Supplemental Online Results:

Functional, phylogenetic, and computational determinants of prediction accuracy using reference genomes

A series of tests determined the relationship between PICRUSt's prediction accuracy and the functional content, phylogenetic composition or annotation accuracy of the genes and metagenomes being inferred. We first assessed several methods for predicting the presence or absence of KOs within genomes, including nearest neighbor and several ancestral state reconstruction-based methods. All methods outperformed our random method that inferred the genome content by taking the average presence or absence for each function across all genomes (mean=0.77 +/- 0.05 s.d). Methods based on ASR significantly outperformed simply taking the functional content of the nearest neighbor among sequenced genomes (Wilcoxon rank sum test p< 2.2e-16; ASR mean=0.955 +/- 0.04 s.d , Nearst Neighbor mean=0.940 +/- 0.06 s.d; Supplemental Fig. 5 and Supplemental Fig. 6). Although the magnitude of the difference in balanced accuracy (1.5%) is modest, it represents a large difference in error rates because both methods perform well. Overall, ASR reduced error rates by ~33%. ASR methods also allow for the calculation of 95% confidence intervals on PICRUSt's gene content prediction. These confidence intervals capture uncertainty in gene content prediction owing to a variety of factors (see Discussion). Characterization when closely related reference genomes are plentiful or unavailable suggest that in either case these confidence intervals are slightly conservative, but otherwise accurately capture uncertainty in gene content predictions (Supplemental Fig. 7). Given these results, the phylogenetically independent contrasts (PIC) method[50] was chosen because of its fast computation time, ability to generate confidence intervals that reflect the strength of evidence for prediction (Supplemental Fig. 7), slight tendency towards specificity over sensitivity (Supplemental Fig. 6), as well as recent successes in accurately predicting 16S rRNA gene copy number using this method[51].

PICRUSt's accuracy in predicting gene contents (COGs or KOs) for each organism correlated with the phylogenetic distance from that organism to a sequenced genome (Spearman r=0.75, p<0.001; Supplemental Fig. 8**)**. This is consistent with the observed correlation between the availability of closely related reference genomes (as captured through NSTI scores) and metagenome prediction accuracy (Figure 3). The 14 genomes with an associated accuracy <0.75 were either poorly annotated draft genomes (n=6), reduced intracellular endosymbionts (n=6) or isolates recently sequenced by the Genomic Encyclopedia of Bacteria and Archaea because of their phylogenetic novelty[52] (n=2). We found that although PICRUSt was robust to substantial rearrangement of the tips of the 16S phylogenetic tree (Supplemental Fig. 9), large errors in phylogenetic placement could still cause poor performance. For example, *Coxiella burnetii* RSA 334 was reconstructed with the worst balanced accuracy (0.61), but closer inspection showed that this was likely due to issues with contamination or incorrect annotation rather than the inference method; its reference phylogenetic placement branches from the Archaea instead of its proper placement within the Gammaproteobacteria.

Prediction of 16S rRNA copy number using PICRUSt improves estimation of microbial relative abundance.

High-throughput sequencing of 16S rRNA marker genes is extensively used to characterize microbial communities. One criticism of such approaches is that marker genes vary in copy number (e.g. between 1 and 15 according to previous studies and IMG annotations)[51, 53, 54]. Therefore, the relative abundance of 16S genes could, in theory, be as much as 15-fold different from the relative abundance of organisms. Because PICRUSt allows inference of the gene copy number of an organism from its 16S rRNA sequence, PICRUSt provides predictions of 16S rRNA copy number that can be used to convert the relative abundance of 16S rRNA sequences into an estimated relative abundance of organisms. These estimates are used by PICRUSt internally when estimating metagenomes, but the same script may also be used separately to normalize 16S rRNA analyses for differences in copy number[51] prior to analysis of the microbial community.

To assess the accuracy of PICRUSt's predictions of 16S rRNA copy number, we employed a cross-validation approach on all finished bacterial and archaeal genomes in the IMG database (1412 annotations total). In this cross-validation, a test dataset was constructed for each genome. Each of these test datasets excluded the genome to be predicted, but contained all other annotated genomes in our reference set. For each test dataset, the 16S rRNA copy number for the test organism was predicted using PICRUSt, and the predicted value compared against the actual copy number (Supplemental Fig. 14).

PICRUSt predictions were well correlated with actual annotations (Supplemental Fig. 14; Pearson r^2 = 0.787; mean absolute error +/- 0.62 copies). We expected that this prediction accuracy would depend on the evolutionary distance separating the predicted organism from its closest relative in which copy number is known. To quantify the effect of this distance, PICRUSt predictions were tested against additional cross-validation datasets which excluded all annotated neighbors within a particular distance of the predicted organism (Supplemental Fig. 15). These datasets were constructed for all distances between 0.0 and 0.30 units of branch length, in increments of 0.03 units of branch length on the reference Greengenes phylogeny. These evaluations revealed that PICRUSt prediction of 16S copy number retains substantial accuracy (Supplemental Figs. 15 and 16; Pearson r^2=0.64; Mean absolute error +/- 1.2 copies), even when copy number annotations are unavailable for close relatives (within 0.06 units of branch length, corresponding to roughly the same genus). Having characterized error in PICRUSt's predictions, we also wanted to test whether distance to the nearest reference genome might bias PICRUSt estimates. To test for bias based on availability of reference data, we plotted the absolute error in PICRUSt's 16S rRNA copy number predictions against the minimal distance to the nearest reference genome (Supplemental Fig. 17). The results show no apparent trend between the minimal distance to the nearest reference genome and the absolute error in the prediction, indicating that PICRUSt predictions are not systematically biased upwards or downwards by the availability of reference genomes.

Supplemental Fig. 15 compares PICRUSt accuracy to the simpler approach of predicting the copy number of the most closely related genome ('nearest neighbor prediction'). These characterizations indicate that depending on the average distance to references, PICRUSt and nearest neighbor approaches may perform identically, or PICRUSt may perform somewhat better on average (Supplemental Fig. 15, Blue vs. Green lines). Therefore nearest neighbor prediction may be useful for rapidly achieving an approximate prediction prior to using PICRUSt's more detailed evolutionary method. Regardless of the exact prediction method selected, the efficacy of PICRUSt's approach on any given dataset can be estimated ahead of time by calculating the average distance to a reference genome (Nearest Sequenced Taxon Index; see main text) using PICRUSt's built-in tools, and comparing against Supplemental Figs. 15 and 16.

In addition to a slight average improvement over nearest neighbor methods, PICRUSt results improve substantially on the default approach of predicting the average copy number (or, equivalently, a random copy number) in terms of both absolute error and correlation. Predicting the average copy number produces an average absolute error of +/- 2.98 16S rRNA gene copies for this dataset (vs. +/- 0.62 copies for PICRUSt predictions). Moreover, while predicting a fixed copy number produces no correlation between true and observed copy number (r^2 = 0.0; Supplemental Fig. 15), PICRUSt predictions explain most variation in 16S rRNA copy number, so long as reasonably related reference genomes are available (NSTI <= 0.09). It is worth noting that the poorly performing fixed number approach is equivalent to the current state-of-the-art for 16S studies, which generally assume that 16S rRNA relative abundance is identical to organismal relative abundance (see Kembel *et al.*[51] for a more detailed discussion of this issue). Indeed even when the nearest annotated relative is constrained to be at least 0.27 units of branch length from the predicted genome (~ 9x as divergent as members of the same 'species-level' OTU), the mean absolute error (+/- 1.97 copies) is still much lower using PICRUSt predictions than predicting a fixed number. We therefore expect that PICRUSt normalization of OTU tables by estimated 16S rRNA copy number should improve estimates of microbial relative abundance for many communities.

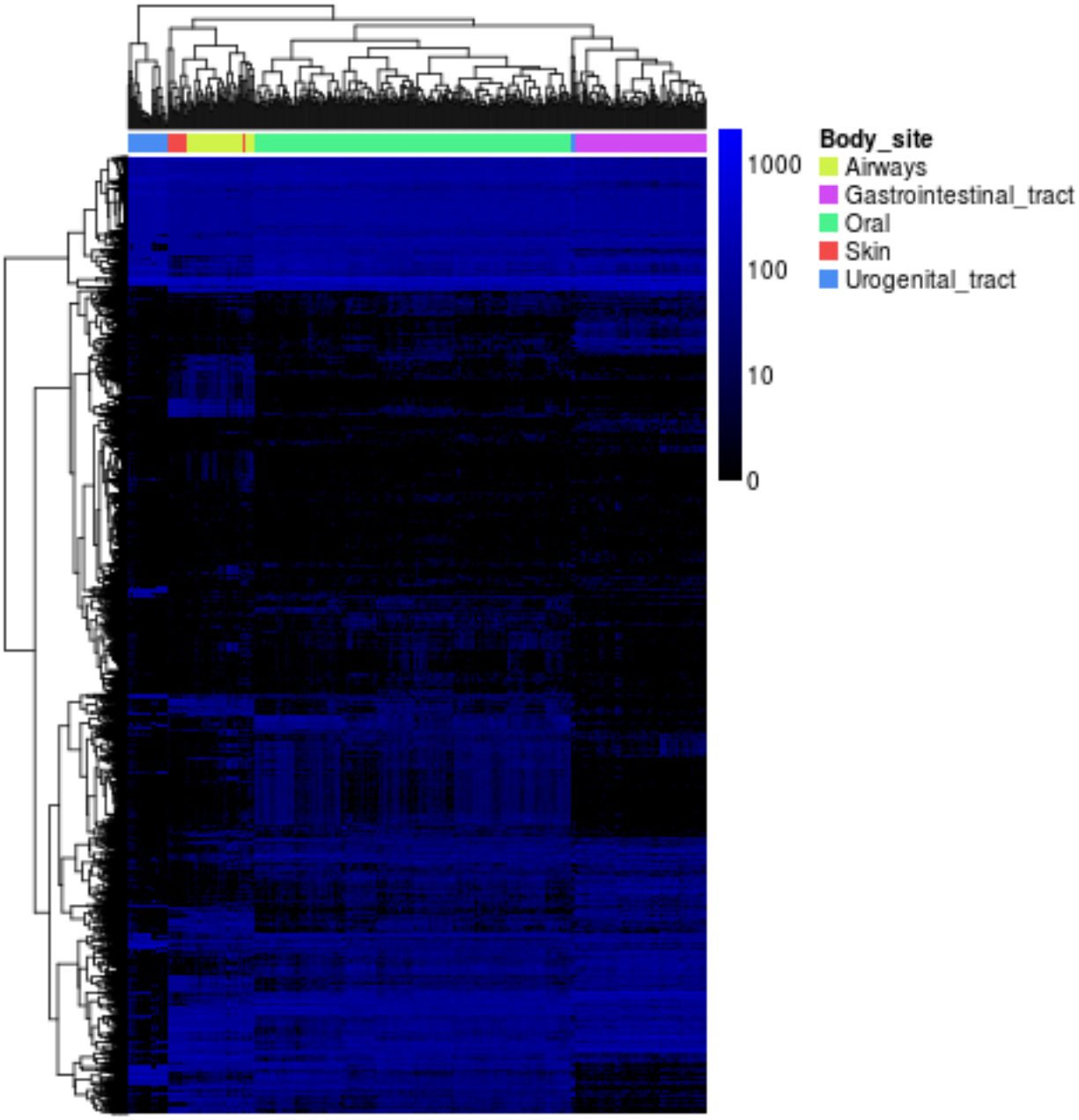
Kembel *et al.*[51] recently reported a method for estimating 16S rRNA copy number in unsequenced bacteria, as well as the application of that estimate to normalize 16S rRNA libraries for variance in 16S copy number between species. The method described relies on an algorithm derived from ancestral state reconstruction that is similar to PICRUSt. Kembel *et al.* report that the Pearson correlation (r) of 0.81 for true vs. estimated copy number, corresponding to an r^2 of ~0.66[51]. This is somewhat lower than PICRUSt results when using all IMG 16S rRNA copy number annotations as a reference (PICRUSt r^2 = 0.79 using 1,412 annotations), but still higher than our results when restricting reference genomes to be outside the genus level (exclusion distance = 0.06 16S rRNA substitutions/site; r^2 = 0.64). Kembel *et al.*[51] used a smaller reference set (881 16S rRNA gene copy number annotations in the full dataset, of which 484 were incorporated into pruned trees for cross-validation), which may tend to result in more distant reference genomes being used in any given prediction. Therefore, it seems likely that the somewhat higher correlation between true and estimated copy numbers when predicting with PICRUSt is due to the larger reference dataset used or possibly the details of the cross-

validation analysis, rather than any algorithmic differences (which appear to be relatively minor). These results therefore provide independent confirmation of the efficacy of 16S rRNA gene copy number estimation using evolutionary techniques.

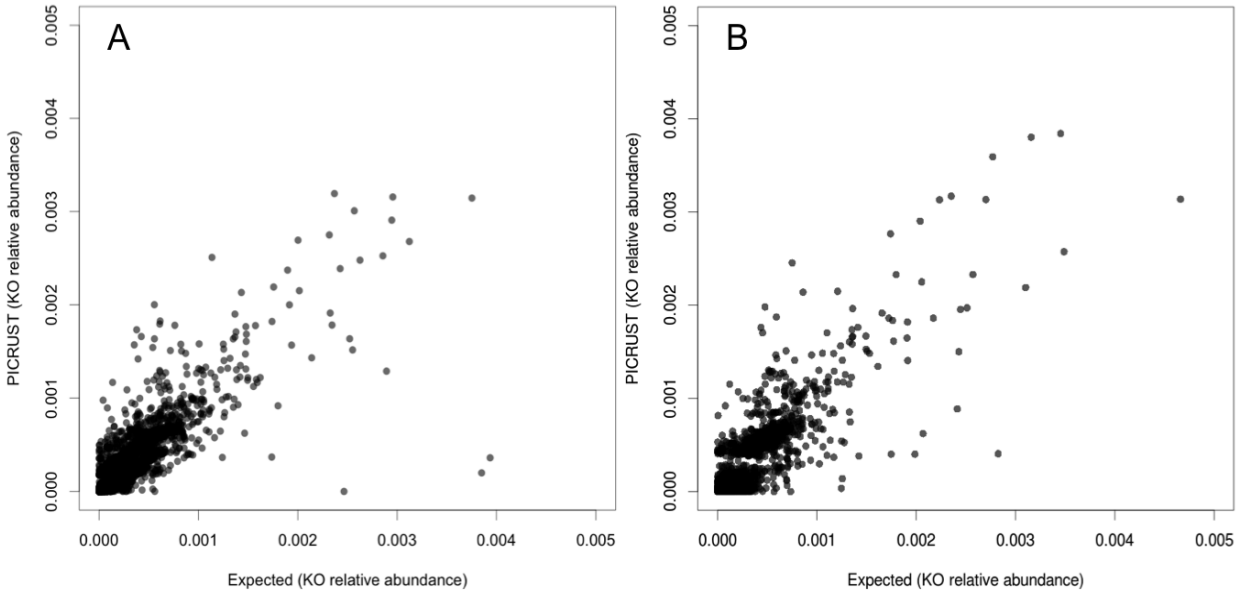
Supplemental References

50. Felsenstein, J. Phylogenies and the Comparative Method. *The American Naturalist* **125**, 1 (1985).
51. Kembel, S.W., Wu, M., Eisen, J.A. & Green, J.L. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS computational biology* **8**, e1002743 (2012).
52. Wu, D. et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056-1060 (2009).
53. Klappenbach, J.A., Dunbar, J.M. & Schmidt, T.M. rRNA operon copy number reflects ecological strategies of bacteria. *Applied and environmental microbiology* **66**, 1328-1333 (2000).
54. Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V. & Polz, M.F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *Journal of bacteriology* **186**, 2629-2635 (2004).

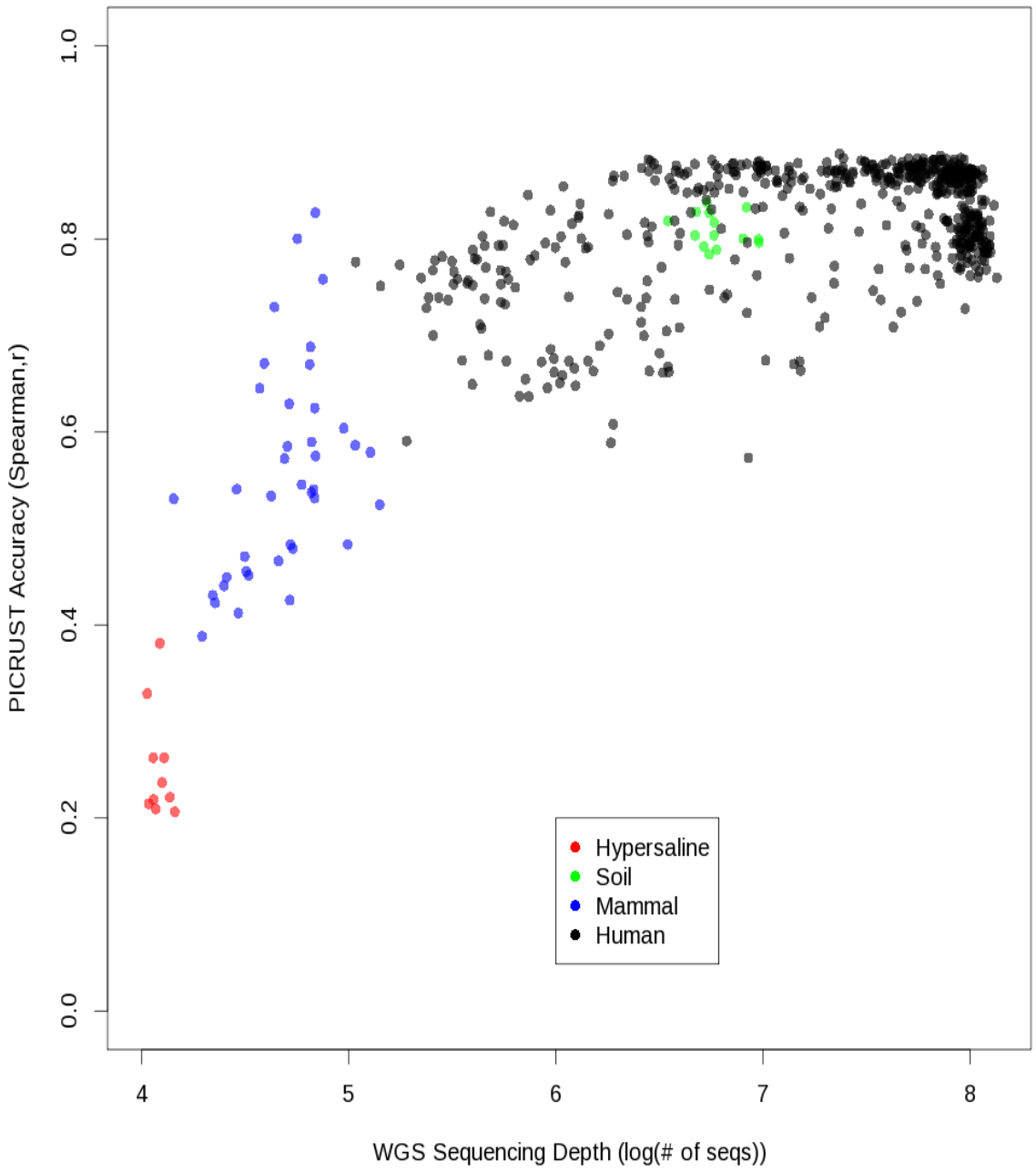
Supplemental Figures



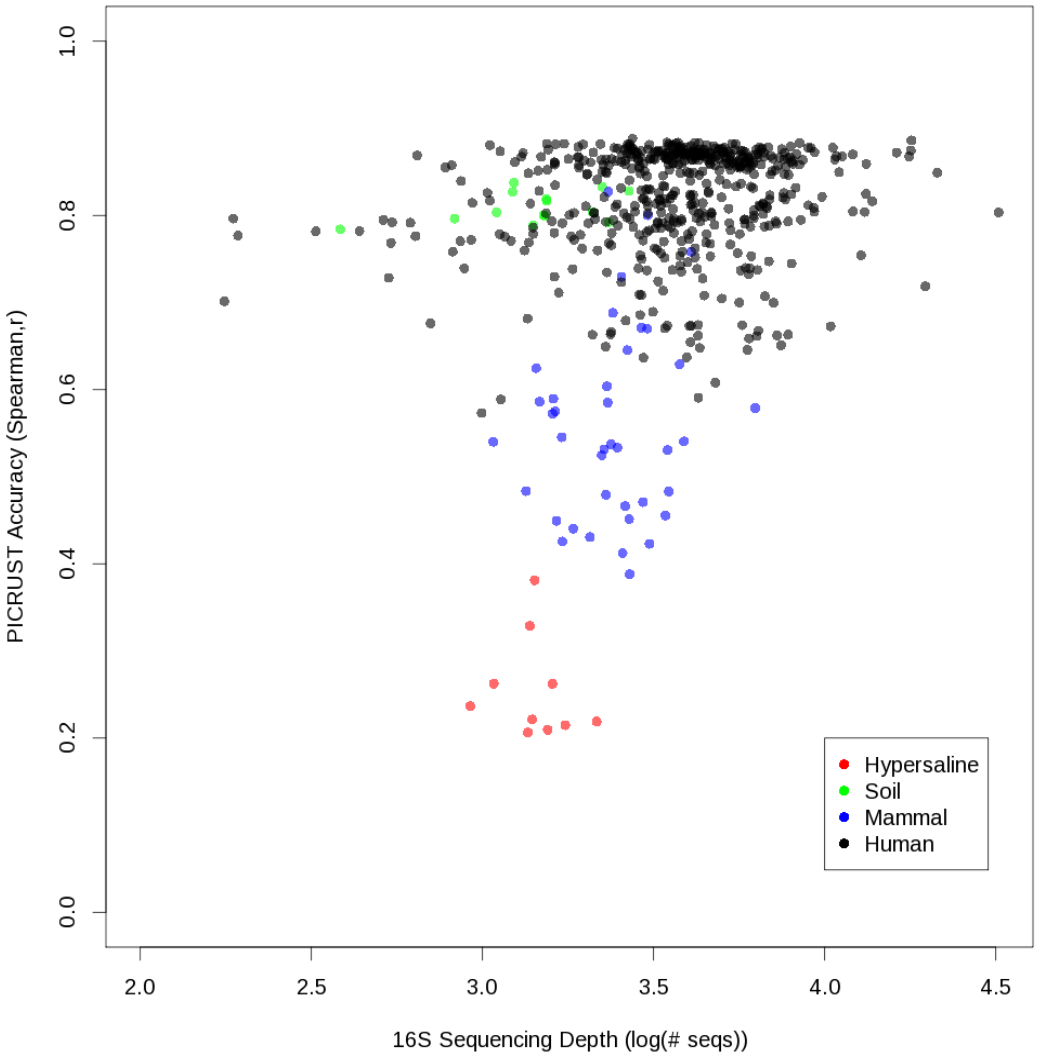
Supplemental Figure 1. Individual gene family (KEGG Ortholog) abundances predicted by PICRUSt in 530 HMP microbiomes. Columns represent samples, and rows represent the 4000 KOs with highest average abundance (for visualization) of 6,885 total gene families (KOs) predicted by PICRUSt for these HMP body sites. Samples and KOs are hierarchically clustered using Euclidean distance and complete linkage. Blue colored intensity represents the abundance of each KO on a log scale (see legend).



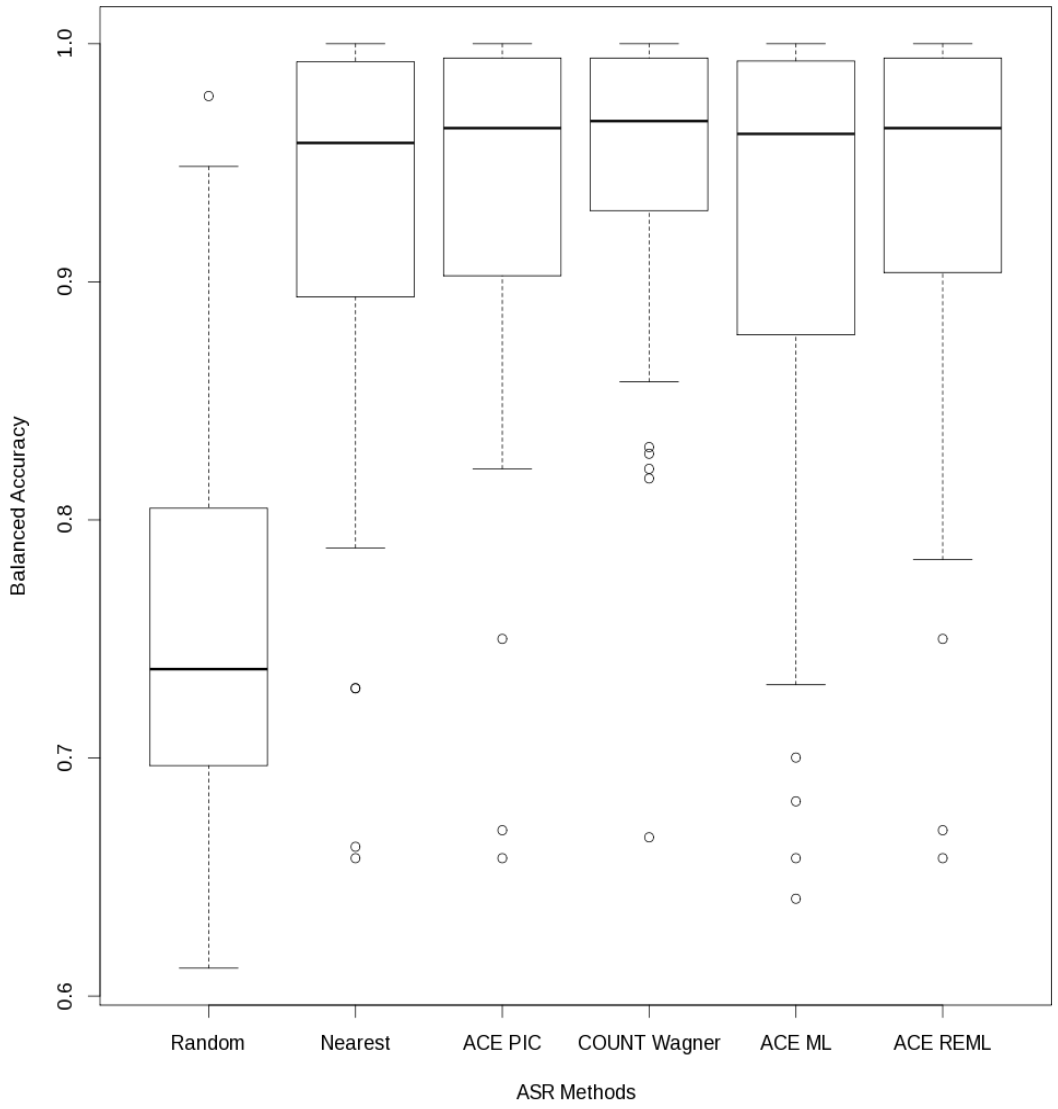
Supplemental Figure 2: Prediction accuracy within the HMP "mock community" data sets. A) "Even" community and B) "staggered" community. Each point is the relative abundance of a KEGG Ortholog (KO) gene family for PICRUSt predictions based on 16S data (y-axis) and that expected from metagenomic sequencing (x-axis).



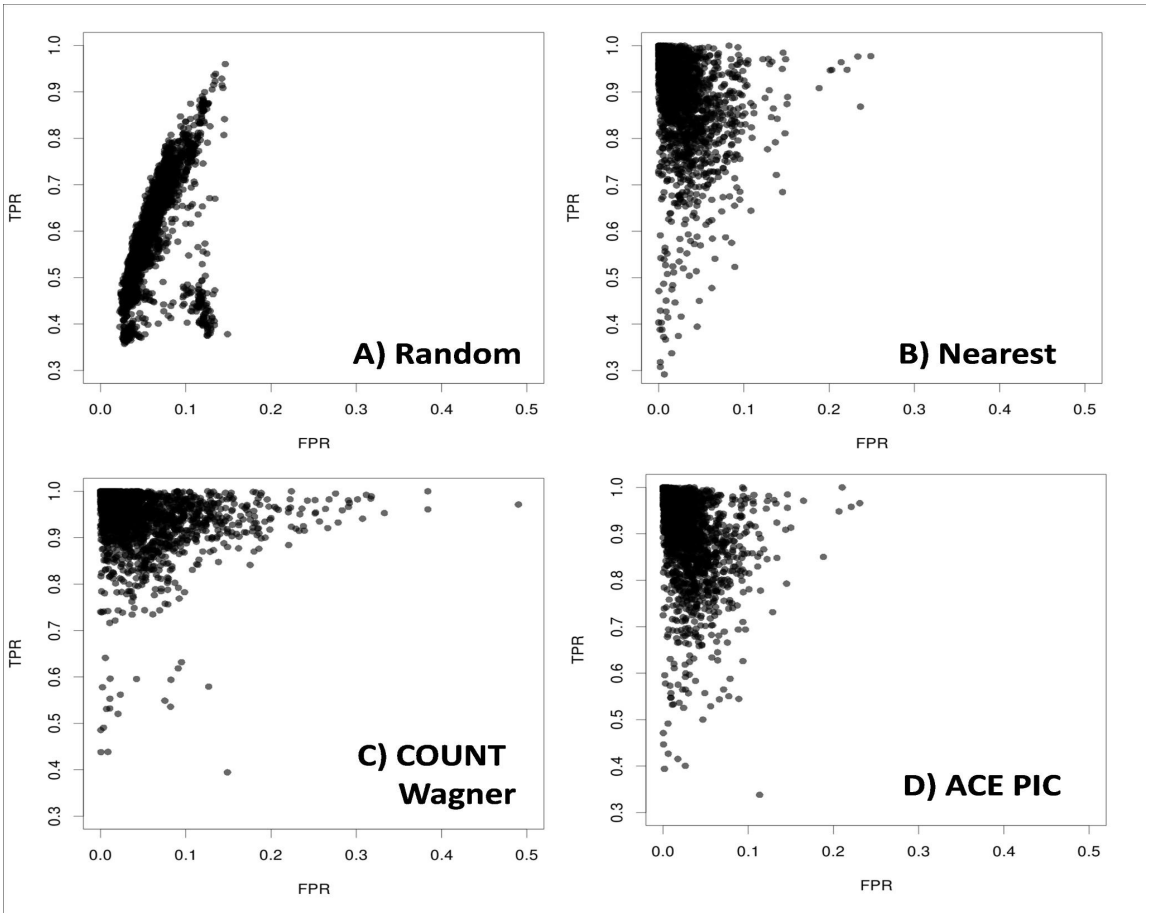
Supplemental Figure 3: **Effect of metagenomic sequencing depth on PICRUSt accuracy.** Paired shotgun metagenomes and 16S rRNA gene surveys are colored by study as indicated in legend. Cross-validation studies presented here use sequenced shotgun metagenomes as a control for PICRUSt predictions. However, undersampling of an underlying biological metagenome by shotgun sequencing can lead to a poor perceived accuracy of PICRUSt in our cross validation analysis, since the resulting sequenced metagenome may not properly represent the true metagenome. See also the metagenome rarefaction analysis (Fig. 4, dashed red line). PICRUSt accuracy values presented here therefore represent conservative estimates.



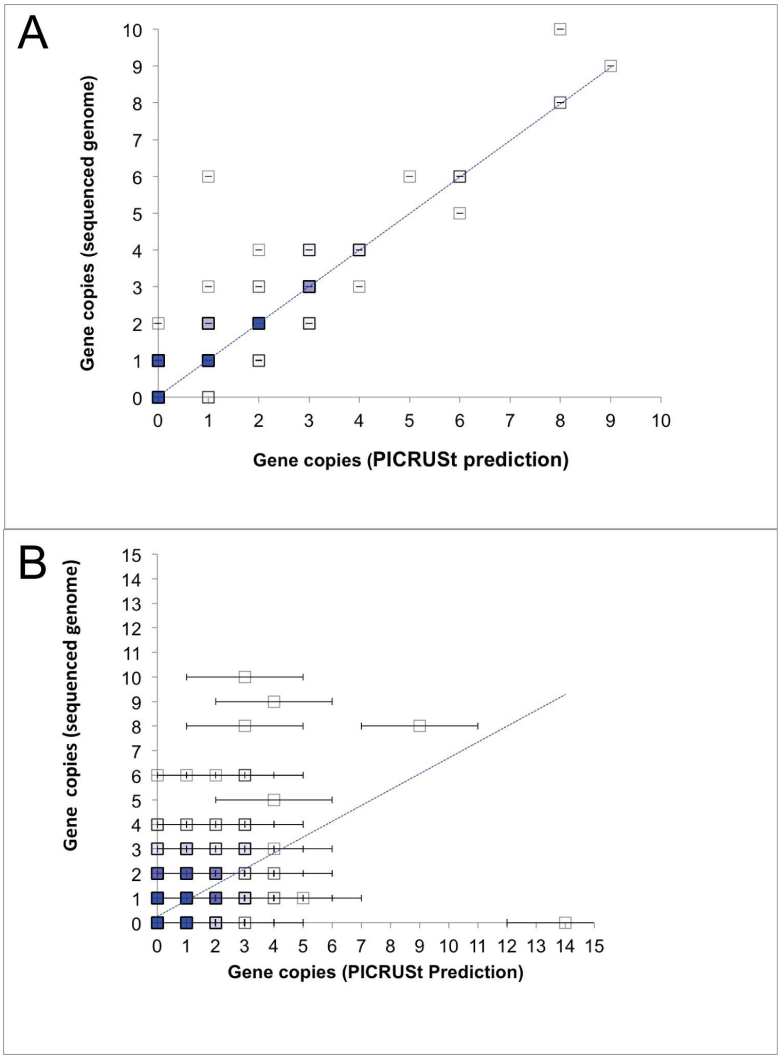
Supplemental Figure 4: **Effect of 16S sequencing depth on PICRUSt accuracy.** Paired shotgun metagenomes and 16S rRNA gene surveys are colored according to study as indicated in legend. Compared to shotgun metagenomic sequencing depth (Supplemental Fig. 3), 16S sequencing depth has relatively little effect on PICRUSt accuracy, likely due to the lower number of distinct 16S sequences in most communities relative to the number of gene that must be sampled across the metagenome (See also the rarefaction analysis in Fig. 4).



Supplemental Figure 5. Influence of ancestral state reconstruction method on PICRUSt genome prediction accuracy. “Random” uses the mean abundance of the KO from a random genome for its prediction (this has the same mean accuracy as predicting the average KO value). “Nearest” uses the KO profile from the nearest reference genome in the reference 16S phylogenetic tree. Other methods use ancestral state reconstruction methods as indicated by their labels, along with a novel weighting method to extend predictions from ancestral nodes to tips in the tree (see Methods S1). Due to the long computation time required to calculate ancestral states across the tree of life using maximum likelihood methods (ACE ML and ACE REML), accuracy was evaluated on a subset of 100 random traits and genomes with accuracy distributions representative of the entire dataset (data not shown). The ACE PIC ASR method was chosen as the default ASR method in PICRUSt due its speed, ability to create confidence intervals for each prediction, and because it is significantly more accurate than the nearest neighbor approach (mean “ACE PIC”=0.955; mean “Nearest”=0.940; Wicoxon rank sum test p< 2.2e-16). Note that all methods are available as options in the PICRUSt software.

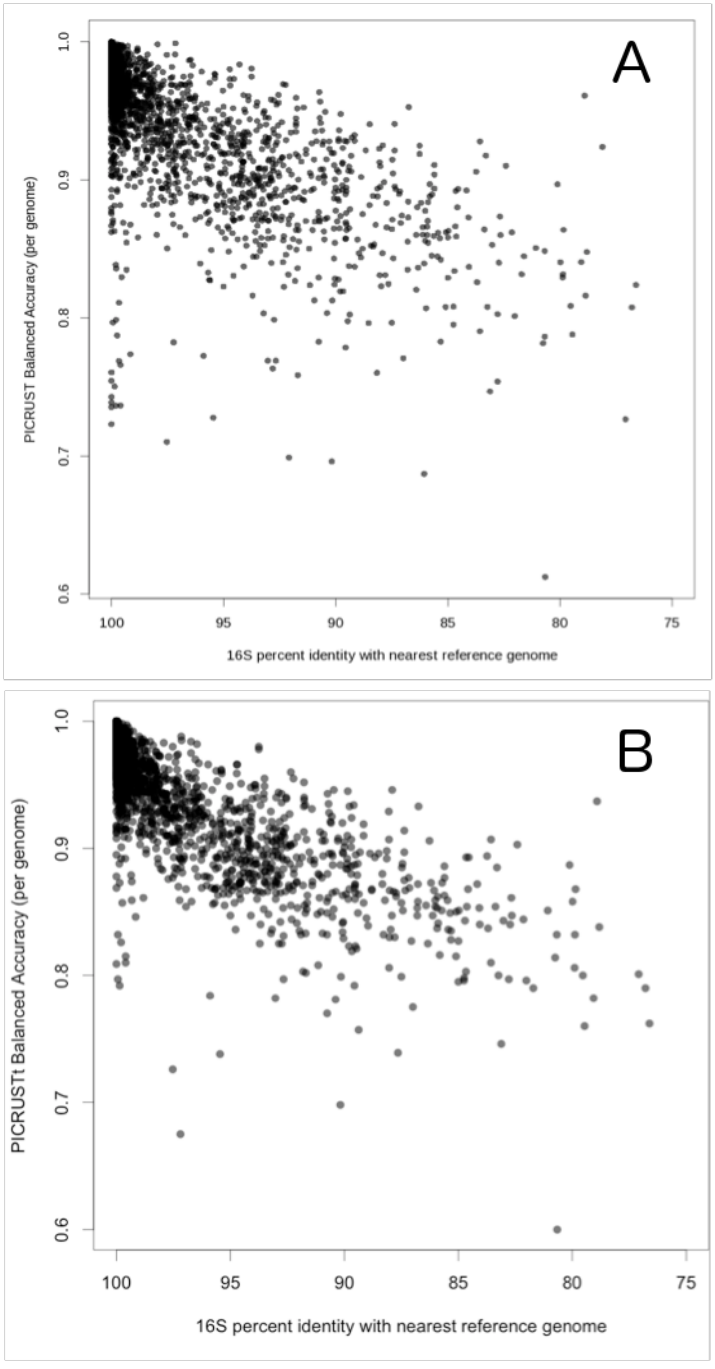


Supplemental Figure 6. True positive rate (TPR) versus false positive rate (FPR) for various methods of genome content prediction. Several methods were tested for their effect on a genome content prediction accuracy with the best performing genomes being in the top left point of each plot (high TPR (y-axis) and low FPR (x-axis)). The **"**ACE PIC" method for ancestral state reconstruction shows generally lower FPR relative to the similarly accurate COUNT Wagner method (see Supplemental Fig. 5). "Random" and "Nearest" are shown for reference and are described in Supplemental Fig. 5.

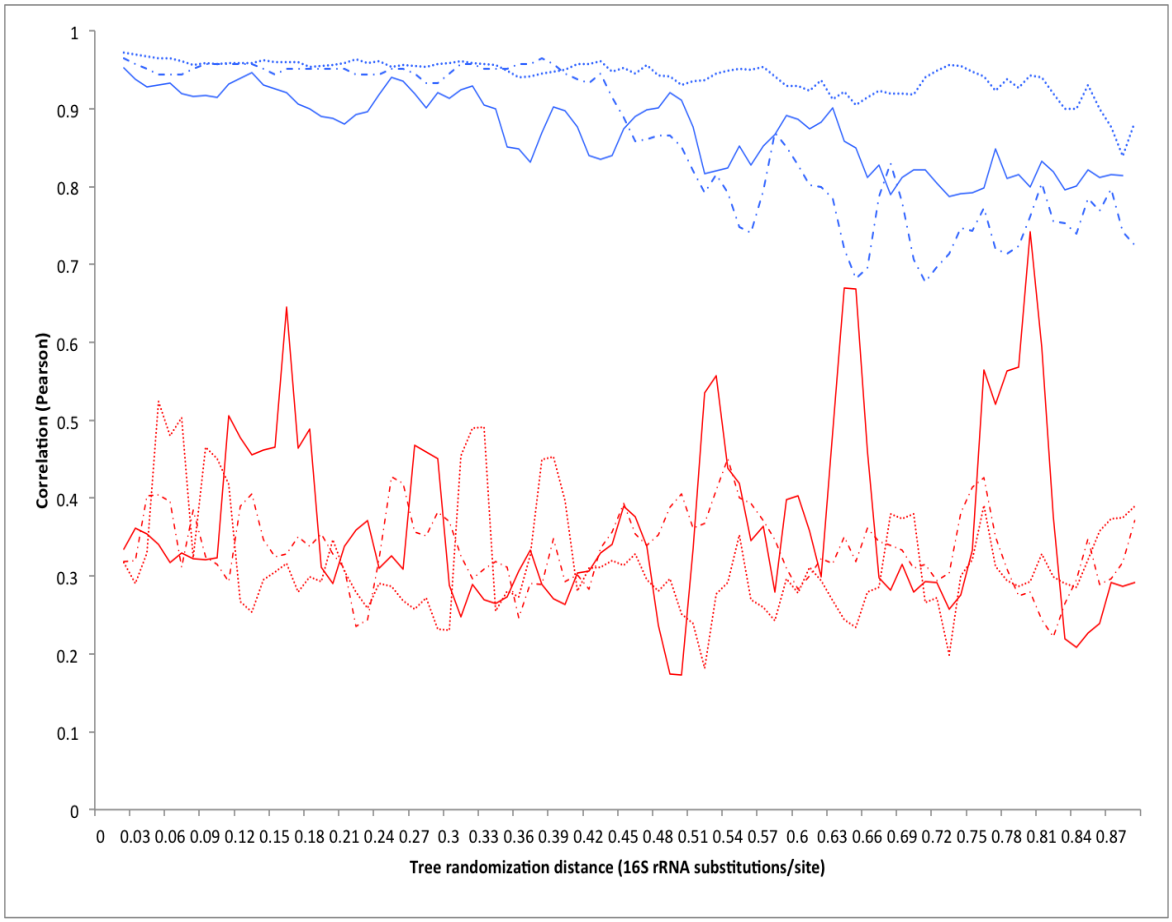


Supplemental Figure 7. Confidence intervals for PICRUSt prediction using sufficient or insufficient reference datasets. Illustration of confidence intervals for per-genome prediction, using prediction of *E.coli* K12 MG1655 (IMG 646311926) as an example. The axes represent true gene copy numbers from genome sequence data vs. PICRUSt predicted gene copy numbers. Points are filled to 1% transparency per occurrence, so darker points reflect common gene copy numbers. Blue dotted lines reflect linear regression of PICRUSt predictions vs. expected values **A)** Prediction using all other IMG bacterial genomes (Pearson r^2 = 0.93; Balanced Accuracy = 0.977). When all other reference genomes are available for prediction, error bars are extremely narrow (< 1 gene copy). 97.7% of genes fall within PICRUSt's 95% confidence intervals for prediction with the full dataset in panel **A**, indicating the CI is slightly conservative **B)** When all genomes within 0.20 16S rRNA substitutions/site are excluded (e.g. a very poor prediction- NSTI >= 0.20; Pearson r^2 = 0.31; Balanced Accuracy = 0.77), error bars widen to reflect uncertainty. Although some individual gene copy numbers are predicted incorrectly, aggregate values are conservative- even using this extremely limited training set (panel **B**), 99.5% of genes fall within the confidence intervals. Similar tests applied to *Bacteroidetes thetaiotaomicron* VPI-4582 (IMG 637000026), and *Pelagibacter ubique* HTCC 1062 (IMG 637000058*)* also produced empirical CIs that were slightly conservative (ranging

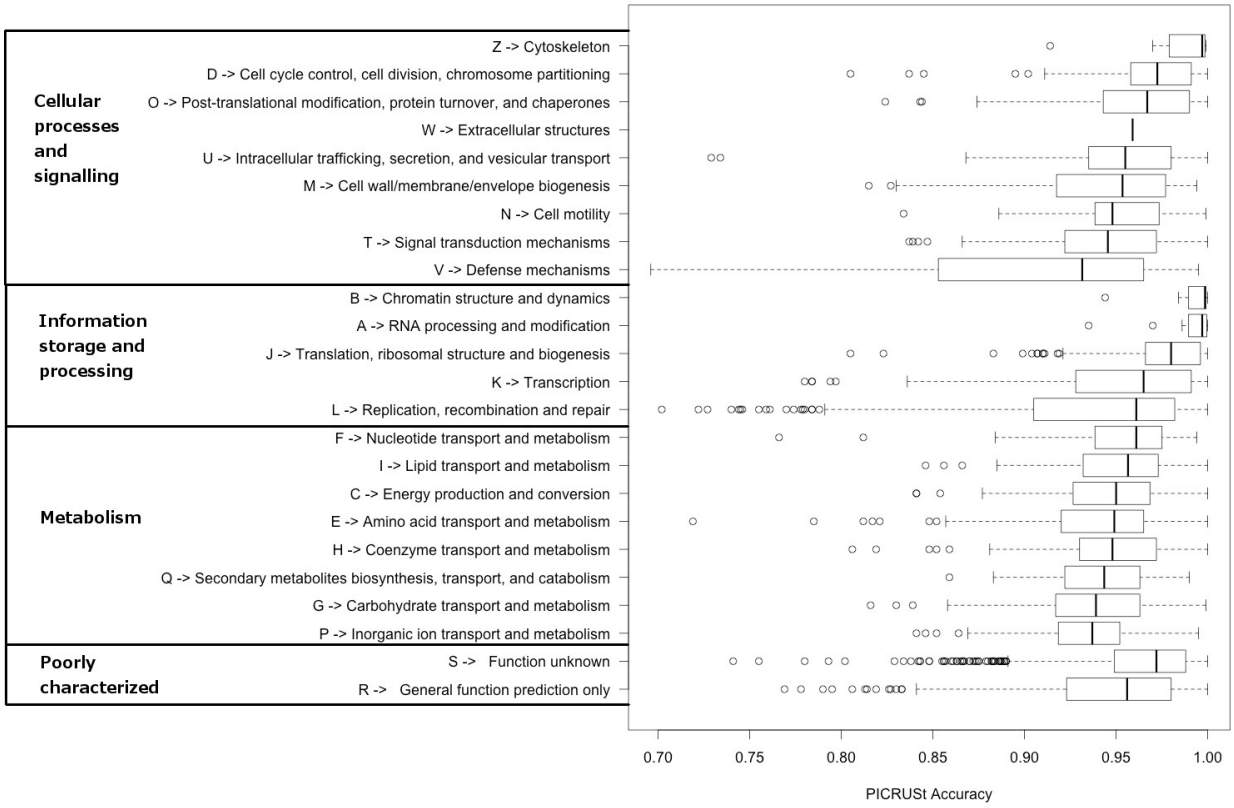
from 97.4-99.9% effective CIs). Separate testing of upper and lower confidence bounds found that in all cases the CIs were generally accurate, though slightly conservative (empirical CIs were in all cases >= 95%), with lower bounds being somewhat more conservative than upper bounds because genes could not be present in fewer than 0 copies.



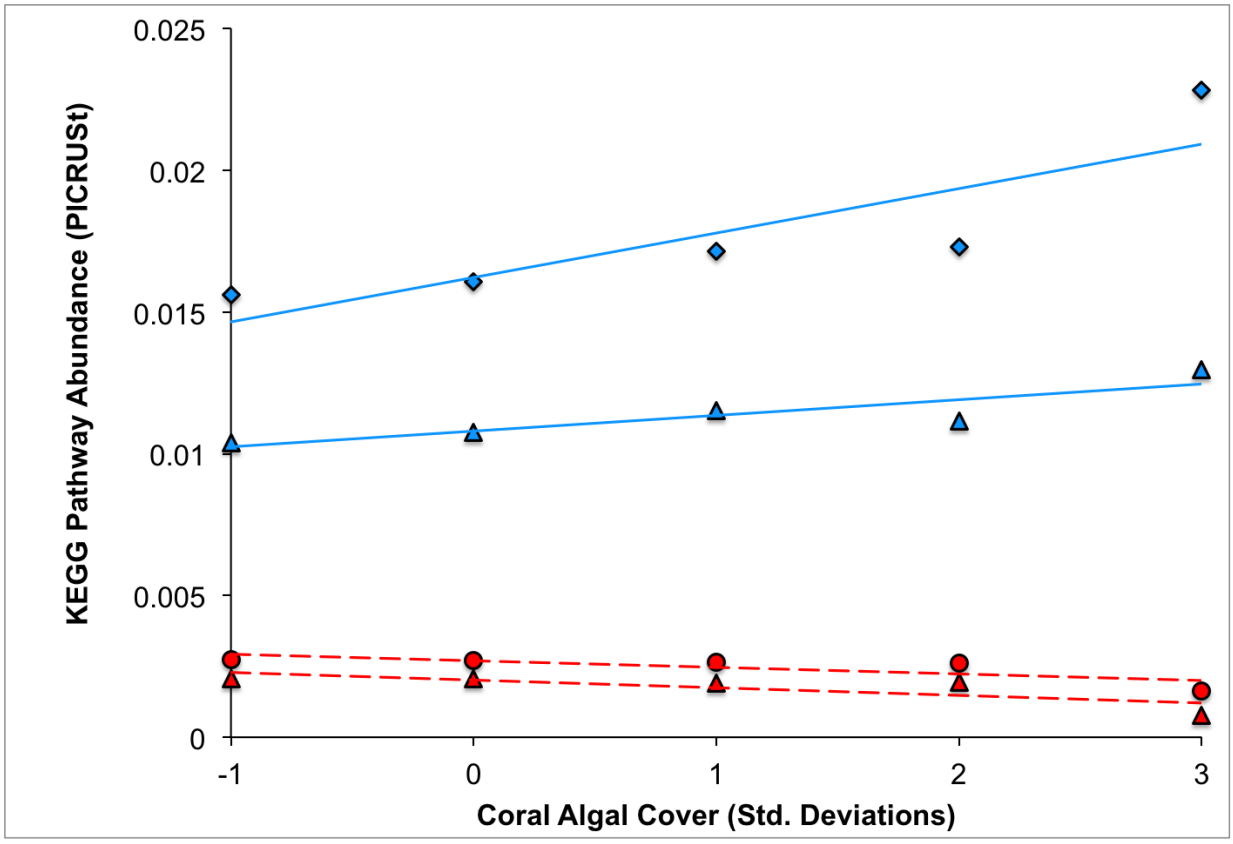
Supplemental Figure 8. Genome prediction accuracy with respect to distance to nearest sequenced reference genome. Plot shows trend of being able to predict the content of each genome in IMG based on the 16S distance to its nearest reference genome (based on phylogenetic distance) for both KOs (A) and COGs (B). Outliers (balanced accuracy<0.75) are a combination of reduced genomes and poorly annotated draft genomes (see text).



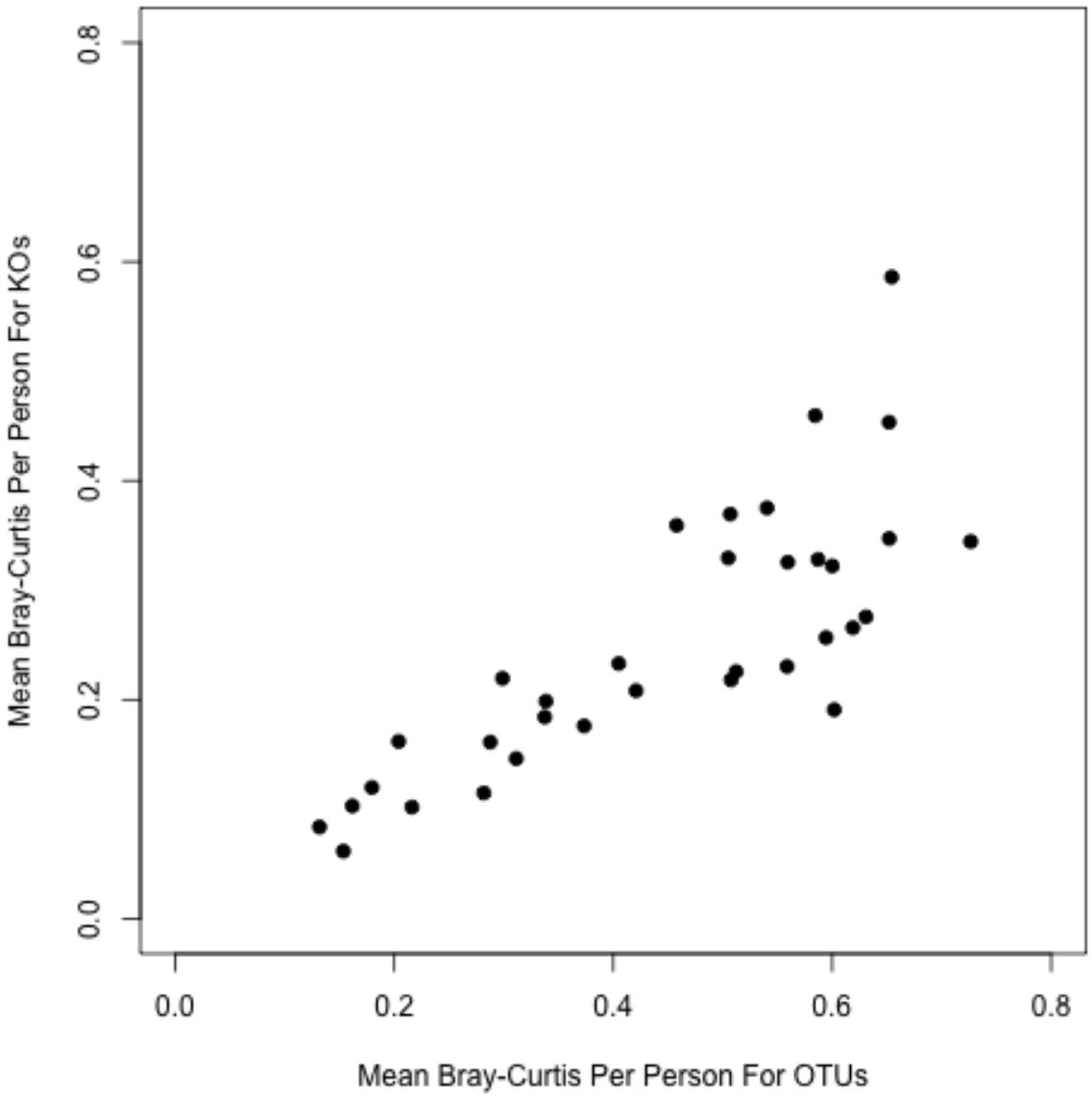
Supplemental Figure 9. Effects of errors in genomes' phylogenetic placement on genome prediction accuracy. Each series represents predictions of the genome of a sequenced organism with increasing levels of added phylogenetic error (*Escherichia coli* K12 MG1655, solid; *Bacteroidetes thetaiotaomicron*, dashed; *Pelagibacter ubique*, dotted). This distance-based holdout procedure has the effect of simulating the difficulty in predicting the contents of a known genome using PICRUSt if the phylogeny describing its relatives was incorrect. Blue lines represent PICRUSt predictions, red lines represent the accuracy of predicting a random genome from the same Greengenes subtree (for reference). For visual clarity each line is a moving average (period 3) of the results). In each trial, neighbors at increasing distances to the genome to be predicted had their phylogenetic placement scrambled (increments of 0.01 units of branch length, across the range 0.0 to 0.90). The x-axis thus indicates the distance within which phylogenetic placement was scrambled, and the y-axis the Spearman correlation coefficient for PICRUSt predictions (blue) of gene family abundance vs. actual (IMG annotated, red) values.



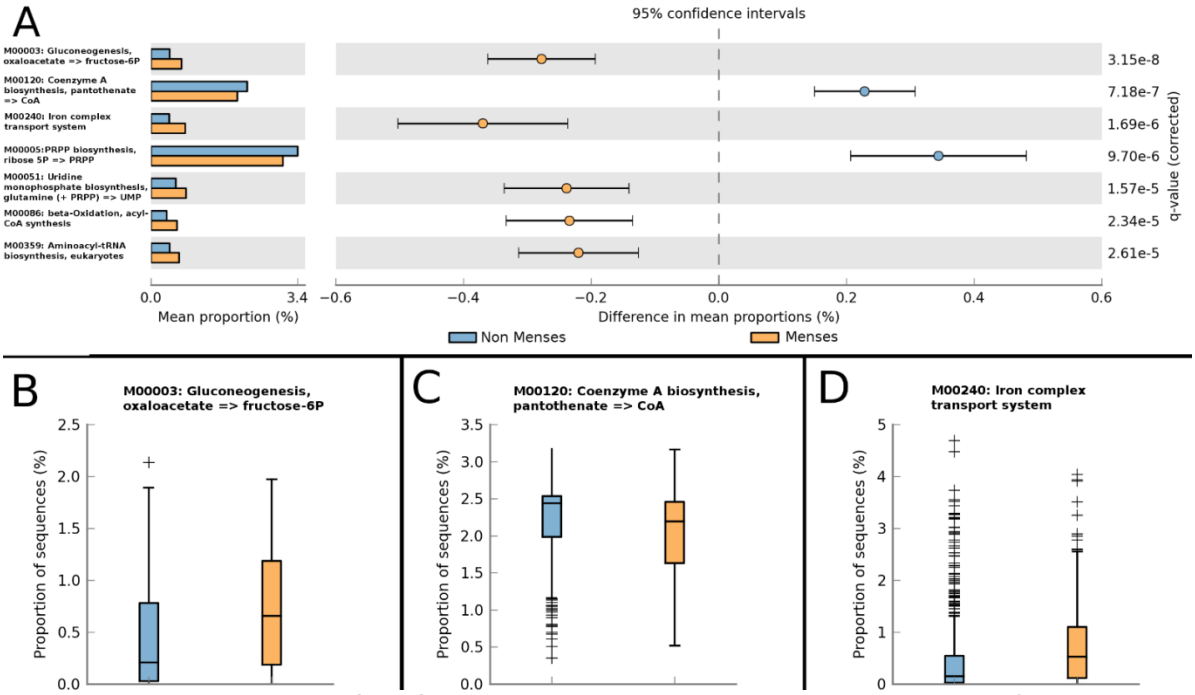
Supplemental Figure 10. Accuracy of COG functions using genome holdout evaluation. The ability of PICRUSt to predict the presence/absence of each individual COG orthologous families at the gene level was evaluated using genome holdouts (see Methods). Each COG family was then grouped into its corresponding higher-level category (i.e. letter code). The resulting distributions of Spearman correlation accuracies is shown here. In agreement with the KO analysis (Fig. 6), ion transport and carbohydrate metabolism have slightly decreased accuracy. In addition, COGs that are likely laterally transferred such as restriction endonucleases, which are found in the defense mechanisms category, also show decreased accuracy.



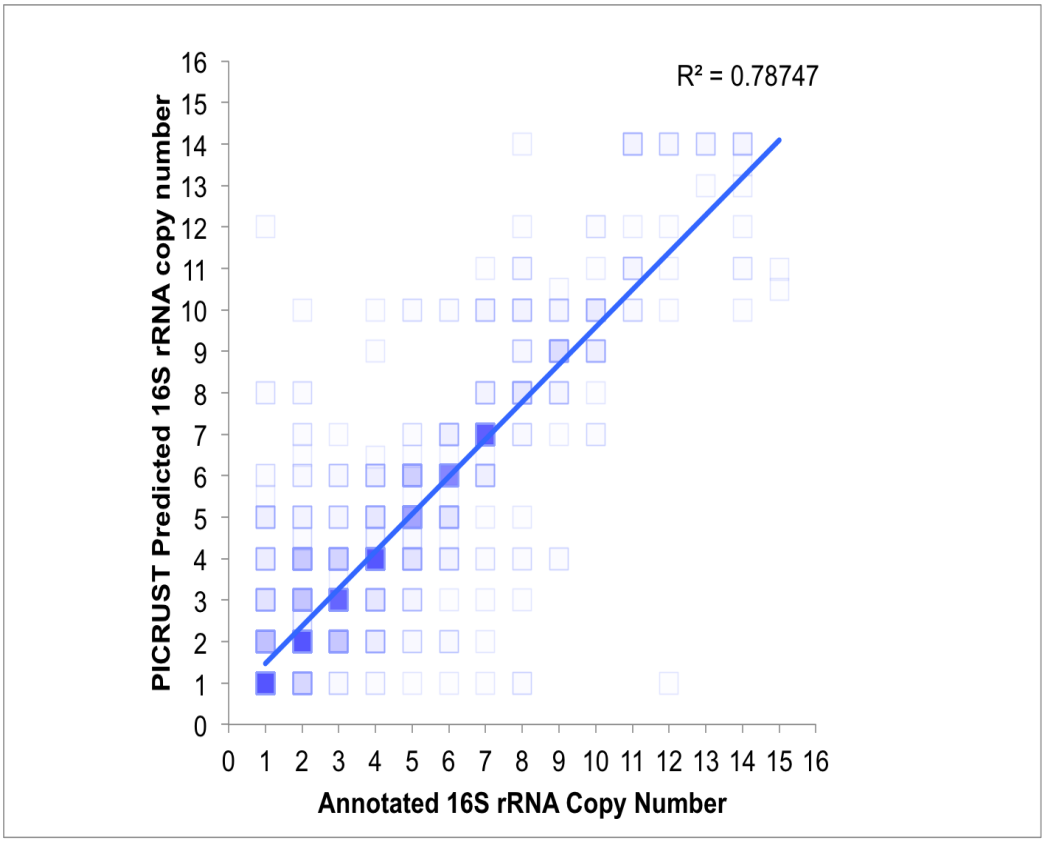
Supplemental Figure 11. PICRUSt analysis of algal cover and predicted gene frequency in reef-building corals. 16S rRNA gene amplicons from 335 coral mucus DNA samples from an ongoing *in situ* experimental intervention in the Florida Keys were sequenced with 454 pyrosequencing, analyzed in QIIME using standard workflows, and converted to predicted gene abundances using PICRUSt (weighted NSTI = 0.12 +/- 0.02 s.d.) Those abundances were then summarized using KEGG Pathways. Relative algal cover reflects benthic quadrant surveys of algal cover. Scores are normalized to mean algal cover and reported as z-scores. Each of these pathways varied significantly by algal cover (FDR-corrected ANOVA; $q < 0.05$). For all pathways shown, Spearman regression against algal cover using transformed data (as shown) was significant for all categories ($p < 0.05$; $r^2 > 0.80$). The raw (non-transformed) algal cover data was also correlated with all KEGG pathways; all shown categories were independently identified as significant in that analysis (FDR-corrected regression $q < 0.05$) with the exception of "Secretion systems", which attained q=0.057. Blue diamonds: secretion systems; Blue triangles: ribosomal biogenesis; Red circles: Carbohydrate metabolism - Galactose metabolism; Red triangles: Carbohydrate metabolism - Ascorbate and alderate metabolism.



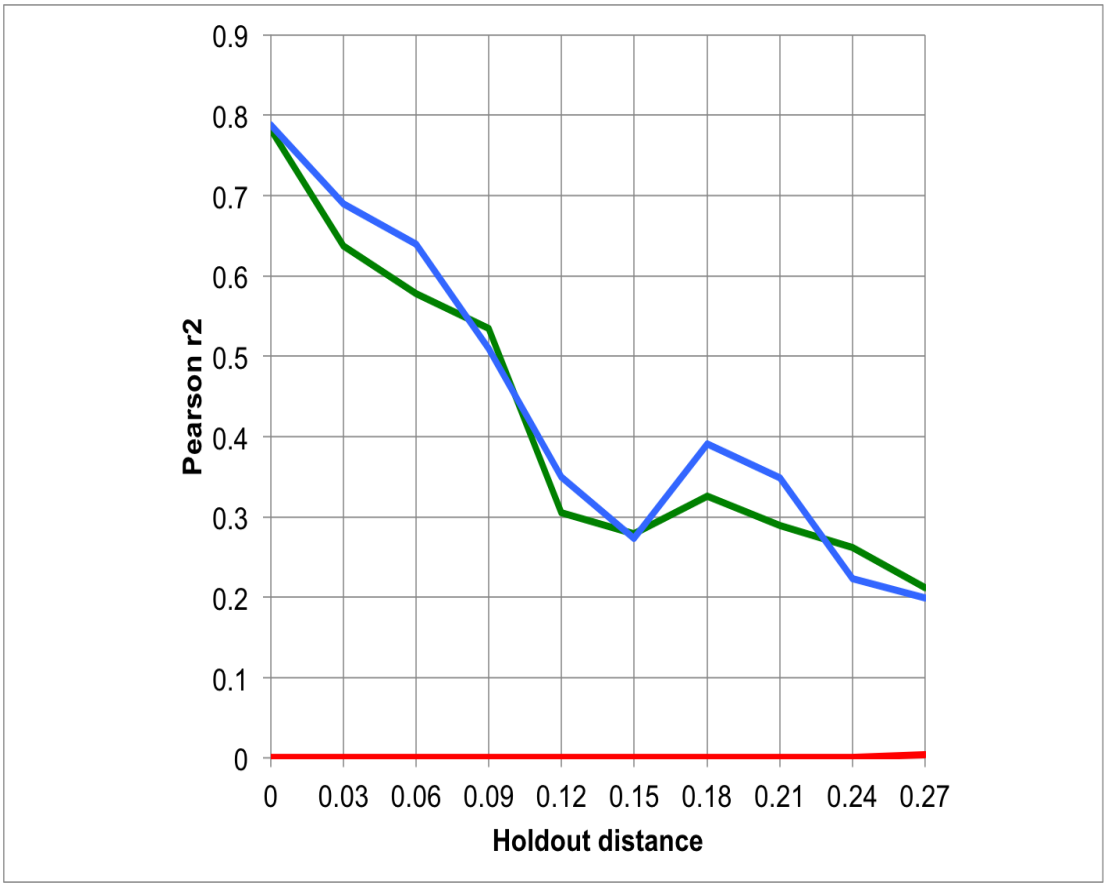
Supplemental Figure 12. Within-subjects beta-diversity of microbial composition versus inferred gene content for vaginal samples. Each point represents the mean Bray-Curtis dissimilarity between all samples from the same individual using either OTUs (x-axis) or PICRUSt predicted KOs (y-axis). In all cases, longitudinal stability is greater when considering KO gene content than when using OTU taxonomic composition (i.e. all points fall below the diagonal).



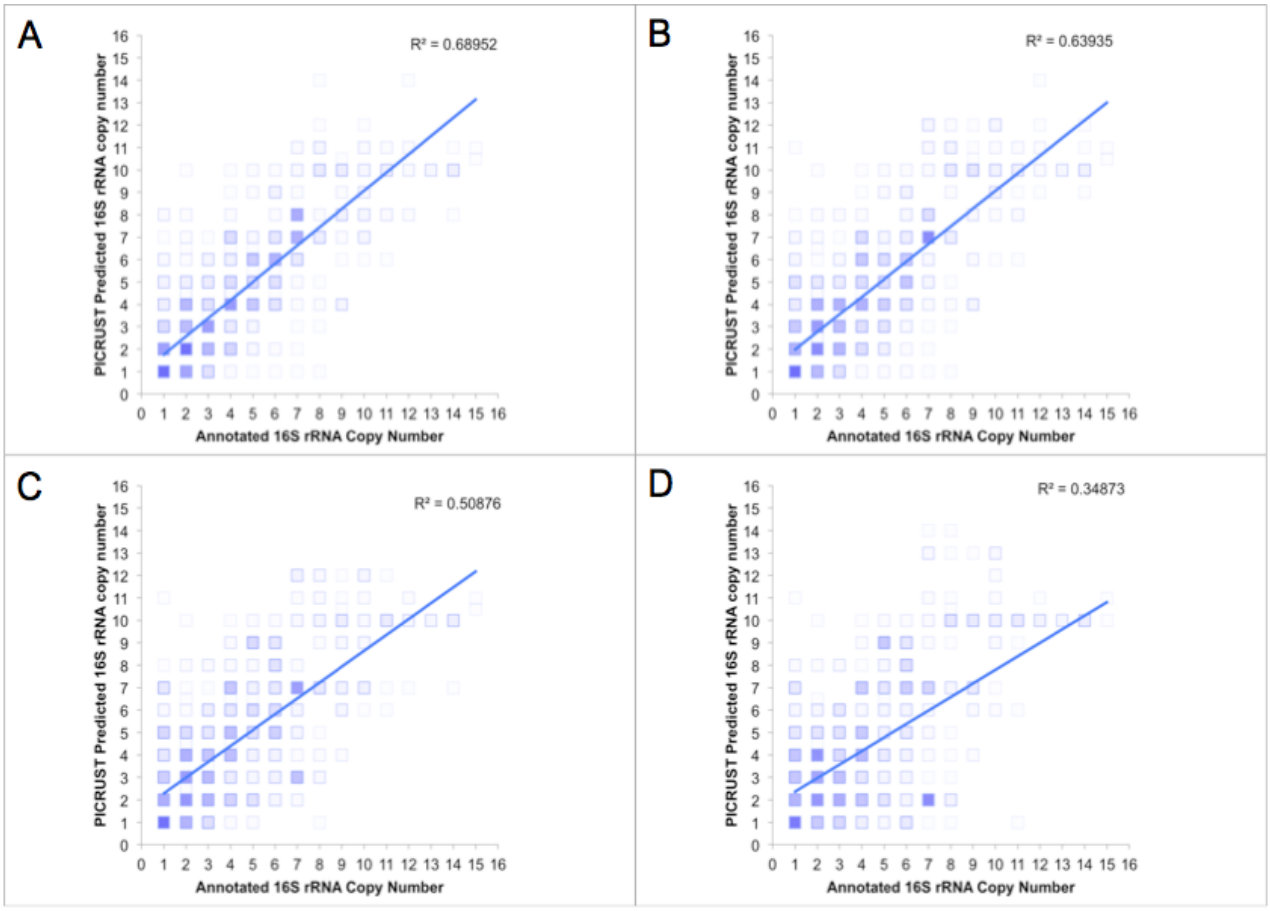
Supplemental Figure 13. PICRUSt predicted metabolic pathways with significant differences in relative abundance during menses. A) Seven KEGG Modules were found to have significant difference (+/- 0.2%) in mean proportions of vaginal samples during menses (Welch t-test with FDR q<0.0001). Boxplots of the three most significant modules show the detailed differences between samples taken during non-menses (blue; n=802) and menses (orange; n=191): B) M00003: Gluconeogenesis, oxaloacetate => fructose-6P, C) M00120: Coenzyme A biosynthesis, pantothenate => CoA, and D) M00240: Iron complex transport system.



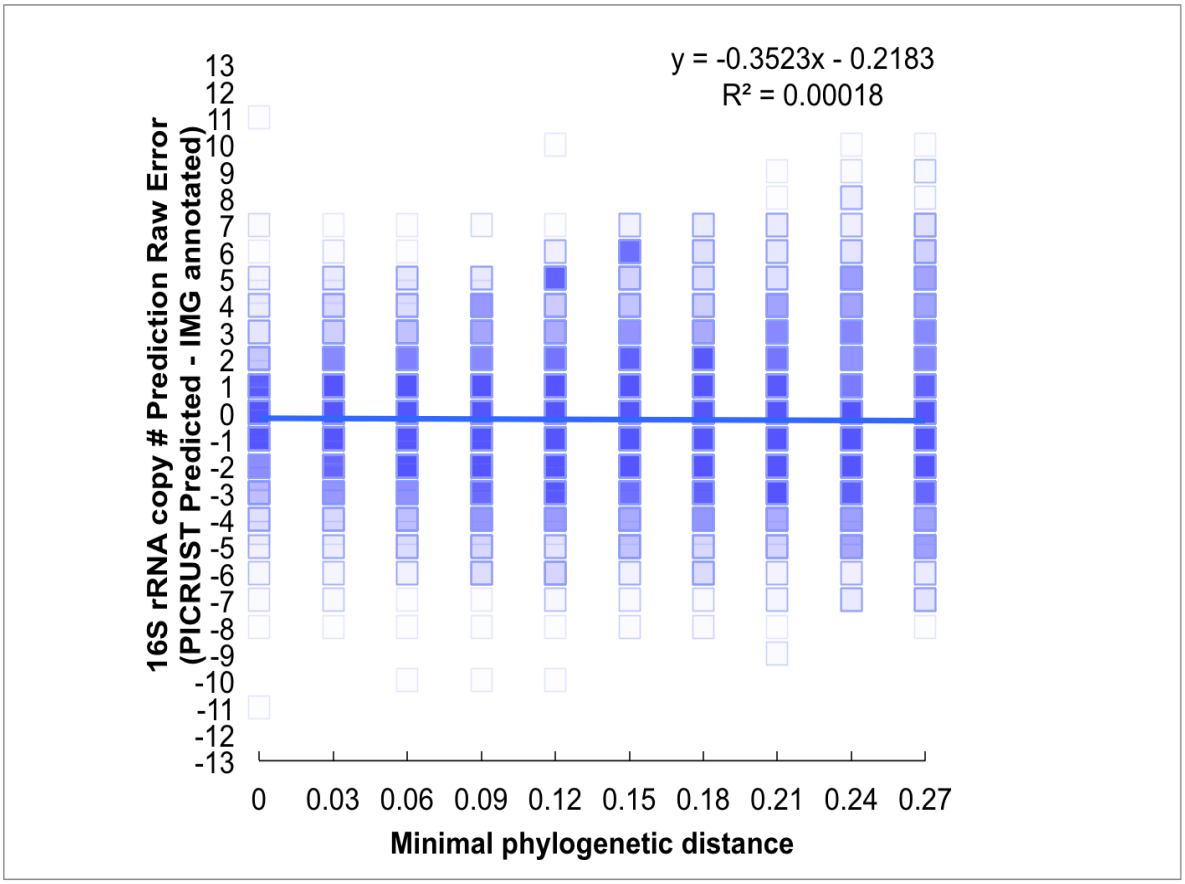
Supplemental Figure 14. PICRUSt accurately predicts 16S rRNA gene copy number for annotated organisms. The effect of sparse annotations on 16S rRNA copy number prediction accuracy was tested using cross-validation of annotated organisms. The chart compares the frequency of predicted vs. actual copy numbers. Each occurrence of a pair of actual vs. predicted values produces 1% saturation, so darker cells represent the most common annotations. Predicted and actual copy numbers were well correlated (Pearson $r^2 = 0.787$; $p<0.001$).



Supplemental Figure 15. Comparison of 16S copy number estimation accuracy for nearest neighbor or PICRUSt estimation with increasing distance to a known reference genome. The y-axis represents the correlation between actual 16S rRNA gene copy numbers and predicted copy numbers. The x-axis displays the distance over which annotated neighbors were removed (simulating sparse annotations of 16S copy number in that portion of the tree). Blue line: PICRUSt estimate using Wagner parsimony reconstructions and exponential weighting. Green line: Nearest Neighbor estimate. Red line: Random Neighbor estimate.



Supplemental Figure 16. Effects of availability of reference genomes on accuracy of 16S rRNA gene copy number prediction. The effect of sparse annotations on 16S rRNA gene copy number prediction accuracy was tested using cross-validation of annotated organisms. Test data excluded all annotated genomes within 0.03 (panel A), 0.06 (panel B), 0.09 (panel C) or 0.12 (panel D) 16S substitutions/site on the Greengenes tree. For each test, PICRUSt predictions for 16S copy number using Wagner parsimony with exponential weighting were compared against the actual IMG copy number annotations. The opacity of data points corresponds to the number of observations, with maximum opacity capped at >= 100 observations.



Supplemental Figure 17. Error in 16S rRNA gene copy number prediction is not biased upward or downward based on the distance to the nearest sequenced genome. Results reflect cross-validation of PICRUSt 16S rRNA gene copy number predictions using 16S rRNA annotations for all finished IMG bacterial and archaeal genomes. For each test, PICRUSt predictions for 16S copy number using Wagner parsimony with exponential weighting were compared against the actual IMG copy number annotations. The y-axis represents the difference between PICRUSt's predictions and actual copy number. The x-axis reflects the distance to the nearest reference genome. The opacity of data points corresponds to the number of observations, with maximum opacity capped at >= 100 observations. Because no trend was observed between distance to a reference genome and error in copy number prediction, PICRUSt estimates of 16S rRNA copy number do not appear to be systematically biased upwards or downwards by increasing distance to a reference genome (although distance to the nearest reference genome increases error).