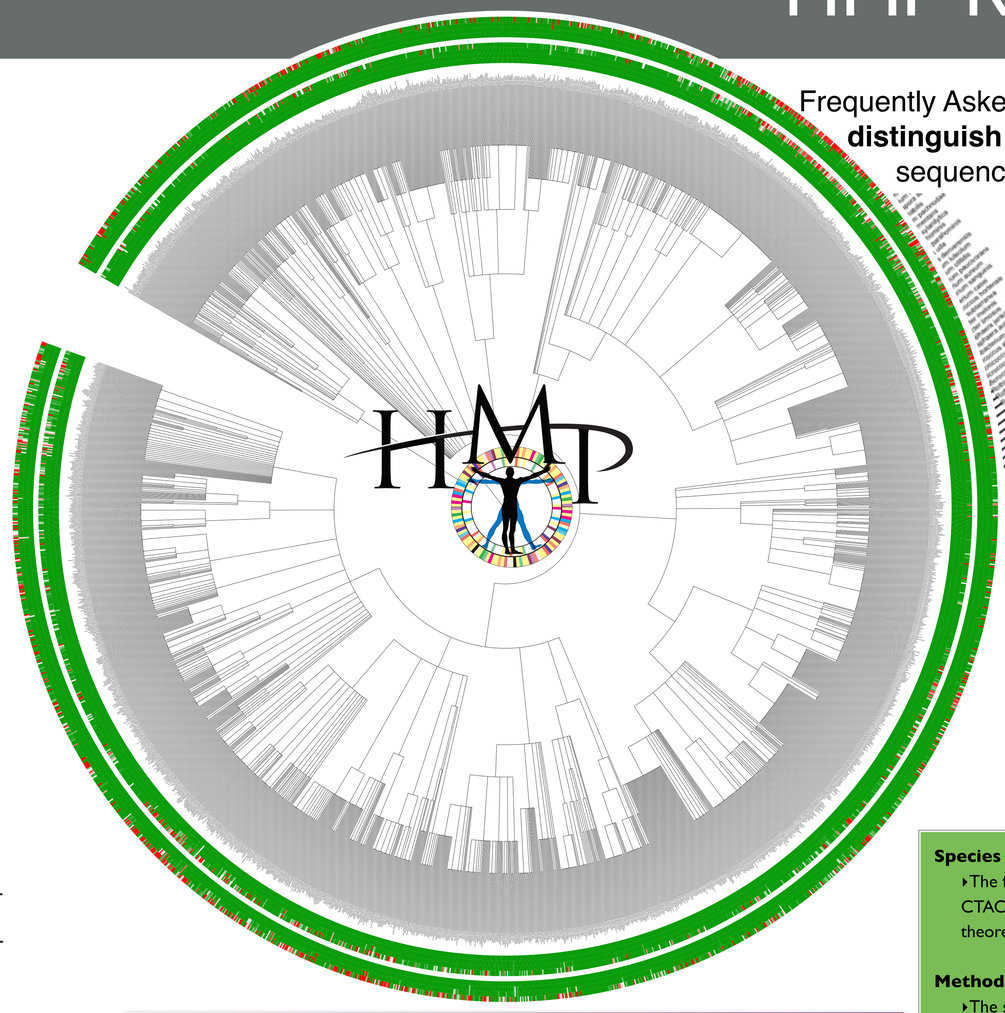
Greengenes 16S Ribosomal RNA Gene Database Update For 2011 HMP Reference Sets And Tools

CONTRIBUTORS	
Todd Z. DeSantis	LBNL & Second Genome, Inc.
Navjeet N. S. Singh, Gary L. Andersen	Lawrence Berkeley National Lab
Jeffrey Werner, Omry Koren, Ruth Ley	Cornell University
Daniel McDonald, Rob Knight	University of Colorado
Alexander Probst	University of Regensburg
Scott Kelly	San Diego State University
Kasthuri Venkateswaran	Jet Propulsion Laboratory
Owen White	University of Maryland
Zhiheng Pei	New York University
Phil Hugenholtz	Australian Centre for Ecogenomics
Strains & Data Analysis Working Groups	HMP Consortium



ABSTRACT

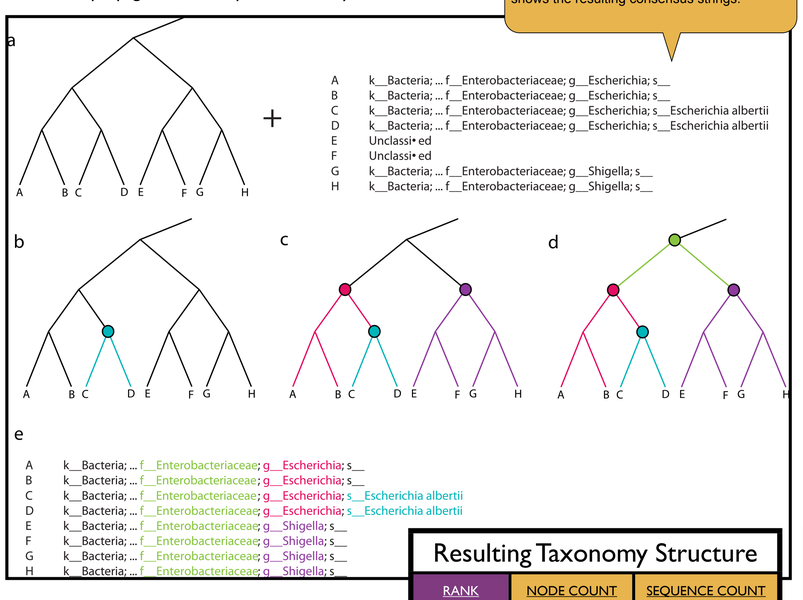
Due to global interest in the human microbiome's role in health and disease, a diverse community of international researchers from the medical, microbiological and computational fields have recently converged to address questions in microbial ecology. Activities such as describing community structure, portraying population dynamics, and depicting diagnostic test candidates all benefit from mapping assay data to high-quality reference sets with useful nomenclature. In popular workflows, ribosomal gene segments are hybridized to probes or sequenced with NGS technology. Annotating the matches to both cultured and vetted uncultured clades reveals trends overlooked by sole reliance on cultured references. The 2011 Greengenes 16S rRNA Taxonomy was created from Infernal-improved NAST alignments, FastTree-validated tree topology, nomenclature reconciliation with NCBI for cultured strains, and manual curation of thousands of yet-to-be cultured groups. The effort resulted in standardized or proposed names for >4000 hierarchical taxonomic nodes. From these relations, existing datasets from 454 (Roche), HiSeq (Illumina), and PhyloChip Assay (Second Genome) and expected datasets from PGM (Ion Torrent) and PacBioRS (Pacific Biosciences) have the opportunity to be compared. Files are available for download in their complete forms as well as subsets suitable for metagenomic pipeline tools. Furthermore, a chromatogram processing and capture tool has been established those desiring to contribute to future reference sets. The Greengenes database is supported by grant UH3CA140233 from HMP of the NIH Roadmap Initiative and National Cancer Institute and NIH common fund contract U01-HG004866, a Data Analysis and Coordination Center for the Human Microbiome Project.

NOVEL TREE - RICH TAXONOMY

- **Methods**
  - Infernal alignment of full length sequences (Nawrocki, 2010)
  - Hypervariable lane mask (Lane, 1991)
  - Uchime chimera filter (Edgar, 2010)
  - Chimera Slayer chimera filter (Haas, submitted)
  - Dual Study Heuristic filter
  - FastTree de-novo tree construction from 407K seqs (Price, 2010)
  - tax2tree node re-mapping (McDonald, in prep)
    - nomenclature mapping from NCBI
    - nomenclature mapping from Greengenes
    - Precision/Recall name conflict resolution
    - Back-filling classifier for unnamed nodes
    - Back-propagation to collapse redundantly named nodes

Overview of the tax2tree workflow. Panel a: Inputs are a phylogenetic tree + taxonomy strings for some (or all) of the tips in the tree including two Unclassified tips (E,F). The taxonomy mapping can optionally have rank abbreviations already appended on (k__, p__, etc...). Note, the consensus strings are shortened (...) for brevity.

Panels b, c, and d: Assignment of species, genus and family nodes, respectively. When decorating genus, we are able to infer in this case that tips E and F are under g__Shigella as the lowest common ancestor with tips G and H has >= 50% relative abundance of the genus name. Panel e) shows the resulting consensus strings.



Resulting Taxonomy Structure		
RANK	NODE COUNT	SEQUENCE COUNT
Kingdom	2	407K
Phylum	90	407K
Class	179	407K
Order	253	402K
Family	460	395K
Genus	1365	314K
Species	2488	112K

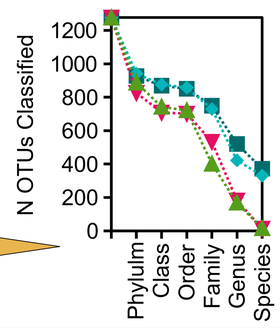
Examples of important nomenclature updates:

Anaerocellum_thermophilum	=>	*Caldicellulosiruptor bescii*
Brevibacterium stationis	=>	*Corynebacterium stationis*
Desulfomicrobium terraneus	=>	*Desulfomicrobium thermophilum*
Marinibacillus marinus	=>	*Jeotgalibacillus marinus*
Clostridium orbiscindens	=>	*Flavonifractor plautii*
Thermoanaerobacter tengcongensis	=>	*Caldanaerobacter subterraneus*
Rhodoferax ferrireducens	=>	*Albidiferax ferrireducens*
Vibrio fischeri	=>	*Aliivibrio fischeri*

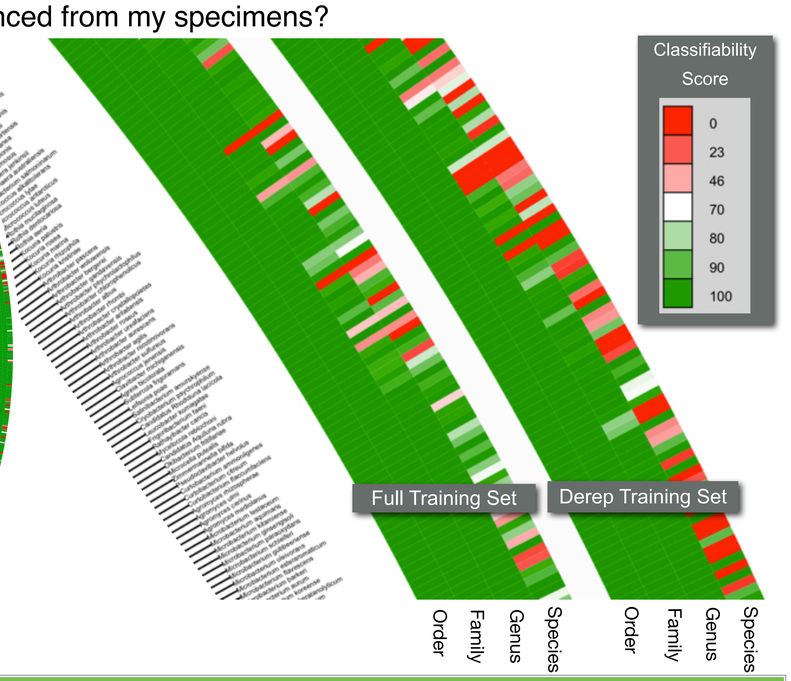
CLASSIFIABILITY

- **Test by classifying 1,200 sequence clusters from human feces against GG and other reference sets.**
- 407K GG reference sequences
  - Derep at 99 or 91.3
  - Trim to amplicon span
  - Train Bayesian classifier
  - Compare number of pyrosequenced OTUs classified at each rank against using GG or other reference sets.

New Greengenes taxonomy allows taxonomic nomenclature to be applied to a significantly greater number of NGS sequences.



Frequently Asked Question: "Which **species** should I expect to **confidently distinguish** considering **my** subsection of the 16S rRNA gene sequenced from my specimens?



Species Classifiability Example

- The foregut microbiome amplicon of interest is 347F (5'-GGAGGCAGCAGTRRGGAAT) to 803R (5'-CTACCRGGGTATCTAATCC) (Nossa, 2010). For each known bacterial species, how confidently could the theoretical amplicon distinguish it from all other nodes (named or not).

Method

- The search pattern, GGAGGCAGCAGTRRGGAATJGGATTAGATACCCYGGTAG, was used in conjunction to the greengenes online sub-alignement locator (http://greengenes.lbl.gov/cgi-bin/nph-probe_locator.cgi) to determine the coordinates of the theoretical amplicon at position 1882 to 4081 (inclusive of both primers). A subsequence from 1917 to 4048 gathers all the DNA sequence exclusive of the primers.
- 407K sub-sequences were obtained.
  - http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/
    - 347to803_gg_norm_unaligned.fasta.gz
- Train the Mothur (Schloss, 2009) classifier using all 407K sequences (Full Training Set) or a non-redundant set of 176K sequences (Derep Training Set).
- 2. Classify each sequence from the training set and record the confidence at each taxonomic rank.
- Find average "classifiability" for each node as the mean confidence of classification *to* that node divided by the set of sequences *belonging* to that node.
- Plot the classifiability of the named species as a heat map using ITOL (Letunic, 2006).

Observations

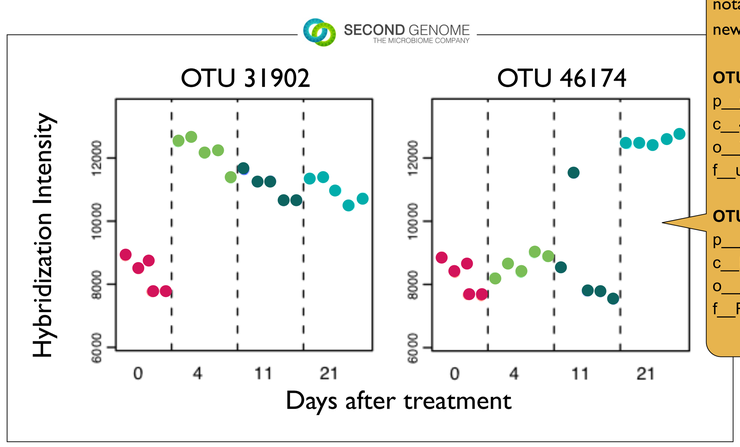
- Family-level classification was highly confident throughout the tree.
- Genus and species level confidence is dependent on the genus and species.
- Confidence was greater using the Full Training Set compared to the Derep Training Set.

For Further Discussion

- Within the trained model the count of taxonomy-by-word intersections will not differ across the two methods but the priors will change.
- More redundant sequences in one node than another affects the model's view of the distribution of a given 8mer across taxonomic nodes.

PHYLOCHIP™ ASSAY ANNOTATION

- **New Greengenes taxonomy facilitates PhyloChip results annotation.**
- 1,016,064 probes
- 59K OTUs, each tracked with multiple probe pairs
- Overcomes sampling effort problem encountered with NGS approaches
  - typical 16S rRNA gene PCR yields 500 to 1,000 ng in a 20 uL volume
    - 1500bp @ 660 g/mole/bp ➡ 5E-14 moles / uL ➡ 3E+10 molecules /uL ➡ 6E+11 total sequences ➡ 600 billion sequences per PCR sample
  - How many should we observe? 600? 60,000 (1 out of every 10 million)?
- Hybridize them all on to a PhyloChip ...
  - Dominant populations do not occlude minority populations



Example output from a PhyloChip (Second Genome, Inc.) experiment tracking bacterial population dynamics in hindgut for 21 days after a specific medical treatment OTU annotations are mapped from the new Greengenes taxonomy:

OTU 31902:
p__Cyanobacteria;
c__4C0d-2;
o__YS2;
f__unclassified

OTU 46174:
p__Bacteroidetes;
c__Bacteroidia;
o__Bacteroidales;
f__Rikinellaceae;

GREENGENES SUBSETS

- All named isolates
- All HMP Genome Strains
- Knight, Caporaso: QIIME-ready reference sets
- Current aligned Core Set for NAST templates
  - 36,550 genes represents the known 16S diversity.
- Older Core Sets
- VISIT: http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files

GENOMES QC

Contigs from genome projects finished or unfinished can be quality filtered using the Greengenes Core Set. Surprisingly over 50 genomes contain zero full-length16S rRNA genes. Finishing effort will be needed to assemble 16S genes in these projects for use as references for future trees.

ACKNOWLEDGMENTS

This study was supported in part by grant UH2/UH3CA140233 from the Human Microbiome Project of the NIH Roadmap Initiative and National Cancer Institute and by NIH common fund contract U01-HG004866, a Data Analysis and Coordination Center for the Human Microbiome Project. Work performed at Lawrence Berkeley National Laboratory is under the U.S. Department of Energy contract number DE-AC02-05CH11231. Manual taxonomic nomenclature verification was supported in part by Second Genome, Inc (San Francisco, CA)

SECOND GENOME THE MICROBIOME COMPANY · BERKELEY LAB · green genes · U.S. DEPARTMENT OF ENERGY