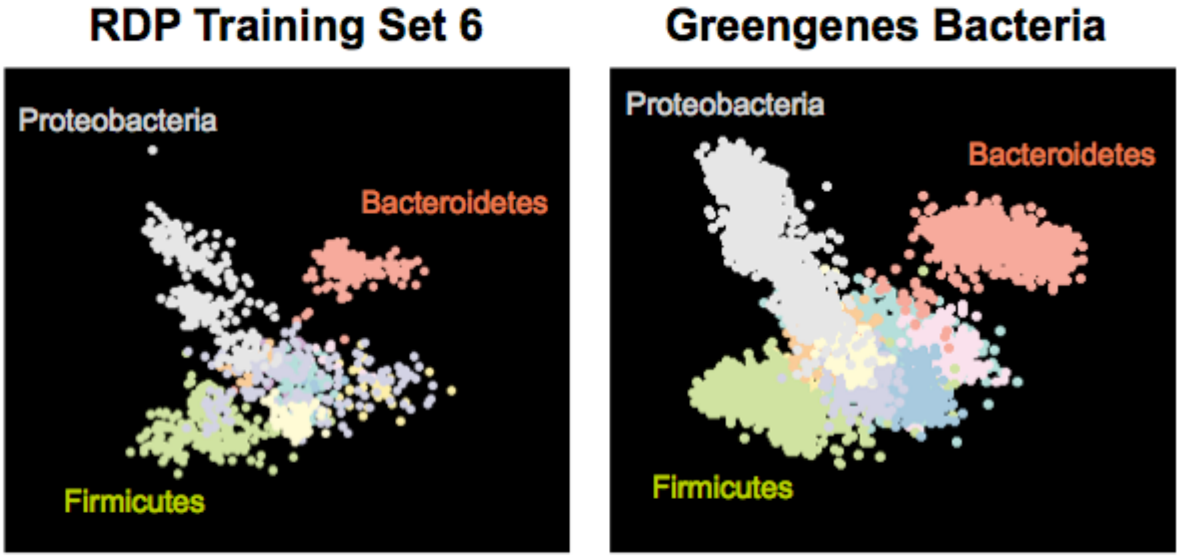
Supplementary Information for

Impact of Training Sets on Classification of High-Throughput Bacterial 16S rRNA Gene Surveys

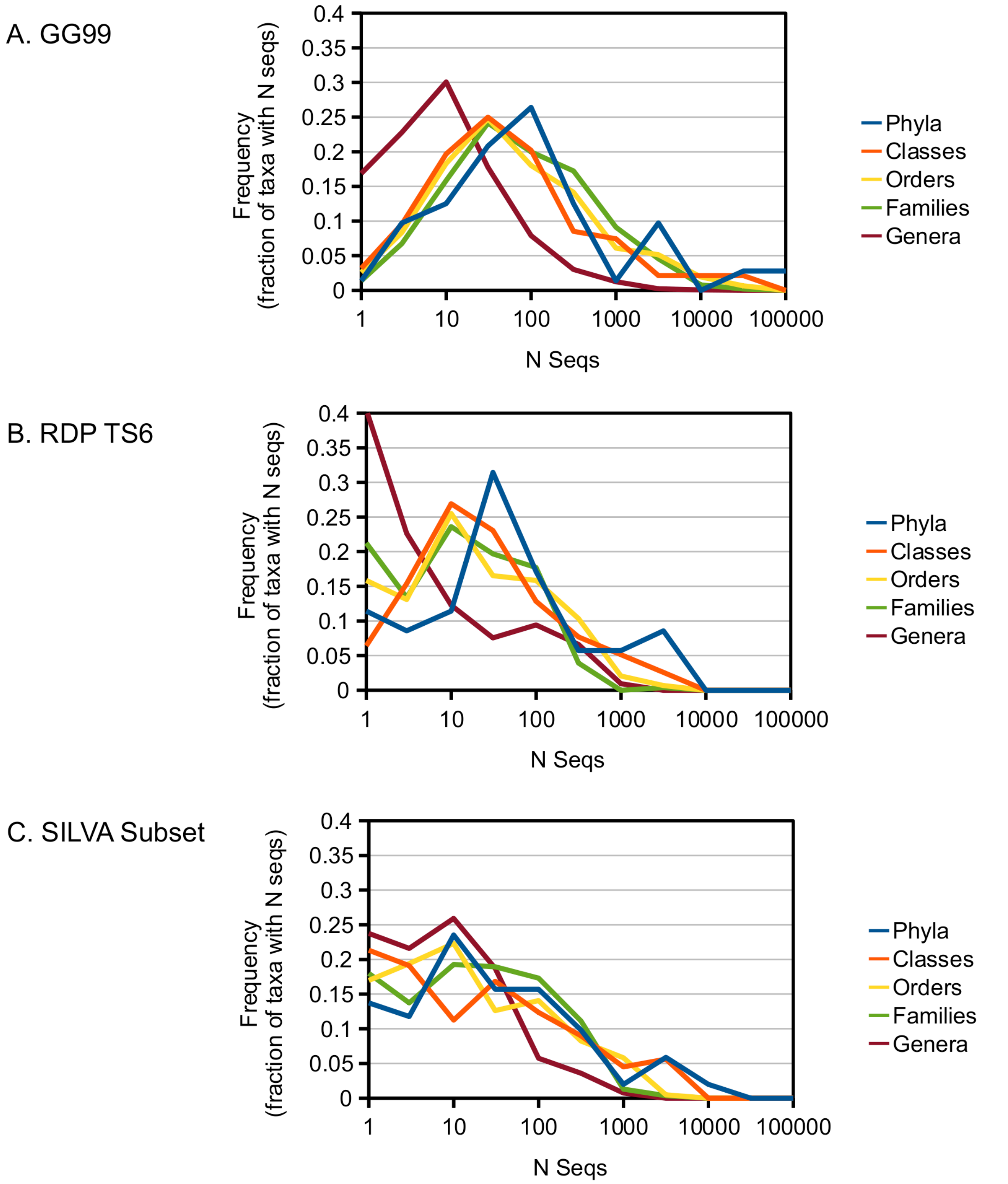
Jeffrey J. Werner[a]*, Omry Koren[b]*, Philip Hugenholtz[c], Todd Z. DeSantis[d], William A. Walters[e], J. Gregory Caporaso[e], Largus T. Angenent[a], Rob Knight[e,f] , Ruth E. Ley[b#]

a. Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY 14850, USA
b. Department of Microbiology, Cornell University, Ithaca, NY 14850, USA.
c. Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia
d. Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
e. Department of Biochemistry and Chemistry, University of Colorado, Boulder, CO 80309, USA.
f. Howard Hughes Medical Institute, University of Colorado, Boulder, CO 80309, USA.

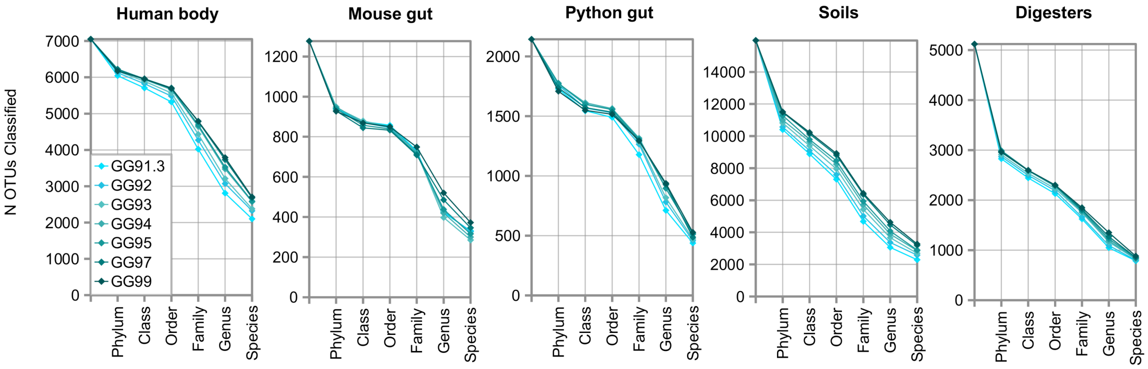
*Authors contributed equally
[#]Email: rel222@cornell.edu



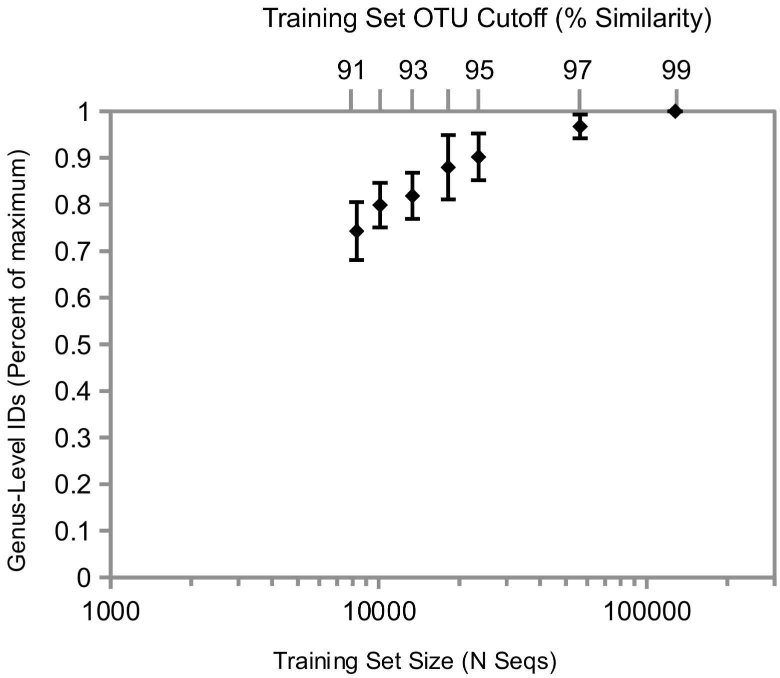
Supplementary Figure S1. Bacterial OTUs clustered at 90% identity (for visualization purposes) from the RDP Training Set 6 and the full Greengenes database. The 10% unique sequences were aligned to the Greengenes core alignment using PyNAST, and distances between sequences were calculated in mothur. The OTU distance matrix was then graphed using principal coordinates analysis (in QIIME) to depict the sequence variation among OTUs in each data set. The plots provide a visual representation of the diversity included in each reference data set.



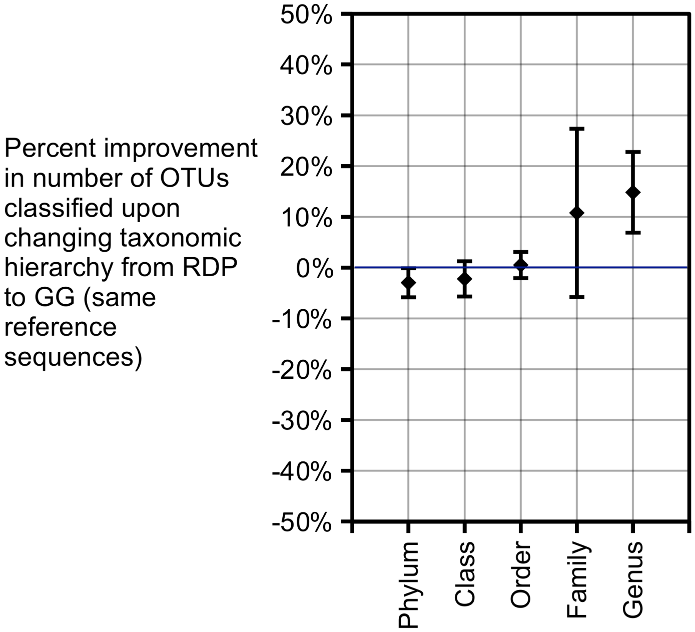
Supplementary Figure S2. Structure of the three primary training sets used in this work, as histograms showing the frequency of taxa containing a given number of sequences: (A) the GG99 training set, (B) RDP TS6, (C) the SILVA Subset.



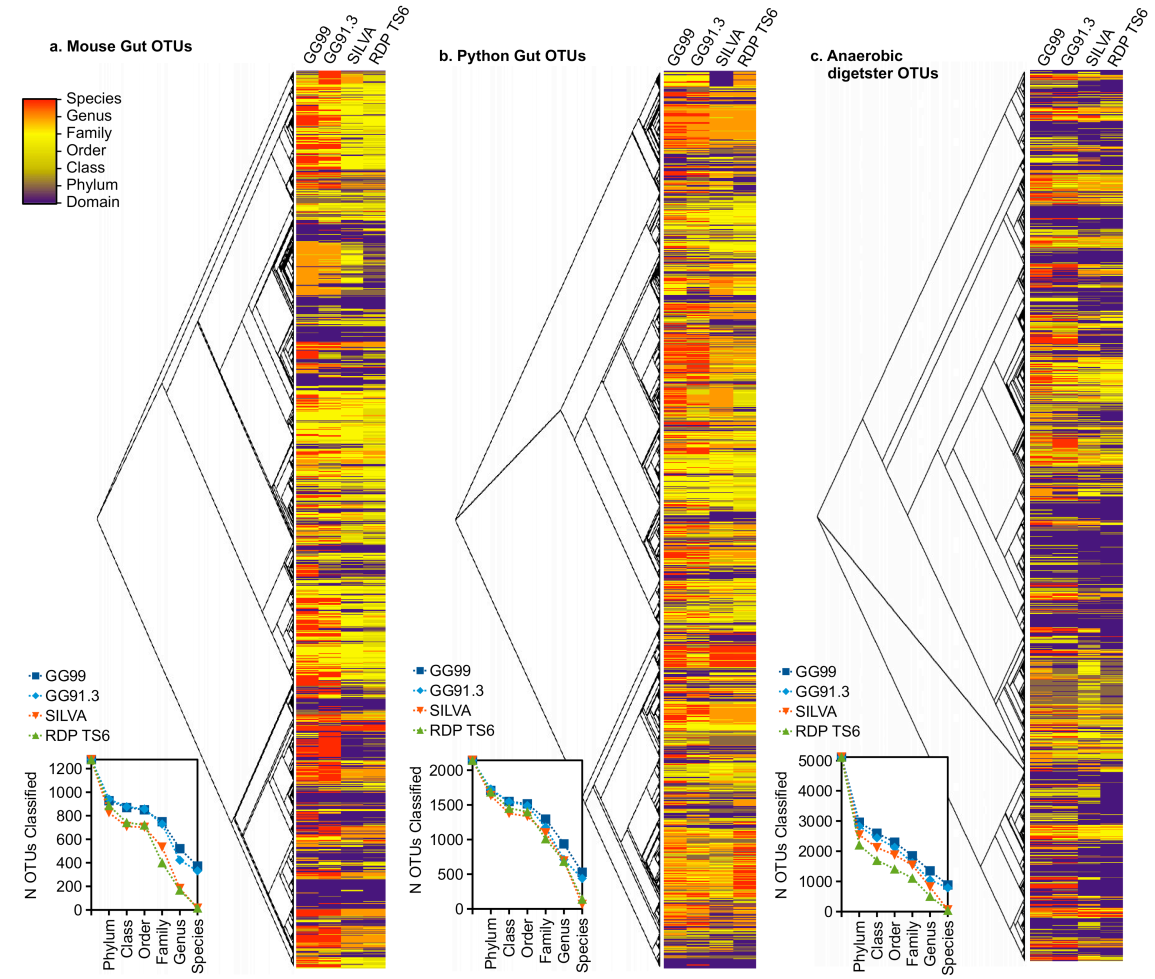
Supplementary Figure S3. Classification depth on a per-OTU basis, as a function of training set size. The Greengenes database was clustered at varying levels of similarity from 90% to 99% (5,832-128,127 total sequences), and was used to classify taxonomy in each of the five data sets.



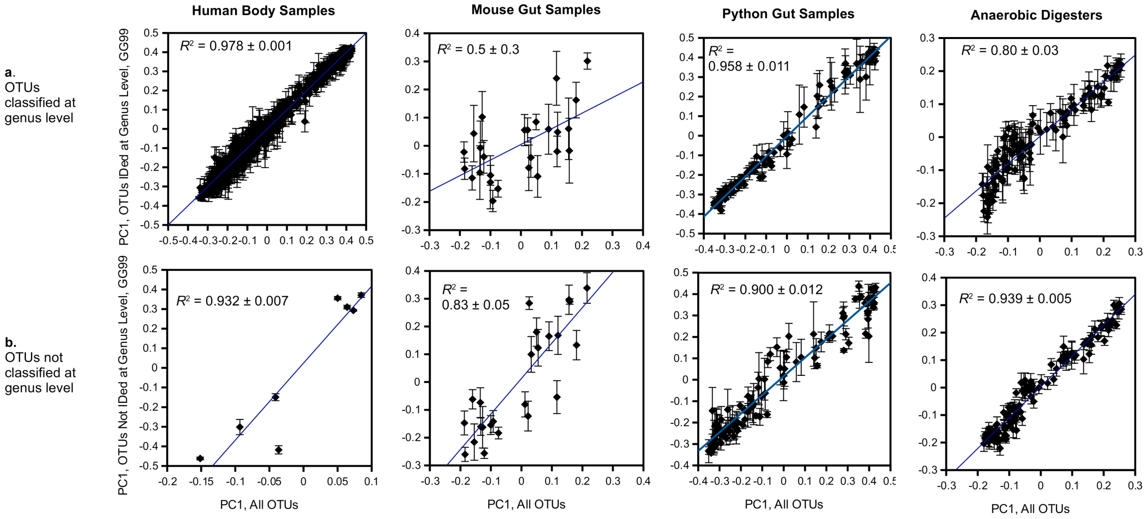
Supplementary Figure S4. The effect of training set size on the relative number of OTUs that were successfully classified at the Genus level, compared to the largest training set (GG99). Error bars represent the s.d. among the five different data sets (human body, mouse gut, python gut, soils, and anaerobic digesters).



Supplementary Figure S5. Comparison of taxonomic classification depth using the same RDP TS6 reference sequences and two different taxonomy outlines: RDP and Greengenes. Change in the number of OTUs classified at each level are plotted as the average percent increase comparing the Greengenes taxonomy to RDP. Error bars represent the s.d. among the five different query data sets (human body, mouse gut, python gut, soils, and anaerobic digesters).



Supplementary Figure S6. Summary of OTU classification depth using each of the three training sets for three of the five studies: (a) mouse gut OTUs, (b) python gut OTUs, and (c) anaerobic digester OTUs (other two data sets are shown in main text, Figure 2). OTUs are organized according to evolutionary history, as determined by the FastTree approximately-maximum-likelihood tree constructed in the default QIIME pipeline. Inset charts summarize the total number of OTUs classified at each taxonomic level (GG99 – dark blue, GG91.3 – light blue, SILVA - orange, RDP TS6 - green).



Supplementary Figure S7. Correlations between PC1 values computed with all OTUs versus only the OTUs that were either (a) identified at the genus level, or (b) not identified at the genus level, by the GG99 training set. Error bars, and errors on R^2 values, represent the s.d. of 10 rarefactions, 200 sequences each. Human body, mouse gut, python gut, and anaerobic digester samples are shown here. Soil samples are shown in the manuscript.