

Statistical Foundations for Analyzing Human Microbiome Data

Patricio S. La Rosa¹, Paul Brooks², Yanjiao Zhou¹, Elena Deych¹, Berkley Shands¹, Ed Boone², David Edwards², Qin Wang², Erica Sodergren¹, George Weinstock¹, and Bill Shannon¹

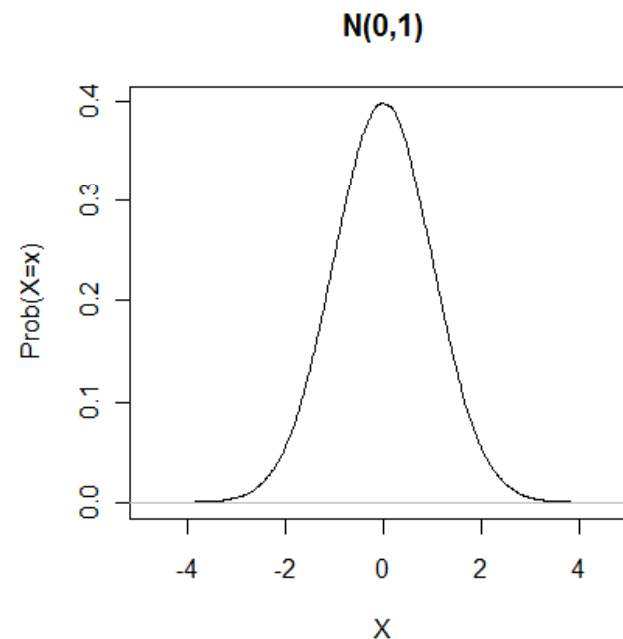
¹Washington University in St. Louis Medical School

²Virginia Commonwealth University

Probability Models Simplify Data

- Replace data by model and parameters
 - Mean and std. dev. defines normal data
 - Statistical tests compare parameters (e.g., t-test)
- What probability models will work for HMP data?

$$P(X_i = x_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-u)^2}{2\sigma^2}}$$

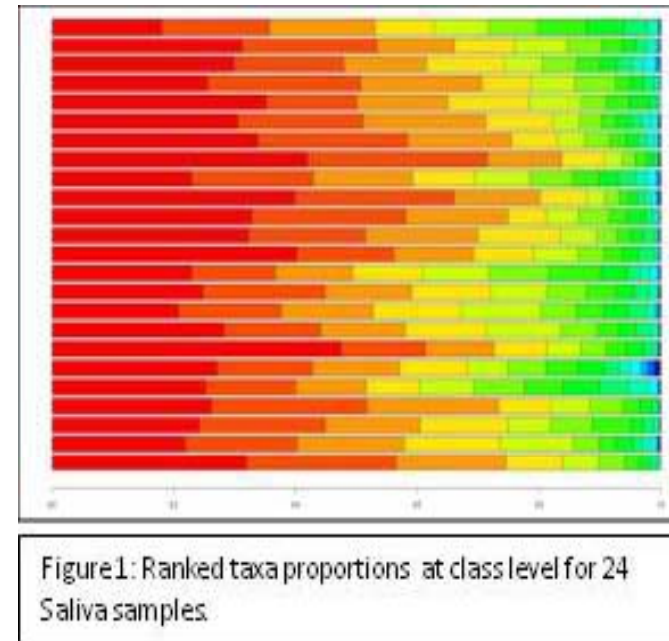


Dirichlet-Multinomial Distribution

- Relative Abundance Data
 - Numbers of individuals observed for each taxon
 - Multivariate descriptor of ecological community

$$P(\mathbf{X}_i = x_i; \{\pi_j\}, \theta) = \frac{N_i!}{x_{i1}! \cdots x_{iK}!} \frac{\prod_{j=1}^K \prod_{r=1}^{x_{ij}} \{\pi_j(1-\theta) + (r-1)\theta\}}{\prod_{r=1}^{N_i} \{(1-\theta) + (r-1)\theta\}}$$

$\{\pi_j\}$ = mean proportion of taxa j, θ = measure of dispersion



Does DM Fit HMP Data?

Table 4. GOF test [2] and overdispersion parameter θ of HMP classes sampled from different body sites.

Body Site	θ	P-value
Saliva	0.006	$< 10^{-12}$
Buccal mucosa	0.012	$< 10^{-12}$
Hard palate	0.012	$< 10^{-12}$
Attached gingiva	0.036	$< 10^{-12}$
Subgingival plaque	0.005	$< 10^{-12}$
Supragingival plaque	0.010	$< 10^{-12}$
Tongue dorsum	0.019	$< 10^{-12}$
Throat	0.014	$< 10^{-12}$
Mid vagina	0.032	$< 10^{-12}$
Stool	0.013	$< 10^{-12}$

- Goodness-of-Fit
 - Power $> 99\%$ to correctly decide data is Dirichlet-Multinomial
 - Size of test to correctly decide data is multinomial $\sim 5\%$
- Simulations indicate DM is good fit to HMP data

What Hypotheses Can We Test?

Table 2: Hypothesis tests and test statistics for microbial community comparison.	
Hypothesis Test	Hypothesis and Test statistic
[3]	$H_0: \boldsymbol{\pi} = \boldsymbol{\pi}_0 \text{ vs } H_A: \boldsymbol{\pi} \neq \boldsymbol{\pi}_0$ $X_{SC} = NP \left(1 + \hat{\boldsymbol{\theta}}(N - 1)\right)^{-1} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)^T (D(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}^T)^{-1} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$
[4]	$H_0: \boldsymbol{\pi}_1 = \dots = \boldsymbol{\pi}_j = \dots = \boldsymbol{\pi}_j \text{ vs } H_A: \boldsymbol{\pi}_i \neq \boldsymbol{\pi}_j$ $X_{MC} = \sum_{j=1}^J N_j \left(1 + \hat{\boldsymbol{\theta}}_j(N_j - 1)\right)^{-1} \sum_{i=1}^K \hat{\boldsymbol{\pi}}_{oi}^{-1} (\hat{\boldsymbol{\pi}}_{ij} - \hat{\boldsymbol{\pi}}_{oi})$

- Test model parameters
 - [3] analogous to 1 sample t-test
 - [4] analogous to 2 sample t-test or ANOVA

Power and Sample Sizes?

Table 3. Comparing RAD means from 2 populations using hypothesis test [5].					
P/Nr	100	500	1000	10000	20000
10	0.78	0.87	0.89	0.90	0.90
20	0.89	0.97	0.98	0.98	0.98
40	0.98	>0.99	>0.99	>0.99	>0.99
60	>0.99	>0.99	>0.99	>0.99	>0.99
100	>0.99	>0.99	>0.99	>0.99	>0.99

Object Data Analysis (ODA)

- Apply probability model to graphical (tree) objects
 - Sequence reads map to paths in a tree
 - Samples map to a tree

$$P(G_i = g_i; g^*, \tau) = c(g^*, \tau) \times \exp(-\tau d(g^*, g))$$

g^* = core microbiome, τ = dispersion, d = distance

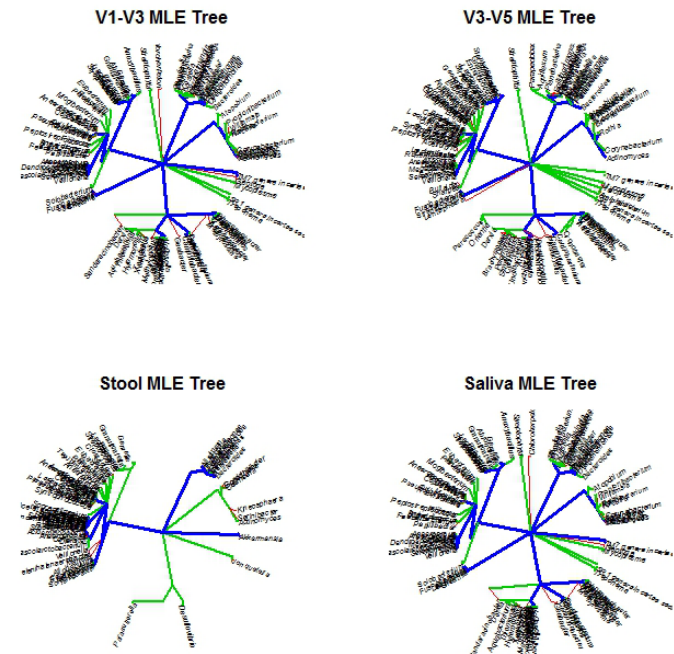
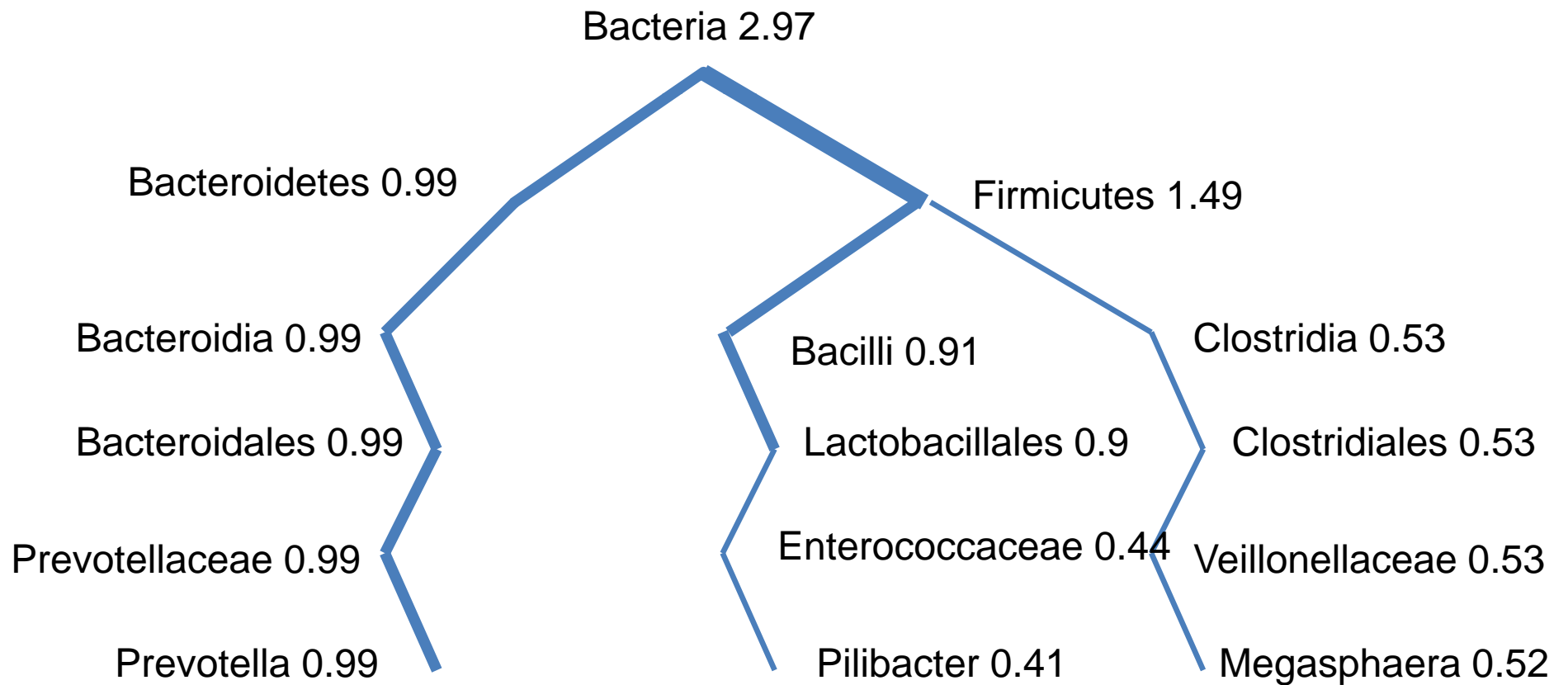
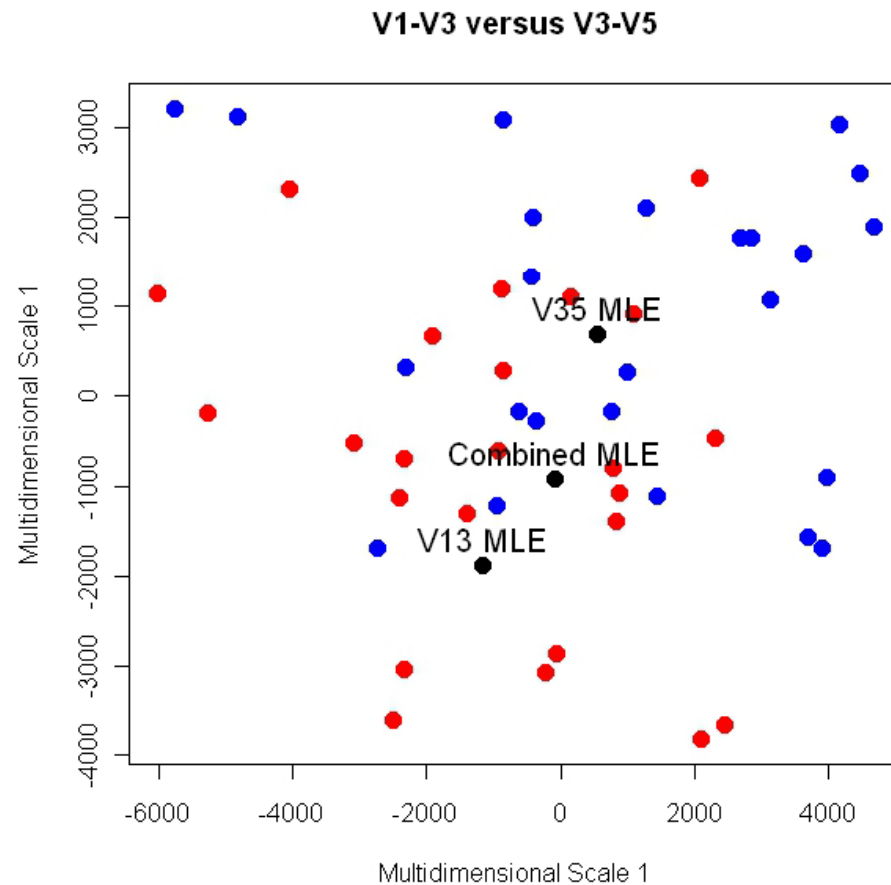


Table 5: Examples of RDP taxonomic assignments for three sequences from a microbiome sample

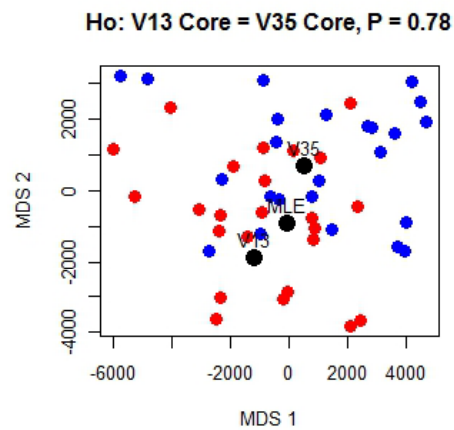
<u>Seq. ID</u>	<u>Kingdom</u>	<u>Phylum</u>	<u>Class</u>	<u>Order</u>	<u>Family</u>	<u>Genus</u>
F51YIRY01BC31	Bacteria:0.99	Bacteroidetes:0.99	Bacteroidia:0.9	Bacteroidales:0.99	Prevotellaceae:0.99	Prevotella:0.99
F51YIRY01DFQI	Bacteria:0.99	Firmicutes:0.53	Clostridia:0.53	Clostridiales:0.53	Veillonellaceae:0.53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



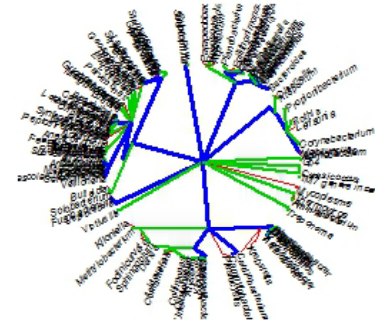
How do we estimate the core?



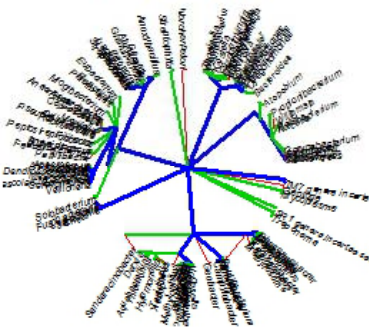
Are Variable Region Cores Equal?



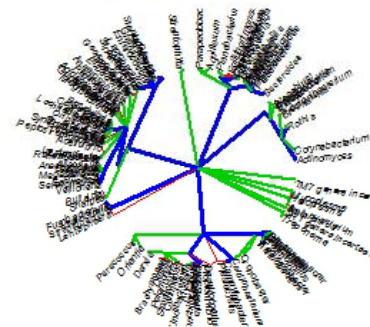
Combined MLE Tree



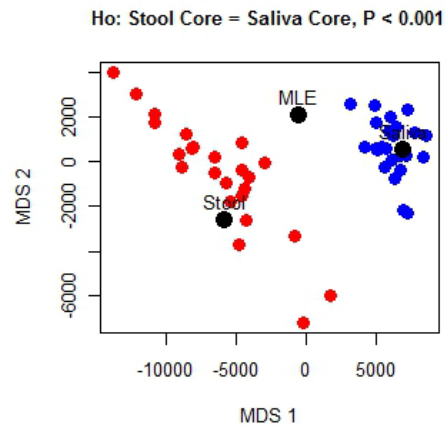
V1-V3 MLE Tree



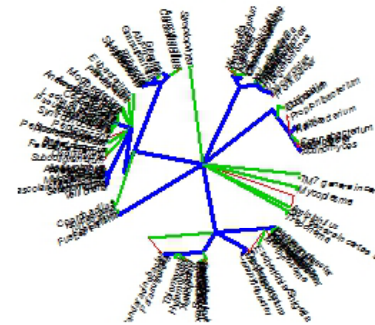
V3-V5 MLE Tree



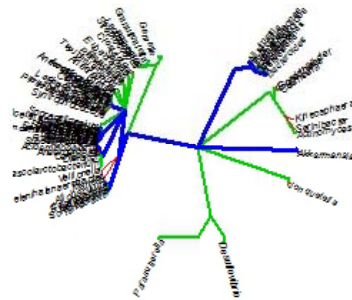
Are Body Site Cores Equal?



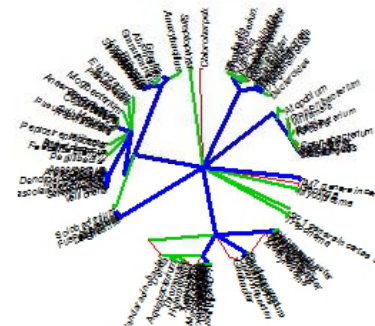
Combined MLE Tree



Stool MLE Tree



Saliva MLE Tree



Why Use Probability Models?

- **Parameters simplify interpretation of data** (e.g., core defined by central graph)
- **Formal hypotheses and P values** (e.g., DM t-test and ANOVA analogs)
- **Existing statistical machinery** (e.g., power calculations for study design)
- **All estimates come with error** (e.g., confidence errors)

Two Posters

- Dirichlet-Multinomial Power Calculations and Statistical Tests for Microbiome Data
 - La Rosa, Brooks, Deych, Boone, Edwards, Wang, Sodergren, Weinstock, Shannon
- Statistical Analysis of Taxonomic Trees in Microbiome Research
 - La Rosa, Zhou, Deych, Shands, Sodergren, Weinstock, Shannon