differences between two samples. SONS and DOTURS compare
34 similarities between samples using OTU-based taxonomies and standard diversity indices of
35 complexity, which provide with quantitative descriptions of a community as well as of its
36 similarity to other communities. Several other methods exist that depend on sequence and
37 phylogeny comparisons (e.g., AMOVA, Tree Climber) [15] or diversity indices and community
38 coverage (e.g., LIBSHUFF [16] and S-LIBSHUFF [17]). All the above methods compare two
39 libraries of sequences, however, because of the computational complexity of calculating
40 phylogenetic trees and generating huge pairwise distance matrices between sequences, these
41 methods are meant mainly to perform pairwise sample comparisons of at most two groups of
42 samples with a restricted amount of sequences per samples. Analysis of groups of large HMP
43 samples has been limited to the application of clustering and ordination techniques based on
44 pairwise distances between samples (see for example, [24]). Other methods and ecological/HMP
45 software packages compare microbiomes based on standard statistical methods such as
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 contingency tables, Fisher's Exact Test, or goodness-of-fit tests to multinomial distributions,
4 bootstrap tests. These packages (e.g., XIPE-TOTEC [18], IMG/M [19], MEGAN [20], Metastats
5 [21], QIIME [22], and STAMP [23] require a significant reduction of the HMP data, often
6 basing the statistical analyses on pairwise comparisons of the abundances of taxa bin or OTUs or
7 other summary statistic features (e.g., community functions).
8
9

10
11
12 In this work a novel parametric statistical inference method based on object-oriented data
13 analysis (OODA) for analyzing HMP data is proposed. OODA is an emerging area of statistical
14 inference where the goal is to apply statistical methods to objects such as functions [25], images,
15 and graphs or trees [26, 27]. In particular, the data objects that pertain to this work are RDP-
16 based taxonomic trees. There is one such object for each habitat sites sampled from each person,
17 and, thus we are interested in modeling and comparing population of trees. The probabilistic
18 modeling approach proposed here has been applied previously to hierarchical clustering trees
19 (dendrograms) and classification and regression trees [26, 28, 29], as well as for constructing
20 Maximum Likelihood Supertrees [30]. This contribution is threefold: first, a weighted tree
21 structure to analyze RDP data is introduced; Second, the unimodal probability measure proposed
22 in [26, 28] is applied to model a set of RDP trees, and the likelihood ratio test statistics for
23 comparing the probability models of two microbiome populations is derived; and third the HMP
24 data is analyzed using the proposed model providing novel insights.
25
26
27
28
29
30
31
32
33
34
35
36

37 Of primary interest to HMP investigators is the estimation of the *core microbiota* from a set of
38 samples. Determining a *core microbiota* aims at finding the organisms (or functions) selected in
39 the host environment, and at studying its correlation with changes in human health. By defining a
40 unimodal probability measure we are able to compute a central taxonomic tree, the maximum
41 likelihood tree, providing an alternative and new definition of the core-microbiome for a set RDP
42 trees samples. Though, in this paper we are focused on analyzing 454 sequencing of 16S rRNA
43 genes with the reads mapped to taxonomic (classification) assignments, the methods are equally
44 applicable to shotgun sequencing data with functional profiling of the microbial community.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

2. DATA STRUCTURE

Human microbiome data analyzed in this paper for illustration are from samples of 24 subjects (male and female), 18-40 years old, from two geographic regions of the US: Houston, TX and St. Louis, MO [2]. For each subject $\sim 1 \times 10^5$ sequences were obtained from two regions of the 16S ribosomal RNA gene, denoted as variable regions V1-V3 and V3-V5, and assigned to bacterial taxa by matching the DNA sequence reads from the next generation sequencing to bacterial reference sequences using the RDP [5]. RDP matches each rRNA sequence to a set of hierarchical taxa following a Linnaean-based taxonomy, and it provides a confidence score, computed via bootstrapping, for each taxonomic classification [5]. The matching is done using a naïve Bayesian rRNA classifier which is trained on the known type strain 16S sequences (and a small number of other sequences representing regions of bacterial diversity with few named organisms) [5]. Generally as a read is assigned further down the taxonomy from kingdom to genus level there is less confidence since reads may show partial matching at more specific taxonomic levels, as well as matching to multiple taxa. This is illustrated in Table 1 where three sequence reads are mapped down to the genus level, with the associated confidence value at each level. Each sequence read's RDP match defines a taxonomic tree path, and when combining them together forms a natural representation of the HMP sample as a Linnaean taxonomic tree. A taxonomic tree is an acyclic rooted graph in which any two vertices or taxa are connected by a path or edge. HMP data is naturally represented as a rooted taxonomic tree with higher taxonomic levels (e.g., phylum or class versus family or genus) closer to the root, and edges weighted by the RDP confidence score. So far the tendency to combine RDP matches consists of applying a hard threshold filter to the confidence scores, usually at 80% or at 50%, and then overlap all branches by adding the filtered confidence scores of common paths. In the above approach, taxa with confidence scores above or equal to the threshold are assigned a score of 1, and below it are reassigned to an unknown taxa category with a score of 1. The path weights of the taxonomic tree obtained after combining the set of RDP matches provides with a measure of the abundance for the descending taxa. The above approach is somehow arbitrary since taxa abundances of known and unknown taxon will depend on the specific threshold level used. In this work, we will combine RDP values without using a threshold filter, which allows us to provide a measure of taxa abundance weighting on the confidence of each taxa assignment, and to avoid creating arbitrarily unknown taxa at each taxonomic level. For the three sequence reads

in Table 1, the tree in Figure 1 is formed by adding confidence values for reads with overlapping paths. In this example all three sequence reads contribute 0.99 confidences to the Kingdom Bacteria ($0.99 + 0.99 + 0.99 = 2.97$), while sequence 2 and 3 contribute to Phylum Firmicutes ($0.53 + 0.96 = 1.49$). Note that all taxonomic levels provide important information to characterize the sample since the aggregated confidence at a parent node in the tree is not necessarily equal to the addition of the aggregated confidences from their children nodes.

Building a taxonomic tree based on adding RDP confidences has several important properties. For example, the resulting tree is consistent with the RDP classification of each sequence where branches closer to the root have higher values than branches closer to leaves. Also, this approach allows us to identify tree branches that have overall higher confidence in each sample. Moreover, as stated above, for any given branch the addition of the confidence values provides with a measure of taxa abundance weighting on the confidence of the resulting RDP taxa assignment. However, one drawback of this approach is that Trees with larger number of sequence reads would tend to have branches with larger weight values, and thus would tend to bias the analysis when modeling a set of Trees, e.g., the computation of the MLE tree. Therefore, to avoid this issue in this work we normalize the number of sequence reads of all samples by a common number of reads.

3. PROBABILISTIC MODEL

A unimodal probability model for graph-valued random objects has been derived and applied previously to several types of graphs (cluster trees, digraphs, and classification and regression trees). [26, 29] In this paper the model is applied to HMP trees constructed from RDP data as described above. Let G be the finite set of taxonomic trees with elements g , and $d: G \times G \rightarrow \mathbb{R}^+$ an arbitrary metric of distance on G . We have the probability measure $H(g^*, \tau)$ defined by

$$\mathbb{P}(g; g^*, \tau) = c(g^*, \tau) \exp(-\tau d(g^*, g)), \quad \forall g \in G, \quad (1)$$

where $g^* \in G$ is the modal or central tree, $\tau \in \mathbb{R}^+$ is a concentration parameter, and $c(g^*, \tau)$ is the normalization constant. The probability model in (1) is known as the Gibbs distribution, the distribution that maximized the entropy providing the greatest sampling diversity [28]. Note that

if $\tau = 0$, $\mathbb{P}(g; g^*, \tau)$ becomes a uniform distribution on \mathcal{G} , and if τ is large then the trees are concentrated around g^* , in which case the data provides information about the central tree.

3.1 Distance metric:

Two broad strategies exist for defining a suitable distance metric d in tree space [29, 31, 32]: one approach focuses on counting the number of times we prune a branch (edge) or add a branch (edge) to transform one tree into another, and the second approach focuses on mapping trees into alternative mathematical structures for which natural metrics already exist. In this work we will focus on the latter approach, more specifically, we will focus on mapping trees into normed spaces.

In general, any finite graph defined on a set of labeled vertices or nodes can be uniquely characterized by mapping it into the space of matrices through the vertex-adjacency matrix $I(g)$ where $I_{ij}(g) := \{1 \text{ if vertex } i \text{ connects to vertex } j \text{ in } g; 0 \text{ otherwise}\}$, and for a weighted graph it can be simply defined as $\tilde{I}_{ij}(g) := \{w_{ij}\}$ where $w_{ij} \geq 0$ is the edge weight linking vertices i and j . If g is an undirected graph its vertex-adjacency matrix is symmetric. The distance metric $d(g_k, g_l)$ is given by the Frobenius norm [33] of the difference between the vertex-adjacency matrices of g_k and g_l i.e. $d_f(g_k, g_l) = \left(\text{tr} \left\{ \left(\tilde{I}(g_k) - \tilde{I}(g_l) \right)^2 \right\} \right)^{1/2}$, where $\text{tr}\{A\}$ is the trace of the matrix A . In the case of the RDP trees, the vertices are taxa labeled according to a Linnaean-taxonomic classification. Therefore not every pair of vertices is connected by an edge, for example, Bacilli is a descendent of Firmicutes but not of Bacteroidetes. This implies that $\tilde{I}(g_i)$ is a sparse matrix with many of its elements being always zero for any RDP tree sample. Moreover, the diagonal elements of $I_{ij}(g)$ and $\tilde{I}_{ij}(g)$ are always equal to zero since the graphs which concern our work are simple graphs which do not have loops. The above observations led us to consider a more efficient representation of RDP trees by mapping the weights of the edges that do exist, according to the Linnaean-taxonomic classification, into a vertex-adjacency vector $v(g) \in \mathbb{R}^P$ where P is the total number of existing edges. The distance metric $d(g_k, g_l)$ is the Euclidean norm of the difference between their adjacency-vectors and it is

1
2
3
4 given by $d_v(g_k, g_l) = \left(\sum_{i=1}^p (v_i(g_k) - v_i(g_l))^2 \right)^{1/2}$. Mapping trees to Euclidean space
5
6 facilitates the analysis and visualization of Tree objects, and the fitting of any probability model
7
8 is computationally simpler in this space. However, one of the drawbacks of this metric is that it
9
10 weight branches with lower range of values (e.g., genus level) do not contribute to differentiate
11
12 one tree from another. Note that it is easy to show that metrics $d_v(g_k, g_l)$ and $d_l(g_k, g_l)$ are
13
14 related as follows: $d_l(g_i, g_j) = \sqrt{2}d_v(g_i, g_j)$. Also, if all the weights are considered to be 0 or 1,
15
16 then $d_v(g_k, g_l)$ is equivalent to the square root of the Hamming distance between the trees,
17
18 which is the number of edges discrepancies between two graphs. In [28] the authors show that
19
20 under this metric the normalization constant is a function of τ only, namely, it is independent of
21
22 the central graph g^* .

23 24 25 3.2 Normalization constant:

26
27 The space of RDP trees is continuous and constrained. In fact, by construction of the RDP tree
28
29 the edge weights (the sum of confidence levels) are monotonically decreasing as we travel from
30
31 the root, the vertex at the kingdom level, to the leaves, the vertices at the genus level. This means
32
33 that $w_{ij} \geq w_{jk}$, where j is the common taxonomical level, i is its respective and unique parent
34
35 node of j , and k denotes any of the descendant nodes of j . Moreover, all weighted edges are
36
37 nonnegative. Therefore, since a weighted edge w_{ij} is an element of the vertex-adjacency
38
39 vector $v(g)$, we have that for any RDP tree g the following vector inequality should be satisfied:
40
41 $Hv(g) \geq 0$, where H is a 0, 1, -1 matrix describing the set of inequalities of the type
42
43 $w_{ij} \geq w_{jk}$ and $w_{ij} \geq 0$. Hence, the normalization constant $c(g^*, \tau)$ can be computed using the
44
45 vertex-adjacency vector mapping as follows:

$$46
47
48 c(g^*, \tau) = \left[\int_{Hv(g) \geq 0} \exp(-\tau \|v(g) - v(g^*)\|) dv(g) \right]^{-1}, \quad (2)$$

49
50
51 where the integral is defined on the subspace S_g formed by the set of inequalities $Hv(g) \geq 0$. It
52
53 is straightforward to show that a lower limit on $c(g^*, \tau)$ is given by,

$$c(\mathbf{g}^*, \tau) \geq \left[\int_{-\infty}^{\infty} \exp(-\tau \|\mathbf{v}(\mathbf{g}) - \mathbf{v}(\mathbf{g}^*)\|) d\mathbf{v}(\mathbf{g}) \right]^{-1} = \frac{\Gamma(P/2) \tau}{\Gamma(P) \sqrt{\pi}^P 2^P}, \quad (3)$$

where P is the dimension of $\mathbf{v}(\mathbf{g})$. The lower limit depends only on τ and on P . Note that since the exponential function in $c(\mathbf{g}^*, \tau)$ is symmetric around $\mathbf{v}(\mathbf{g}^*)$, then $c(\mathbf{g}^*, \tau)$ would tend to the lower limit as $\mathbf{v}(\mathbf{g}^*)$ moves away from the boundaries of S_g (including the origin). For a given $\mathbf{v}(\mathbf{g}^*)$, the difference between $c(\mathbf{g}^*, \tau)$ and the lower limit depends on the concentration parameter τ , i.e., the larger its value the smaller the difference.

4. MODEL PARAMETERS ESTIMATION

To estimate (\mathbf{g}^*, τ) from a set of sample trees we use a maximum likelihood estimate (MLE) approach. In particular, for a random sample of n observed trees, $S_n := \{\mathbf{g}_1, \dots, \mathbf{g}_n\} \subset G$, the log-likelihood is given by

$$\ln L(\mathbf{g}^*, \tau; S_n) = -n \ln \left[\int_{H_{\mathbf{v}(\mathbf{g}) \geq 0}} e^{-\tau \|\mathbf{v}(\mathbf{g}) - \mathbf{v}(\mathbf{g}^*)\|} d\mathbf{v}(\mathbf{g}) \right] - \tau \sum_{i=1}^n \|\mathbf{v}(\mathbf{g}_i) - \mathbf{v}(\mathbf{g}^*)\|, \quad (4)$$

and the MLE $(\hat{\mathbf{g}}^*, \hat{\tau})$ are such that $\ln L(\hat{\mathbf{g}}^*, \hat{\tau}; S_n)$ is maximum, $\hat{\mathbf{g}}^* \in G$, and $\hat{\tau} \geq 0$. Note that Banks and Constantine [29] pointed out, the fact that the likelihood equation (4) contains two terms whose importance depends on the value of τ . If $\tau \rightarrow 0$ then (4) is dominated by the first term of the equation, and the points (trees) becomes less important. In this case the likelihood is a non-linear function of the distances between the trees in a neighborhood around \mathbf{g}^* . However, for large τ , the second term dominates the likelihood, and the data points (trees) are of primary importance. In this case, the likelihood is a linear combination of the distances between the observed trees and the current estimate of \mathbf{g}^* .

Following the analysis of Banks and Constantine [29], it can be shown that the MLE $(\hat{\mathbf{g}}^*, \hat{\tau})$ must satisfy the following equation:

$$\frac{1}{n} \sum_{i=1}^n \|\mathbf{v}(\mathbf{g}_i) - \mathbf{v}(\hat{\mathbf{g}}^*)\| = \frac{\int_{H_{\mathbf{v}(\mathbf{g}) \geq 0}} \|\mathbf{v}(\mathbf{g}) - \mathbf{v}(\hat{\mathbf{g}}^*)\| e^{-\hat{\tau} \|\mathbf{v}(\mathbf{g}) - \mathbf{v}(\hat{\mathbf{g}}^*)\|} d\mathbf{v}(\mathbf{g})}{\int_{H_{\mathbf{v}(\mathbf{g}) \geq 0}} e^{-\hat{\tau} \|\mathbf{v}(\mathbf{g}) - \mathbf{v}(\hat{\mathbf{g}}^*)\|} d\mathbf{v}(\mathbf{g})} \quad (5)$$

where

$$\hat{g}^* = \operatorname{argmin}_{\hat{g} \in G} \left\{ n \ln \left[\int_{H_{\mathbf{v}(\hat{g})} \geq 0} e^{-\hat{\tau} \|\mathbf{v}(\hat{g}) - \mathbf{v}(g)\|} d\mathbf{v}(g) \right] + \hat{\tau} \sum_{i=1}^n \|\mathbf{v}(g_i) - \mathbf{v}(\hat{g}^*)\| \right\} \quad (6)$$

Solving the above equations for $(\hat{g}^*, \hat{\tau})$ is hard, and thus it requires numerical search algorithm to obtain an approximation solution. Shannon and Banks in [26] developed an iterative algorithm to compute an approximate MLEs $(\hat{g}^*, \hat{\tau})$ using equations (5) and (6), however their approach is specific to the set of unweighted trees (classification trees) and to the distance metric based on a weighted sum of the number of discrepant path of a certain length across all possible path lengths (See [26] for more details.) An algorithm considering the normalization constant would imply solving a multidimensional integral of an exponential function defined on the whole tree space, which does not have a closed-form solution. A Monte-Carlo integration approach incorporating Importance Sampling [34] could be attempted however at the cost of introducing a high computational complexity. Here, instead, we will approximate the likelihood function in (4) by replacing the normalization constant with its lower limit as given in (3), and thus the approximate MLEs $(\hat{g}^*, \hat{\tau})$ are given by,

$$\hat{g}^* = \operatorname{argmin}_{\hat{g} \in G} \left\{ \sum_{i=1}^n \|\mathbf{v}(g_i) - \mathbf{v}(\hat{g}^*)\| \right\}, \quad (7)$$

$$\hat{\tau} = \left(\frac{1}{n} \sum_{i=1}^n \|\mathbf{v}(g_i) - \mathbf{v}(\hat{g}^*)\| \right)^{-1}. \quad (8)$$

Note that solving the minimization problem in (7) with respect to \hat{g}^* is equivalent to solving it with respect to the vertex-adjacency vector $\mathbf{v}(\hat{g}^*)$ since \hat{g}^* is uniquely characterized by $\mathbf{v}(\hat{g}^*)$. The unconstrained minimization problem with respect to $\mathbf{v}(\hat{g}^*)$ is also known as the Fermat-Weber location problem [35], and its solution is given by the geometric median of the vertex-adjacency vectors $\mathbf{v}(g_i)$, for $g_i \in S_{\mathcal{N}}, i = 1, \dots, n$. If the set of $\mathbf{v}(g_i), i = 1, \dots, n$ are not collinear, the expression $\sum_{i=1}^n \|\mathbf{v}(g_i) - \mathbf{v}(\hat{g}^*)\|$ is strictly convex and hence there is a unique minimum, $\mathbf{v}(\hat{g}^*)$. If $\mathbf{v}(g_i), i = 1, \dots, n$ are collinear, the minimum is given by the dimension-wise median and hence it may not be unique [35]. There are no analytic solutions to compute the geometric median. Here we use the Weiszfeld iterative algorithm (see [35] for more details) with initial solution given by the mean tree, i.e., the average of the set of adjacency-vectors

$v(g_i)$, $i = 1, \dots, n$. Though the proposed MLE approximation is computationally attractive, improving upon them is an interesting problem for future research.

5. TWO-SAMPLE TEST COMPARISON

We are interested in assessing whether the distributions $H(g_1^*, \tau_1)$ and $H(g_2^*, \tau_2)$ from two metagenomic populations are the same or different, which is equivalent to evaluating whether their respective parameters are the same or different. The corresponding hypothesis is given as follows:

$$H_0: (g_1^*, \tau_1) = (g_2^*, \tau_2) = (g_0^*, \tau_0) \text{ vs } H_A: (g_1^*, \tau_1) \neq (g_2^*, \tau_2) , \quad (9)$$

where (g_0^*, τ_0) is the common parameter vector. Since the parameters under both hypothesis are unknown, we use the likelihood-ratio test (LRT) to evaluate (9), which is given by,

$$\lambda = -2 \ln \left(\frac{L(g_0^*, \tau_0; \{S_{1n}, S_{2m}\})}{L(g_1^*, \tau_1; \{S_{1n}\}) + L(g_2^*, \tau_2; \{S_{2m}\})} \right), \quad (10)$$

where $S_{1n} := \{g_{11}, \dots, g_{1n}\}$ and $S_{2m} := \{g_{21}, \dots, g_{2m}\}$ are the sets containing n and m random samples of trees from each metagenomic population, respectively. We assume that the model parameters are unknown under both the null and alternative hypothesis, therefore, we estimate these using the ML procedure described in Section 4, and compute the corresponding p-value using non-parametric bootstrap (see [28, 36] for more details.)

6. APPLICATION TO HMP DATA

We apply the HMP taxonomic tree OODA methods developed here to existing HMP data formed by 24 subjects [2]. In all our analyses below we only selected samples that have more than 1000 reads and, as a consequence, we excluded the right and left antecubital fossa sites since these ended up having 3 and 4 samples each.

For a given set of HMP stool samples we show in Figure 2 a multidimensional scaling (MDS) plot based on the pairwise Euclidean distance of all taxonomic trees. The MDS plot shows the distribution of the taxonomic trees in 2 dimensions. Individual taxonomic trees are shown around

1
2
3 the MDS plot to illustrate how the tree structure varies. The trees are displayed using a circular
4 graphical representation in which the root of the tree is at the center and same taxonomic-level
5 nodes are placed at a fix radius around the center at a fix order, allowing each taxonomic lineage
6 to be represented in a fixed consistent position in each tree. In this plots only genera are listed
7 around the circumference of the plot. The tree branches are color coded to represent different
8 ranges of weight values (see color table at the bottom left side of the figure). Blue denotes the
9 branches with the highest confidence among all, and red denote the branches with lowest
10 confidence. The maximum likelihood estimation, \hat{g}^* , summarizes the tree distribution, and is the
11 tree structure (microbiota) that maximizes the likelihood of seeing the data we observed. In
12 Figure 3 we illustrate the corresponding MLE tree for the saliva samples illustrated in Figure 2.
13
14
15
16
17
18
19
20
21

22 To illustrate differences within a body site but across variable regions of the 16S rRNA gene,
23 stool samples for 24 subjects were sequenced at variable regions V1-V3 and V3-V5, mapped to
24 the RDP database, and a taxonomic tree estimated for each sample. Figure 4 (a) shows a
25 multidimensional scaling used to display the distribution of these 48 trees showing V1-V3 (blue)
26 and V3-V5 (red) samples are overlapping; Figure 4 (b) shows the MLE tree estimated after
27 combining the two groups and Figure 4 (c) and Figure 4 (d) illustrates the MLE trees for
28 samples corresponding to V1-V3 and V3-V5 regions, respectively. The LRT test of the null
29 hypothesis that the microbiota distributions in variable regions V1-V3 and V3-V5 are the same is
30 not rejected with p-value = 0.26, based on 1000 bootstraps, and is confirmed visually by the
31 similarities in the MLE trees for these two regions. (Note that the structure of these trees places
32 the same bacteria taxon at the same location on the tree, so visual branching comparisons are
33 valid.) We conclude from this analysis that the central trees are the same for V1-V3 and V3-V5
34 and the combined MLE tree should be used as the best estimate of the central tree of the stool
35 samples. Table 2 shows the p-values of the LRT statistics applied to each body site to test for
36 similarities between samples from regions V1-V3 and V3-V5. In general, all of the sites except
37 for attached-gingivae, buccal-mucosa, palatine-tonsils, and saliva, did not reject the null
38 hypothesis that the distributions of samples from regions V1-V3 and V3-V5 are same based on a
39 statistical significance of 5%.
40
41
42
43
44
45
46
47
48
49
50
51
52
53

54 To illustrate differences across body habitats, stool and saliva samples for 24 subjects were
55 sequenced, mapped to the RDP database, and a taxonomic tree estimated for each sample. In
56
57
58
59
60

1
2
3 Figure 5(a) we computed the Euclidean pairwise distance matrix between the trees (region V3-
4 V5) and multidimensional scaling was used to display the distribution of these 48 trees showing
5 stool (red) and saliva (blue) samples are visually distinct. Figure 5 (b) displays the MLE tree
6 estimated for the two body site groups combined, and the MLE trees estimated for stool and
7 saliva separately are shown in Figure 5 (c) and (d), respectively. The LRT test of the null
8 hypothesis that the distribution parameter in stool and saliva are the same is rejected with p-value
9 $< 10^{-13}$, based on 1000 bootstraps, and is confirmed visually by the differences in the MLE trees
10 for these two body habitats. We conclude from this analysis that the distributions are different for
11 the two groups, and the MLE trees fit separately to the body habitats are the best estimates of
12 their corresponding central trees.
13
14
15
16
17
18
19
20
21

22 In Figure 6 we illustrate a multidimensional plot, using Euclidean pairwise distances, of MLE
23 trees estimated from each body site, and in Table 3 we show the p-values of the LRT statistics
24 applied to all possible pairwise comparisons of body habitats. It can be seen that, based on a
25 statistical significance of 5%, the LRT statistic rejects the null hypothesis that the distribution
26 parameters are same on pair of samples from body habitats located on anatomical regions
27 physically separated, while in the case of some body habitats sharing the same anatomical
28 region, e.g., vaginal sites, the null hypothesis is accepted. Note that multiple hypothesis testing
29 correction can be performed here, however, the example is for illustration purposes only and so
30 we did not emphasize the biological implication of our results. Also, the MLE trees (Figure 6)
31 group following the same criterion of anatomically proximity; however, we cannot conclude that
32 the MLE trees are different using the p-values of the LRT statistics (only similarities can be
33 concluded) since this test statistic assess for changes in both the central mode tree θ^* and the
34 dispersion parameter τ simultaneously.
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

51 7. DISCUSSION

52 We propose a novel parametric statistical inference method for analyzing HMP data which is
53 naturally represented in the form of a taxonomic tree. Using methods from Object Oriented Data
54 Analysis (OODA), we applied classical statistical methods for inference and hypothesis testing
55 to the analysis of HMP RDP data. In particular, we applied a unimodal probability model which
56
57
58
59
60

1
2
3 depends on a dispersion parameter and central mode tree. We introduce an approximate MLE
4 procedure for estimating model parameters and we derive LRT statistics for comparing the
5 distributions of two metagenomic populations.
6
7

8
9
10 Within the framework of representing HMP data by taxonomic trees there are currently two basic
11 approaches for defining (estimating) the core: First, a consensus tree can be built by combining
12 common branches from the samples and removing unusual samples, i.e., *the intersection tree*.
13 This approach defines the core as the set of organisms that are present in a particular body site in
14 all or in a vast majority of individuals [37]. However, this graph-theoretic approach would
15 eliminate taxa that are alternate sources of the same biological function and thus would each be
16 present in some but not all tree samples. Moreover, it assumes that every tree sample is correct
17 and no estimates of error or methods for inference are available; and second, all sequence reads
18 from each sample can be combined and a single taxonomic tree constructed, i.e., *the union tree*.
19 This approach will likely produce a biased estimate since rare sequence matches and spurious
20 matches caused by error in sequencing or incorrect matching to taxa with conserved genetic
21 regions will be retained. Therefore, the *MLE tree*, \hat{g}^* , proposed in this work, stands as a more
22 appropriate definition of core since it corresponds to the microbiome most likely to be observed,
23 and is "core" in that sense. The great comparative advantage of using \hat{g}^* as a core definition is
24 that by using a probabilistic model it deals with the variability of organisms present in the
25 samples across subjects avoiding a highly constrained and deterministic definition of the core. In
26 a future work we will study the biological insights that the *MLE tree* can provides as a core when
27 analyzing metagenomic samples of specific habitat sites.
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42 The approximate MLE tree tends to incorporate more branches than any sample tree since the
43 MLE tree correspond to the geometric median of all the samples, and thus it is the results of a
44 linear combination of all sample tree. After reviewing this point more carefully, we realized that
45 the MLE tree has the properties of a supertree, potentially a much larger tree than some or all of
46 the original tree data. Again, we believe this is an important component of the definition of the
47 core microbiome.
48
49
50
51
52

53
54
55 Our approach is based on the assumption that a unimodal model fits the set of tree samples,
56 which might not always be valid [38]. Goodness-of-fit test for the unimodal model applied to
57
58
59
60

1
2
3 binary trees has been discussed in [28] and [29]. However, the Pearson Chi-Square approach
4 proposed in [28] applies only to the set binary trees which are discrete objects and unweighted,
5 thus not applicable to the tree objects of our paper. To the best of our knowledge goodness-of-fit
6 test statistics for this model applied on weighted trees has not be derived yet. We are currently
7 working on deriving more general models such as finite mixture model of the unimodal
8 probability model to assess for the existence of several modes in the data, e.g., due to the
9 presence of subgroups of trees within the data that correspond to sample subgroups. The
10 estimation of the corresponding core microbiomes of each subgroup can be obtained by using the
11 conventional EM algorithm combined with the MLE search algorithm proposed in this work.
12 Though, this is formally a modeling selection approach, it will provide a sense of how well the
13 unimodal assumption holds in the data compared with multimodal alternatives.
14
15
16
17
18
19
20
21
22
23

24 The application of the LRT statistics to real HMP data formed by 24 subjects allowed testing for
25 differences of core microbiomes across body habitats and variable regions within the same body
26 site. These results illustrate the potential that our method has in guiding the analysis of the vast
27 amount of HMP data is currently being generated (samples from 300 subjects, 18/15 body
28 habitats, multiple visits, and multiple sequencing platforms), and in helping to bridge the
29 transition from HMP technology development to clinical applications.
30
31
32
33
34
35
36
37
38
39

40 ACKNOWLEDGMENTS

41 This work was supported by

42 U54 HG004968 “Human Microbiome Project Consortium Sequencing of Healthy People”
43 (Weinstock)

44 WUSM Dept. of Medicine Biostatistical Consulting Center (Shannon)

45 UH3 AI083265 “The Neonatal Microbiome and Necrotizing Enter colitis (Tarr)

46 Children’s Discovery Institute, St. Louis Gut Neonatal Microbiome Initiative (Warner)

47 1U01 HL101465 “Influence of the Enteric Microbiome on the Genesis of Bronchopulmonary
48 Dysplasia” (Hamvas)
49
50
51
52
53
54
55
56
57
58
59
60

REFERENCES

1. Group TNHW, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M. The NIH Human Microbiome Project. *Genome Research* 2009; 19: 2317-2323. DOI 10.1101/gr.096651.109.
2. NIH. <http://commonfund.nih.gov/hmp>.
3. Schloss PD. Evaluating different approaches that test whether microbial communities have the same structure. *ISME J* 2008; 2: 265-275.
4. Sul WJ, Cole JR, Jesus EdC, Wang Q, Farris RJ, Fish JA, Tiedje JM. Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. *Proceedings of the National Academy of Sciences* 2011. DOI 10.1073/pnas.1111435108.
5. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research* 2005; 33: D294-D296. DOI 10.1093/nar/gki038.
6. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol.* 2006; 72: 5069-5072. DOI 10.1128/aem.03006-05.
7. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies Jr, Glöckner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* 2007; 35: 7188-7196. DOI 10.1093/nar/gkm864.
8. Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Relman DA, Sogin ML. Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing. *PLoS Genet* 2008; 4: e1000255.

- 1
2
3 9. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. Metagenomic
4 Pyrosequencing and Microbial Identification. *Clin Chem* 2009; 55: 856-866. DOI
5 10.1373/clinchem.2008.107565.
6
7
- 8
9 10. Wooley JC, Godzik A, Friedberg I. A Primer on Metagenomics. *PLoS Comput Biol*; 6:
10 e1000667.
11
- 12 11. Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining
13 operational taxonomic units and estimating species richness. *Appl Environ Microbiol* 2005;
14 71: 1501-1506. DOI 71/3/1501 [pii] 10.1128/AEM.71.3.1501-1506.2005.
15
16
- 17 12. Schloss PD, Handelsman J. Introducing SONS, a tool for operational taxonomic unit-based
18 comparisons of microbial community memberships and structures. *Appl Environ Microbiol*
19 2006; 72: 6773-6779. DOI 72/10/6773 [pii] 10.1128/AEM.00474-06.
20
21
- 22 13. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA,
23 Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ,
24 Weber CF. Introducing mothur: open-source, platform-independent, community-supported
25 software for describing and comparing microbial communities. *Appl Environ Microbiol*
26 2009; 75: 7537-7541. DOI AEM.01541-09 [pii] 10.1128/AEM.01541-09.
27
28
- 29 14. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial
30 communities. *Appl Environ Microbiol* 2005; 71: 8228-8235. DOI 71/12/8228 [pii]
31 10.1128/AEM.71.12.8228-8235.2005.
32
33
- 34 15. Schloss PD, Handelsman J. Introducing TreeClimber, a test to compare microbial
35 community structures. *Appl Environ Microbiol* 2006; 72: 2379-2384. DOI 72/4/2379 [pii]
36 10.1128/AEM.72.4.2379-2384.2006.
37
38
- 39 16. Singleton DR, Furlong MA, Rathbun SL, Whitman WB. Quantitative Comparisons of 16S
40 rRNA Gene Sequence Libraries from Environmental Samples. *Appl. Environ. Microbiol.*
41 2001; 67: 4374-4376. DOI 10.1128/aem.67.9.4374-4376.2001.
42
43
- 44 17. Schloss PD, Larget BR, Handelsman J. Integration of Microbial Ecology and Statistics: a
45 Test To Compare Gene Libraries. *Appl. Environ. Microbiol.* 2004; 70: 5485-5492. DOI
46 10.1128/aem.70.9.5485-5492.2004.
47
48
- 49 18. Rodriguez-Brito B, Rohwer F, Edwards RA. An application of statistics to comparative
50 metagenomics. *BMC Bioinformatics* 2006; 7: 162. DOI 1471-2105-7-162 [pii]
51 10.1186/1471-2105-7-162.
52
53
54
55
56
57
58
59
60

19. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen I-MA, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Research* 2008; 36: D534-D538. DOI 10.1093/nar/gkm869.
20. Mitra S, Klar B, Huson DH. Visual and statistical comparison of metagenomes. *Bioinformatics* 2009; 25: 1849-1855. DOI btp341 [pii] 10.1093/bioinformatics/btp341.
21. White JR, Nagarajan N, Pop M. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Comput Biol* 2009; 5: e1000352.
22. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunencko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* 2010; 7: 335-336.
23. Parks DH, Beiko RG. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 2010; 26: 715-721. DOI btq041 [pii] 10.1093/bioinformatics/btq041.
24. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. *Science* 2009; 326: 1694-1697. DOI 1177486 [pii] 10.1126/science.1177486.
25. Ramsay JO, Silverman BW. *Functional data analysis*. Springer, 2005.
26. Shannon WD, Banks D. Combining classification trees using MLE. *Stat Med* 1999; 18: 727-740.
27. Wang H, Marron JS. *Object Oriented Data Analysis: Sets of Trees*. *The Annals of Statistics* 2007; 35: 1849-1873.
28. Banks D, Carley K. Metric Inference For Social Networks. *Journal of Classification* 1994; 11: 121-149.
29. Banks D, Constantine GM. Metric Models for Random Graphs. *Journal of Classification* 1998; 15: 199-223. DOI 10.1007/s003579900031.
30. Steel M, Rodrigo A. Maximum Likelihood Supertrees. *Systematic Biology* 2008; 57: 243-250. DOI 10.1080/10635150802033014.

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
31. Felsenstein J. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Zoology* 1973; 22: 240-249.
32. Margush T, McMorris FR. Consensus N-Trees. *Bulletin of Mathematical Biology* 1981; 43: 239-244.
33. Golub GH, Loan CFV. *Matrix computations*. Johns Hopkins University Press, 1996.
34. Gould H, Tobochnik J, Christian W. *An introduction to computer simulation methods: applications to physical systems*. (3rd edn). Pearson Addison Wesley: San Francisco, 2007.
35. Vardi Y, Zhang C-H. A modified Weiszfeld algorithm for the Fermat-Weber location problem. *Mathematical Programming* 2001; 90: 559-566. DOI 10.1007/pl00011435.
36. Efron B, Tibshirani R. *An introduction to the bootstrap*. Chapman & Hall, 1993.
37. Turnbaugh PJ, Gordon JI. The core gut microbiome, energy balance and obesity. *The Journal of Physiology* 2009; 587: 4153-4158. DOI 10.1113/jphysiol.2009.174136.
38. Holmes S. Statistical approach to tests involving phylogenies In *Statistical approach to tests involving phylogenies* Gascuel O (ed). Oxford University Press: New York, 2005.

Table 1 Example of a Linnaean taxonomic classification of three sequences. Each taxonomic assignment shows the estimated classification reliability computed via bootstrapping.

<u>Seq. ID</u>	<u>Kingdom</u>	<u>Phylum</u>	<u>Class</u>	<u>Order</u>	<u>Family</u>	<u>Genus</u>
F51YIRY01BC31	Bacteria:0.99	Bacteroidetes:0.99	Bacteroidia:0.9	Bacteroidales:0.99	Prevotellaceae:0.99	Prevotella:0.99
F51YIRY01DFQI	Bacteria:0.99	Firmicutes:0.53	Clostridia:0.53	Clostridiales:0.53	Veillonellaceae:0.53	Megasphaera:0.52
F51YIRY01CLKP	Bacteria:0.99	Firmicutes:0.96	Bacilli:0.91	Lactobacillales:0.90	Enterococcaceae:0.44	Pilibacter:0.41

For Peer Review

Table 2 P-values of the two sample test comparison, using LRT statistics and 1000 bootstraps, to test for similarities across samples from variable regions V1-V3 and V3-V5 of the 16S rRNA gene, within a body site.

Body Habitats	P-value
anterior-nares	0.15
attached-gingivae	<E-06
buccal-mucosa	0.02
hard-palate	0.12
l-antecubital-fossa	0.53
l-retroauricular-crease	0.47
mid-vagina	0.23
palatine-tonsils	0.03
posterior-fornix	0.22
saliva	0.02
stool	0.26
subgingival-plaque	0.12
supragingival-plaque	0.12
throat	0.10
tongue-dorsum	0.05
vaginal-introitus	0.20

Table 3 P-values of two sample test comparison using LRT statistics and 1000 bootstraps on HMP data formed by 24 subjects, region V3-V5.

p-value	anterior-nares	attached-gingivae	buccal-mucosa	hard-palate	l-retroauricular-crease	mid-vagina	palatine-tonsils	posterior-fornix	r-retroauricular-crease	saliva	stool	subgingival-plaque	supragingival-plaque	throat	tongue-dorsum	vaginal-introitus
anterior-nares	1															
attached-gingivae	0.00	1														
buccal-mucosa	0.00	0.10	1													
hard-palate	0.00	0.04	0.07	1												
l-retroauricular-crease	0.00	0.00	0.00	0.00	1											
mid-vagina	0.00	0.00	0.00	0.00	0.00	1										
palatine-tonsils	0.00	0.01	0.00	0.27	0.00	0.00	1									
posterior-fornix	0.00	0.00	0.00	0.00	0.00	0.60	0.00	1								
r-retroauricular-crease	0.00	0.00	0.00	0.00	0.46	0.00	0.00	0.00	1							
saliva	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1						
stool	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1					
subgingival-plaque	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1				
supragingival-plaque	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	1			
throat	0.00	0.00	0.00	0.01	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	1		
tongue-dorsum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1	
vaginal-introitus	0.00	0.00	0.00	0.00	0.00	0.47	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

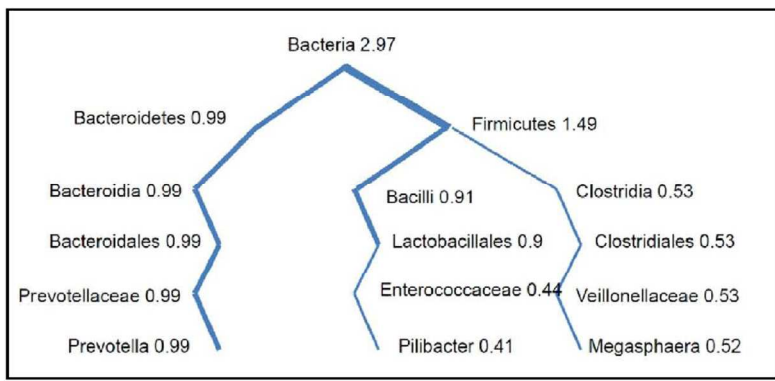


Figure 1 Taxonomical tree build from adding three RDP classifications of sequences as shown in Table 1.

165x73mm (200 x 200 DPI)

Peer Review

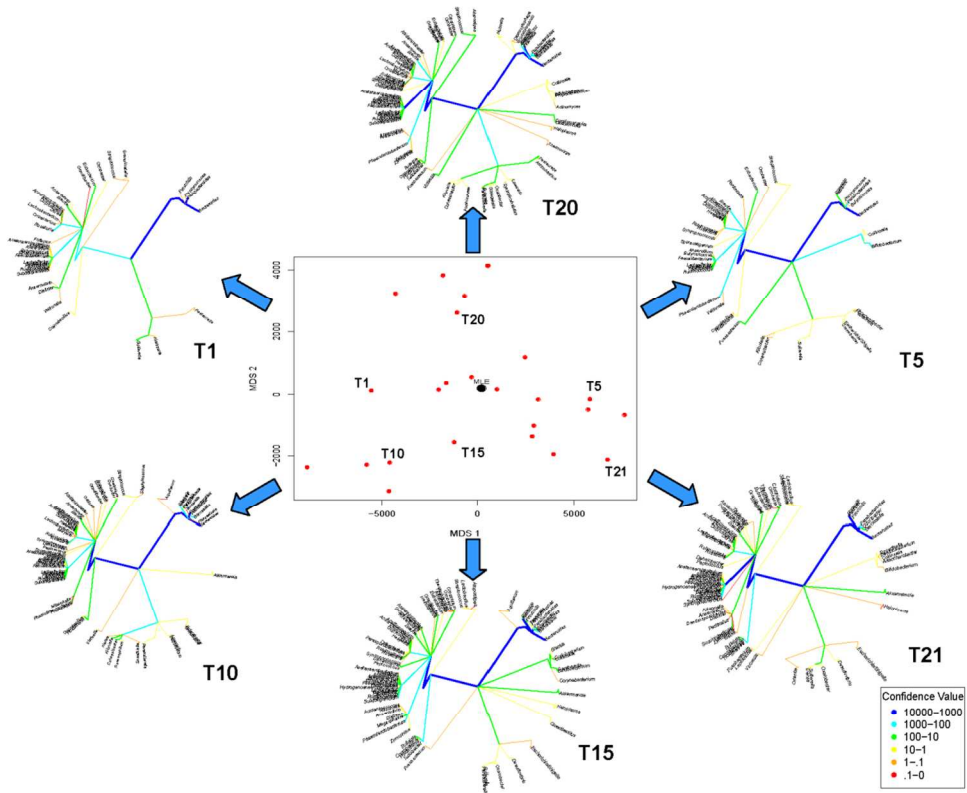


Figure 2 MDS plot showing the distribution of the taxonomic trees corresponding to stool samples sequenced at region V3-V5. The tree branches are color-coded to represent their weight values (sum of confidence) according to the reference table at the bottom left side of the plot. Blue denote the branches with the highest confidence among all while red denote the branches with lowest confidence. Note here that the tip of each branch represents a genus, and the location of each genus is the same on all trees

171x177mm (200 x 200 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

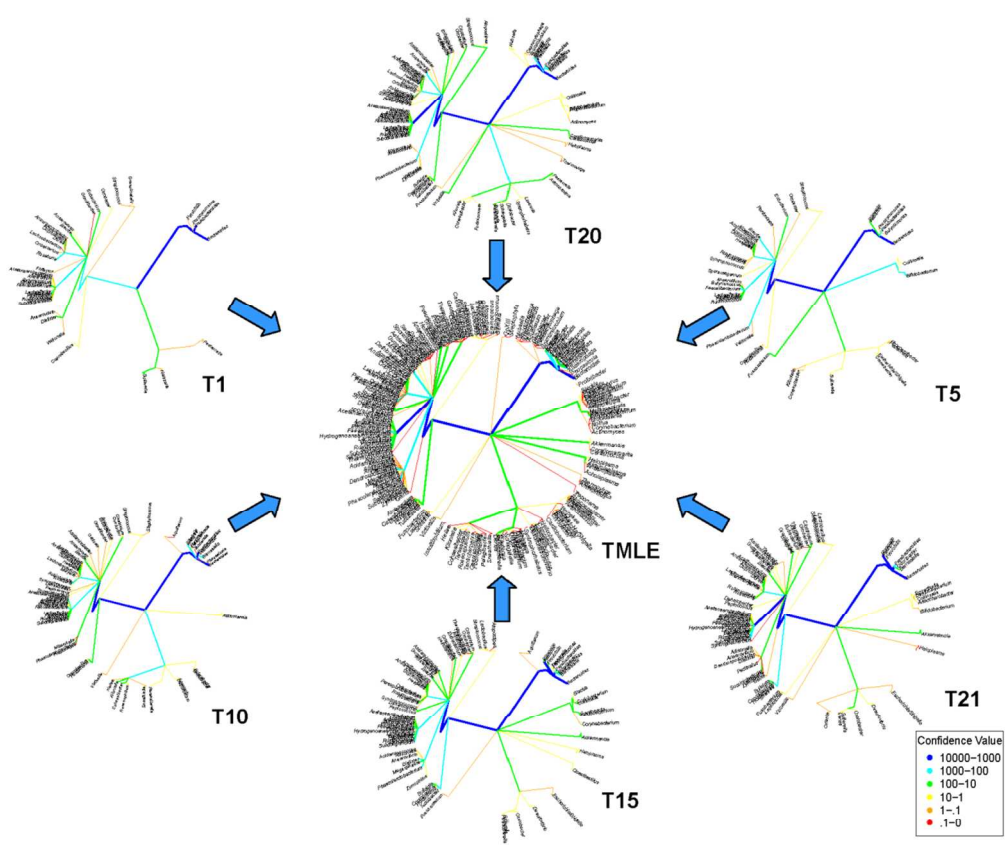


Figure 3 Illustration of the MLE tree for stool samples, region V3-V5.

165x152mm (200 x 200 DPI)

ew

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

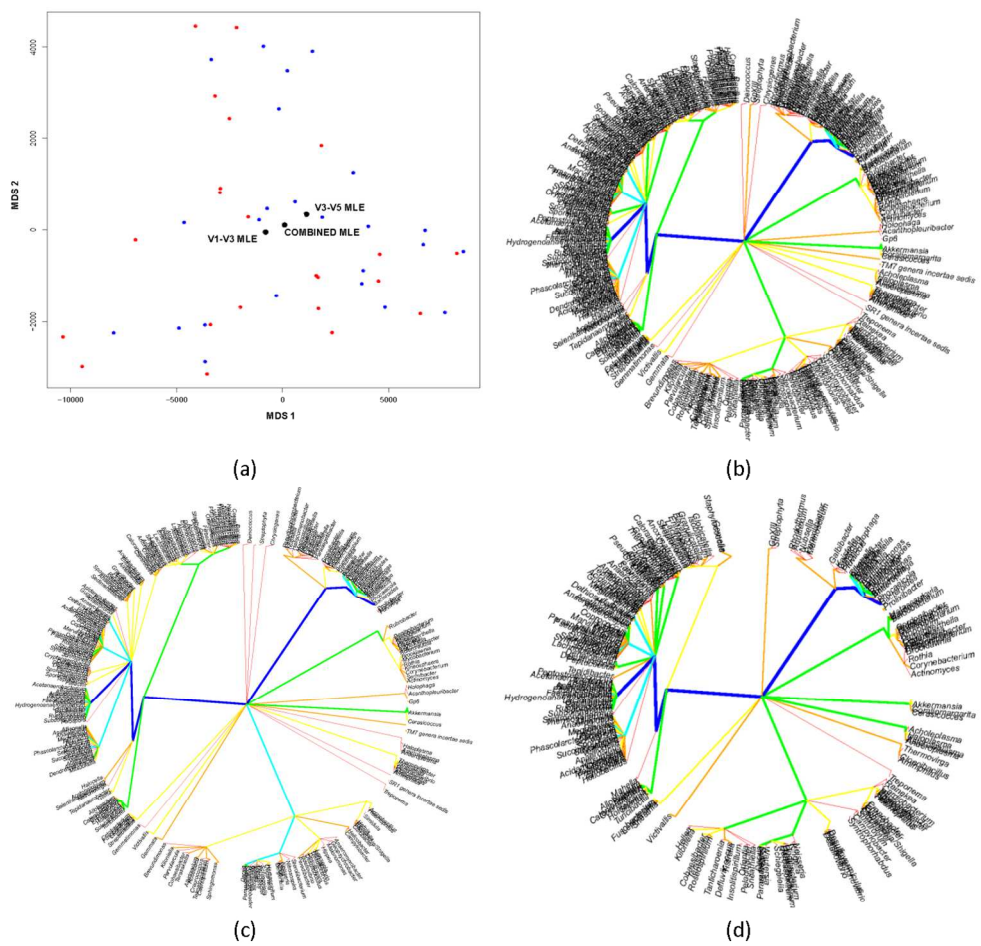


Figure 4 Analysis of stool samples for 24 subjects sequenced at variable regions V1-V3 and V3-V5, mapped to the RDP database: (a) A pairwise distance matrix was generated using Euclidean distance, and multidimensional scaling was used to display the distribution of these 48 trees showing V1-V3 (red) and V3-V5 (blue) samples are overlapping; (b) MLE tree for the 48 trees; (c) and (d) MLE tree for trees corresponding to V1-V3 and V3-V5 regions, respectively.

171x190mm (220 x 220 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

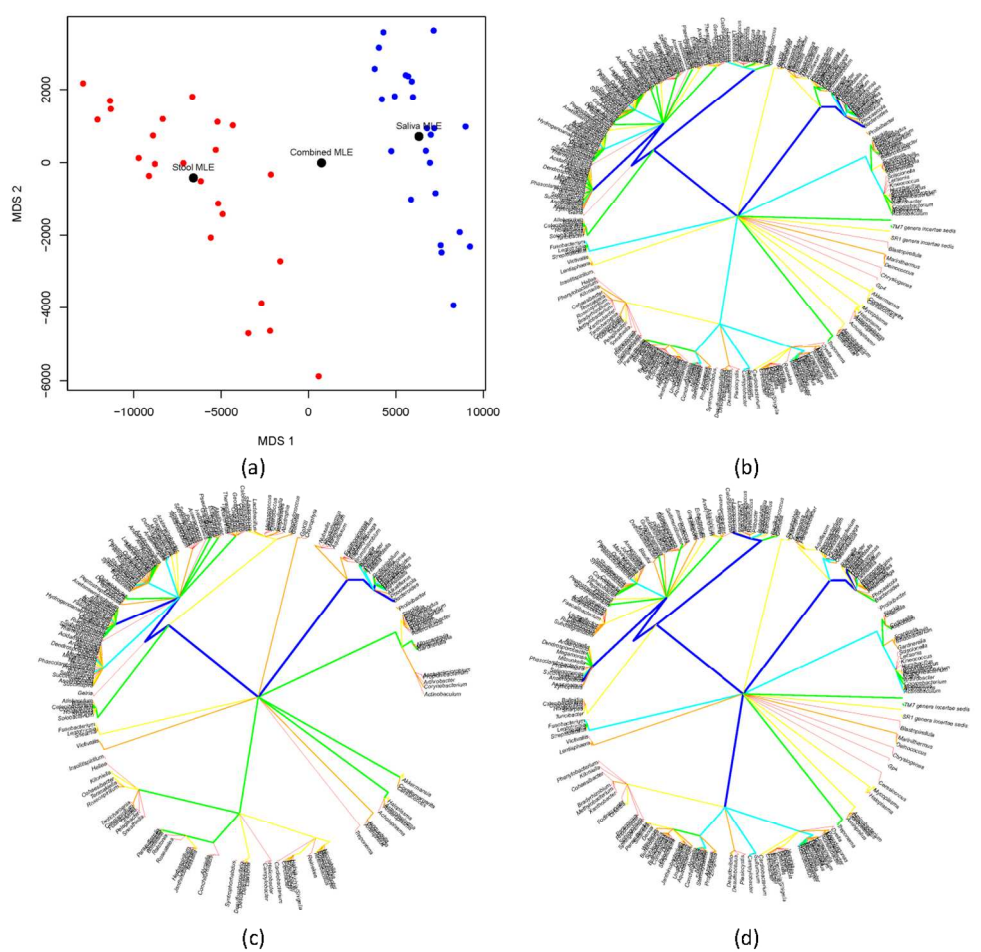


Figure 5 Analysis of saliva and stool samples for 24 subjects sequenced at variable regions V3-V5, mapped to the RDP database: (a) A pairwise distance matrix was generated using Euclidean distance and multidimensional scaling was used to display the distribution of these 48 trees showing stool (red) and saliva (red) samples do not overlap; (b) MLE tree for the tree samples combined; (c) and (d) MLE tree for trees from stool and saliva samples, respectively.

171x193mm (220 x 220 DPI)

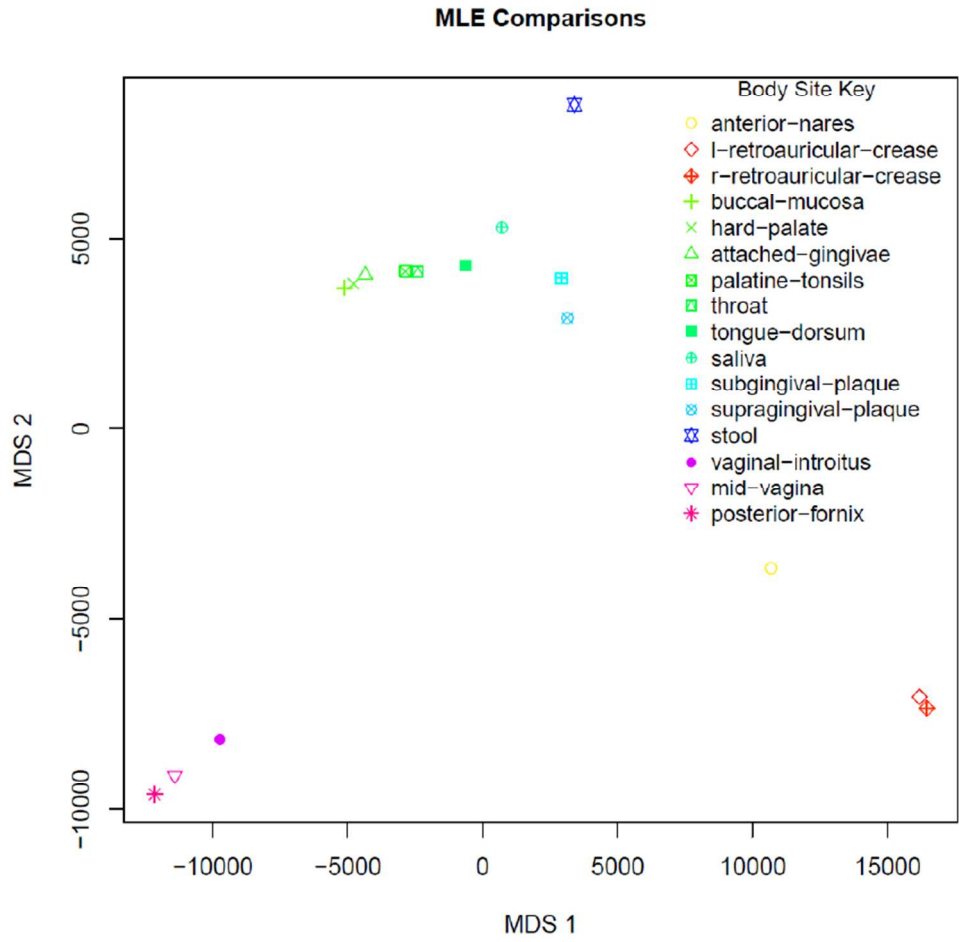


Figure 6 Multidimensional plot, using Euclidean pairwise distances, of MLE trees estimated from the HMP data formed by 24 subjects, region V3-V5.

234x252mm (96 x 96 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60