Dear Dr. Jones,

Many thanks for your positive feedback on this manuscript, and for your time and suggestions for improving it during revision. We’ve looked closely at both your editorial summary and the comments from reviewers, and a response to all points is included here. We also edited the manuscript based on this input and include references to the modified text when possible. Major revisions include:

* The inclusion of predictions and biological conclusions from two new 16S-only datasets, a series of 34 vaginal microbiome time courses (993 total samples) from Gajer et al STM 2012 and 335 newly-published coral mucus microbiomes. We have also further highlighted our previously included novel analysis of ~6,000 16S-only HMP samples for which no gene function information was available prior to analysis in PICRUSt.
* A demonstration of PICRUSt’s use with gene and pathway catalogs other than KEGG Orthologs, Pathways, and Modules, specifically by providing downloadable tools for use with the COG catalog and a newly-introduced validation/demonstration of COG analysis in PICRUSt.
* Clarification of the fact that PICRUSt performs gene-level inferences, not pathway-level, although the resulting gene abundances are typically grouped post hoc into pathways or other gene sets amenable to biological interpretation.
* We’ve provided a more detailed description of how PICRUSt can distinguish high-confidence predictions (e.g. for “core” genome elements) vs. lower-confidence predictions (e.g. for “pan” genomes), through the use of confidence intervals (provided in user output). We also discuss in more detail how PICRUSt can be combined with metagenomic sequencing and other assays and analysis tools as part of overall experimental design for microbial communities.
* A number of additional gene and genome-level validations of the ancestral state reconstruction process and an extension of our comparison of PICRUSt inferences to metagenomic sequencing.
* Additional minor figure improvements, text revisions, and supplement additions throughout.

Many thanks again for your feedback, and we continue to be open to any suggestions you or the reviewers can provide to further finalize the manuscript.

**Editorial comments**

*The reviewers raise several concerns that must be addressed by revision of the text and not only in a rebuttal. Also, after discussing the reviewers comments and setting this paper in the context of other similar papers that we have published, we require that you show that PICRUST works for more than high-level KEGG analysis (that is sources other than KEGG)*

We’ve provided two clarifications of this point in the revision, specifically:

* PICRUSt’s main output is a gene-level inference, which we then summarize using pathway-level visualizations for examples like Figures 2 and 6. We’ve emphasized this by adding a heatmap of individual gene abdundances (KOs) for the HMP dataset (new Supplemental Fig. 1) and revised the following text:

“Any functional classification scheme can be used with PICRUSt; here, we demonstrate the use of the popular KEGG Orthology (KOs) and Clusters of Orthologs Groups (COGs).”

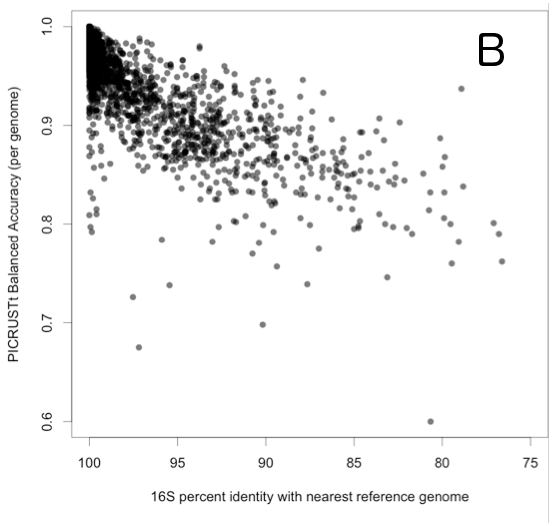
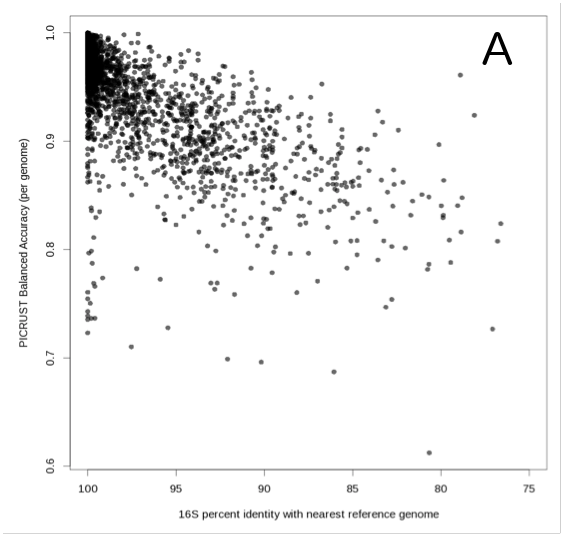
...

“The final output from the system is thus an annotated table of predicted gene-level counts (e.g. KO or COG ortholog groups) for each sample, which can optionally be collapsed into pathway-level categories...”

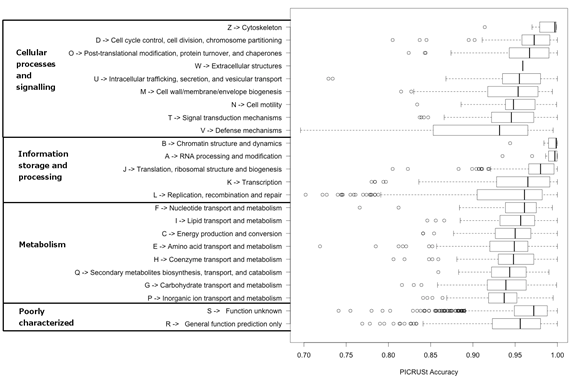
## 

**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).

* To demonstrate PICRUSt’s compatibility with gene catalogs other than the KEGG Orthology and pathway catalogs other than KEGG Modules and Pathways, we’ve added a complete dataset and analysis based on COG gene families. These appear in Supplemental Fig. 8B, which shows PICRUSt accuracies using individual COG families for each genome, and Supplemental Fig. 10, which groups these to assess accuracies of overall COG categories. These results support our previous findings using KO gene families and KEGG pathways.



**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).



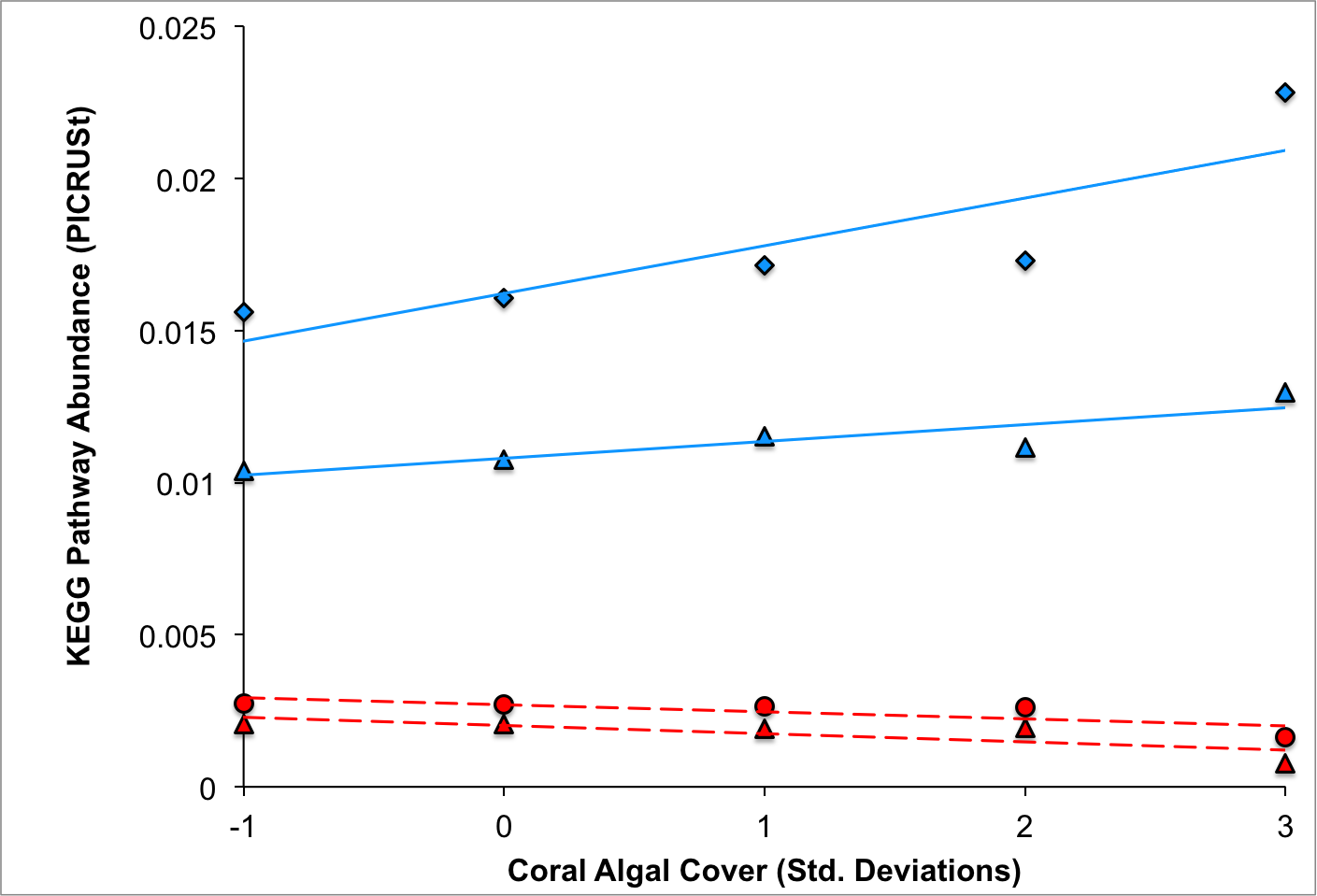
**Supplemental Figure 10. Accuracy of COG functions using genome holdout evaluation.** The ability of PICRUSt to predict the presence/absence of each individual COG orthologous family at the gene level was evaluated using genome holdouts (see Methods). Each COG family was then grouped into its corresponding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

*...and that use of PICRUST can lead to novel biological insights. The latter might be through ongoing projects, or, reanalysis of an available dataset to show more clearly compelling findings that could serve to illustrate the benefits of using your approach (that is to say, findings that had not previously been reported for that dataset).*

In order to demonstrate PICRUSt’s applicability to novel datasets and biological discovery, we have included two new sets of inferences. The first describes a previously unpublished dataset of 335 samples from coral surfaces with varying levels of algal cover (Supplemental Fig. 11). It has been hypothesized that coral decline is in part a result of algal overgrowth due to eutrophication. Algae secrete excess dissolved organic carbon, which in turn encourages overgrowth of fast-growing microbial opportunists. PICRUSt analysis of corals mucus in experimental plots varying in algal cover detect several trends consistent with this hypothesis: increases in both growth-related pathways and secretion systems with increasing algal cover, but decreases in multiple categories of carbohydrate metabolism (associated with coral mutualists that feed on e.g. galactose secreted by the coral). This is described in the newly-added text:

“In the second example, we applied PICRUSt to generate functional predictions for ecologically-critical microbial communities associated with reef-building corals. The system under study is subject to an experimental intervention simulating varying levels of eutrophication and overfishing. One hypothesis to explain the role of algae in global coral decline posits that eutrophication favors algal growth, which in turn increase dissolved organic carbon (DOC) loads. DOC favors overgrowth of fast-growing opportunist microbes on the coral surface, outcompeting more typical commensal microbes, depleting O2, and ultimately causing coral disease or death. This is known as the DDAM Model: Dissolved organic carbon, Disease, Algae, and Microbes; direct algal toxicity through secreted allelochemicals also appears to play a role. To shed light on this hypothesis using PICRUSt, we predicted metagenomes for 335 coral mucus samples collected in situ from corals in experimental plots with varying levels of algal cover (Supplemental Fig. 11). Consistent with algae-driven increases in opportunistic pathogen loads, secretion system genes were strongly correlated with relative algal cover (Spearman r=1.0, p=0.0), with 46% enrichment in corals from high- vs. low- algal cover plots. Algal cover also produced significant variation in ribosomal biogenesis genes (ANOVA raw p=1.6 x 10-4; Bonferroni-corrected: 0.049; FDR q= 0.0047), indicating an effect on generally faster-growing organisms. This variation was strongly correlated with relative algal cover across plots and timepoints (Spearman r=0.90, p=0.037) and represented a 25% increase in this gene category between corals in plots with the highest vs lowest algal cover. Further supporting evidence for a decrease in typical consumers of coral mucus carbohydrates in favor of fast-growing opportunists was provided by significant depletion of two categories of carbohydrate metabolism genes (Spearman r=-1.0; p=0.0 “Galactose metabolism”; Spearman r=-0.90, p=0.037 “Ascorbate and alderate metabolism”). As the weighted NSTI in this case was 0.12 (+/- 0.02 stdev.), these results suggest that PICRUSt may provide biologically actionable hypotheses even in challenging environments with fewer available reference genomes.”

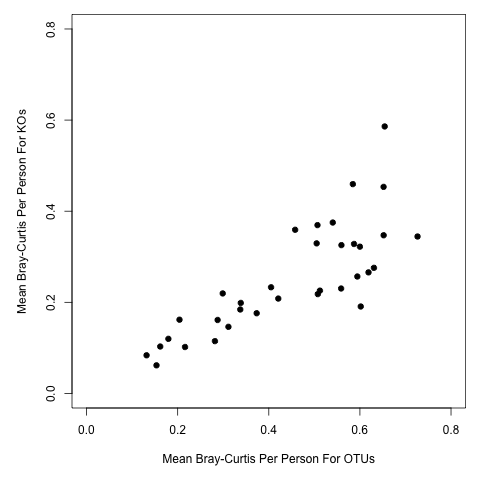
This is accompanied by the new Supplemental Figure 11:



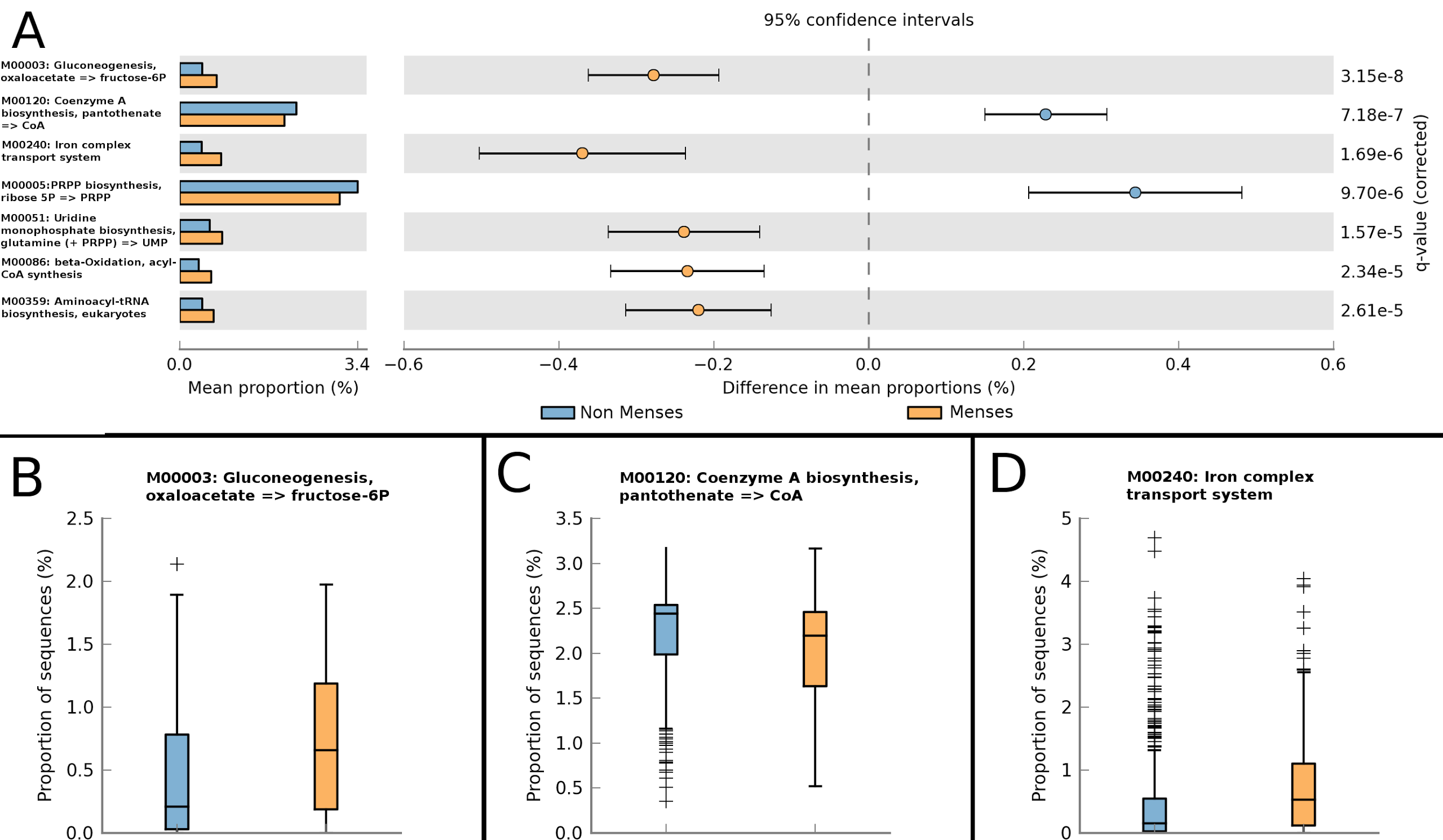
**Supplemental Figure 11. PICRUSt analysis of algal cover and predicted gene frequency in reef-building corals.** 16S rRNA gene amplicons from 335 coral mucus DNA samples from an ongoing in situ experimental intervention in the Florida Keys were sequenced with 454 pyrosequencing, analyzed in QIIME using standard workflows, and converted to predicted gene abundances using PICRUSt (weighted NSTI = 0.12 +/- 0.02 stdev.) 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; r2 > 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.

The second newly added novel application is to a published vaginal microbiome study that utilized only 16S sequencing. PICRUSt predictions for all 993 samples (from ~3mo time courses spanning 34 women) allowed us to quickly identify at least two previously unavailable biological results. Specifically, the first new observation was that the functional stability (relative to organismal variation) previously seen cross-sectionally among individuals is also maintained longitudinally within individuals (new Supplemental Fig. 12). The second novel point was the identification of KEGG metabolic modules with significant differences in mean abundance for samples taken during menses (new Supplemental Fig. 13), including genes involved in iron transport. This provides what may be the first culture-independent description of rapid vaginal microbial metabolic adaptation at the metagenomic level to the host menstrual cycle. These results have been summarized in the manuscript as follows:

“Finally, we assessed 993 samples from time courses covering approximately 16 weeks each from 34 individual subjects’ vaginal microbiomes. These samples have been previously analyzed only in the context of longitudinal changes in microbial taxonomic composition over time; PICRUSt provided the first insights into what additional putative microbial pathway changes might explain or accompany this compositional variation. The first observation this enabled was a comparison of community beta-diversity within subjects over time, contrasting the degree of similarity of microbial composition over time with the similarity of the accompanying inferred metagenomes. In all cases, the mean Bray-Curtis diversity using KOs predicted by PICRUSt was more stable over time than when using OTU composition (Supplemental Fig. 12). This provides the first longitudinal results mirroring the functional stability in metagenomes that has been observed cross-sectionally. Second, we identified 7 KEGG modules that had significant differences in mean abundances in samples taken during menses (Supplemental Fig. 13). The KEGG module with the largest significant increase in mean proportion during menses was “M00240: Iron complex transport system”, suggesting a shift in the microbiome that might be explained by pathways utilizing the iron-rich environment provided during menstruation.”



**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 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.

Finally, we also clarified the material regarding the inclusion of 6,431 newly-analyzed samples from the Human Microbiome Project that had not previously been functionally profiled. These were underemphasized in the previous draft, and we call additional attention to them in this revision with the following text:

“As a final illustration of PICRUSt’s computational efficiency and ability to generate new biological knowledge, we applied PICRUSt to three large 16S rRNA datasets to generate new functional insights. In the first example, all 6,431 16S samples from the HMP were used to predict metagenomes using PICRUSt, requiring <10 minutes of runtime on a standard desktop computer. One of the many potential applications of such data is in functionally explaining shifts in microbial phylogenetic distributions between distinct habitats. Previous culture-based studies had detected higher frequencies of aerobic bacteria in the supragingival plaque relative to subgingival plaque, and an analysis of HMP 16S rRNA sequences detected taxonomic differences between these two sites. Analysis of the PICRUSt-predicted HMP metagenomes revealed an enrichment in the metabolic citrate cycle (M00009) genes in supragingival plaque samples in comparison to subgingival plaque (p<1e-10; Welch’s t-test with Bonferroni), supporting previous claims that aerobic respiration is more prevalent in the supragingival regions.”

We can also provide additional information for the review process based on three ongoing collaborations that have used PICRUSt for additional datasets, each with their own manuscript outside the scope of this initial publication:

1) With Merete Eggesbo, we used PICRUSt in a cohort of 500 Norwegian babies to identify functional genes that vary with age, delivery mode, health outcomes (too large a dataset to be practical with shotgun sequencing).

2) With Kenny Simpson at Cornell, we paired metabolomes from 80 Crohn's, colitis, and healthy subjects with 16S PICRUSt inferences and obtained good correlations between microbial pathway abundances and products.

3) With Jon Braun at UCLA, we also paired metabolomes from 47 CD, UC, and healthy subjects at two biopsy locations with PICRUSt data, and again obtained good agreement between inferred pathway abundances and the metabolites they produce.

Because our current manuscript focuses on the PICRUSt methodology, its motivation, use, and especially validation, we feel that it would not be appropriate to include additional non-validated inferences directly. However, we hope that the new datasets, these examples, and the descriptive text below integrating PICRUSt with other functional assays are together enough to describe its role in biological discovery.

*Finally, you need to more clearly report in the paper how this type of approach compares with other methods to find function (to better profile the advantages) and suggest in the Discussion what types of projects could employ PICRUST to achieve goals that might not be feasible otherwise (larger sample sizes, confirmation of metabolomics or other data and so on).*

This is an excellent suggestion, and we have added the following Discussion text to guide users of PICRUSt in its strengths, weaknesses, and integration into broader microbial community analysis workflows:

“To best leverage the strengths both of (meta)genomic sequencing and of PICRUSt, we recommend its incorporation into marker gene studies using a deliberate, tiered approach. Because phylogenetic dissimilarity among environmental organisms and sequenced genomes (as captured by NSTI) affects PICRUSt accuracy, NSTI values can be calculated from preliminary 16S rRNA data to assess whether reference genome coverage is sufficiently dense to allow for accurate PICRUSt prediction. If adequate reference genomes are not available, additional genome sequences can be collected to fill in phylogenetic “gaps” in the reference database and allow for accurate prediction. This can be performed either through traditional culture-based techniques, single-cell genomic approaches, or deep metagenomic sequencing of samples targeted based on 16S data. If NSTI appears sufficient but additional controls are desired, a preliminary set of paired 16S rRNA and shotgun metagenomic samples can be compared using PICRUSt’s built-in tools to empirically test prediction accuracy on the sample types of interest. Based on such validations from select samples, PICRUSt can then be used to extend approximate functional information from a few costly metagenomes to much larger accompanying 16S rRNA gene sequence collections.”

*You should also explicitly discuss limitations of your approach (reviewer 1 makes excellent points in this regard).*

In addition to the discussion of PICRUSt’s capabilities above, we have more explicitly described its limitations in the Discussion section, discussing examples of poor annotation (acetogenesis genes) or variable distribution (methane oxidation genes) raised by Reviewers 1 and 3:

“However, the limitations of this approach must be considered in interpreting PICRUSt predictions. For example, only 16S marker gene sequences corresponding to bacterial and archaeal genomes are currently included; thus this version of the system does not infer viral or eukaryotic components of a metagenome. PICRUSt’s ability to detect patterns also depends on the input data used: the software cannot distinguish variation at the strain level if the marker gene sequence used is identical among strains, and cannot detect genes families (or summarize them into pathways) if those genes are not included in the input genomic data used, or if pathway annotations are currently poor (e.g. for acetogenesis genes). However, because PICRUSt can accept trees produced by alternative marker genes or gene/pathway annotations, users have the flexibility to customize the tool to meet the needs of their system. Although high overall accuracy was obtained despite microbial lateral gene transfer and other processes of gene gain and loss, gene families or pathways (e.g. methane oxidation) with highly variable distribution throughout the tree of life can still lead to incorrect predictions in individual cases. PICRUSt thus provides confidence intervals for each functional abundance prediction that reflect the degree of variation in that function among sequenced phylogenetic neighbors of predicted (meta)genomes, with wide confidence intervals indicating a high degree of uncertainty (Supplemental Fig. 7). If individual gene abundances (rather than aggregate patterns) are of interest, users can choose to either discard predictions with low confidence, or confirm them experimentally..”

**Reviewer #1**

*In this paper, the authors present a tool, PICRUSt, that can be used to predict functional potential of a community from taxonomic data, namely the 16S ribosomal RNA gene, suggesting a "predictive metagenomic" approach to analyzing microbial communities. The ability to make this taxonomic to functional link would address both cost and depth-of-sampling challenges in functional metagenomics. Additionally, this method allows for estimates of the contribution of each OTU to a given gene function, something that cannot be easily done with shotgun metagenomic data. The authors evaluate the performance of this tool across a few different biomes and many different datasets and introduce dataset metrics, such as the Nearest Sequenced Taxon Index (NSTI) that correlate well with accuracy. This paper is well written, the analyses are statistically rigorous and the software is freely available and well documented. This paper and the software it introduces, PICRUSt, could be an important tool to address some questions in functional microbial ecology without the need for shotgun metagenomes. The paper itself does not provide novel results, but does describe in detail the potential value of the developed software.*

*This is primarily a tools paper, and a very good one, so it does not necessarily need to include a novel insight, but addressing one particular question with this tool, that either could not have been answered before or was answered but at a much higher cost, might highlight good use cases for the software. As the software is described now, it is clear that using the information from a 16S to functional mapping, can determine if the overall functional potential of one community is different than another, in most cases. Figure 6 also provides a nice accuracy profile for different functions of interest, suggesting that certain processes might be better targets for comparison. Figure 2 also nicely highlights some processes that differ between environments. With this approach discoveries like this could be made, and given the existence of so many 16S datasets, these datasets could be re-evaluated with this tool to give a potential functional perspective.*

*However, this tool might not be well suited to particular questions, such as looking at processes in methane oxidation or denitrification that are broadly taxonomically distributed and where the process continues to be discovered in new taxonomic groups. Like any analysis, the approach a researcher takes should depend on the question they are asking. The authors do point out some of the limitations of the tool in the discussion, but it might be worth discussing at what level the analyses can be done. For instance, this is all done with KEGG groupings. How might it perform at a gene level, such as in MG-RAST? Or would this not be an appropriate use of the software? For instance, as far as I can tell, this tool would not be used to discover rhodopsin in the sea as DeLong did. It's not expected that one tool can do everything. It just might be useful to discuss it's specific purpose, or to have a use case that shows its utility.*

*The development of this tool also shows the importance of comprehensive reference databases. This is one of the best cases I've seen made for that. The authors do a comprehensive analysis of how the number of genomes correlates with accuracy. They might also use it as an opportunity in their discussion to call for more reference genomes.*

Many thanks to the referee for their kind words, thorough summary, and thoughtful feedback. We have addressed specific comments below, but to touch on several of these high-level areas that have become themes of this response:

* We agree regarding the balance between validation and new discovery needed to best describe PICRUSt in this manuscript, and we hope that the algal and vaginal datasets added above addresses this point.
* Likewise, PICRUSt certainly has limits to resolution and accuracy that would preclude it being the “only” assay used in a functional exploration of a microbial community. We have detailed this along with other anticipated limitations in the accompanying added text.
* We hope that we have clarified that PICRUSt does operate primarily at the gene (or more precisely gene-family) level, and in addition to expanding our discussion of KEGG Orthology genes grouped into pathways, we’ve added COG-based gene families and pathway groupings.
* Finally, we have also described PICRUSt’s place in broader computational and experimental workflows that might be used to explore microbial community function.

*Specific points*

*I think many microbial ecologist will be uncomfortable at first with the basic premise of this paper, that taxonomy can recapitulate function. We see so many cases where functions are very taxonomically diverse, where there's lateral gene transfer or where closely related organisms, even organisms that are the same species by 16S, have different functional capabilities. I think what the authors are trying to demonstrate, and do effectively show, is that broadly, taxonomy does have some functional predictive power. It just may be that many of the specific functions researchers are interested in, do not behave that way.*

This is an excellent point and represents a balance that we have aimed to achieve when describing PICRUSt: microbial function is neither randomly distributed across the tree of life, nor is it 100% predictable from phylogeny. Previous quantitative estimates suggest that the proportion of variation in gene content explained by phylogeny is somewhere between 70-90%, depending on factors like heterogeneity of environments, inclusion or exclusion of intracellular endosymbionts (which can be more problematic than lateral transfer), etc. It was these types of descriptive studies (Chaffron et al. 2010, Zaneveld et al. 2010, Konstantinidis and Tiedje, 2005, and Konstantinidis and Tiedje, 2005), along with the repeated observation of high correlation between metagenomic function and 16S phylogeny found in the literature (e.g. Muegge et al. 2011) that led us to believe that PICRUSt might be possible in the first place.

The reviewer’s point regarding functional variants within species is a particularly important one. In the gut, we have previously estimated that ~9% of gene content variation occurs below the 97% OTU thresholds (4-17% found using various 250bp tags). The variation in the estimate is due to differences in the rate of evolution between 16S regions and corresponding differences in which genomes are in estimated 97% OTUs) (Zaneveld et al. 2010). It is possible that more accurate ‘sub-species’ level placement may be possible with improved read lengths which would, in turn, improve predictions. Internally, we have discussed testing a number of alternative read-placement/OTU picking protocols upstream of PICRUSt, including pplacer. Such alternative protocols would integrate readily with our analysis step and will be updated through the web page if they prove useful. As read lengths from sequencing improve, fine-scale phylogenetic resolution for environmental organisms should be easier to achieve in high-throughput analyses. For now, if phylogenetic resolution is essential but a comprehensive exploration of alpha diversity is not, there is no reason that PICRUSt could not be run with Sanger reads and full-length 16S rRNA trees; our validations indicate that predictable functional content converges rapidly with number of 16S sequences, which suggests that sub-species predictions with hundreds or thousands of Sanger reads are feasible. If even finer resolution is required to distinguish fine-scale differentiation amongst strains 100% identical in the 16S rRNA gene, other fast-evolving markers can be sequenced and used straightforwardly with the system as an alternative.

An interesting future project, for example, might be artificially evolving divergent strains across conditions of interest. Sequencing could be followed by PICRUSt analysis to detect the extent to which condition-dependent genomic changes (due to gene gain, loss or transfer) differed from the ‘neutral’ prediction using phylogenies built using either very fast-evolving marker genes or the known experimental history of the populations.

*The introduction is much stronger than the discussion about making the claim that taxonomy is related to function, stating that "phylogeny and biomolecular function are strongly correlated in all divisions of life". However, they cite papers that are focused on linking 16S genes with housekeeping genes or are conducting surveys in more well-studied systems, like the human gut. The discussion is more muted on this point. I think the authors could take a bit more of the tone in the discussion to the introduction, so microbial ecologists understand that the authors understand the limitations even while creating the tool. It is likely to make a reader less defensive at the start of the reading, because I think microbial ecologists might be reluctant to embrace this taxonomy/function relationship at such a broad scale when they're first starting to read the paper.*

We have revised the wording of this section to indicate that the degree of correlation indeed varies by trait. We have also added a sentence to indicate different degrees of expected correlation, anticipating the analysis of accuracy versus functional module we later show in Figure 6:

“Although marker gene and shotgun sequencing strategies differ in the type of information produced, phylogeny and biomolecular function are strongly, if imperfectly, correlated. Phylogenetic trees based on 16S closely resemble clusters obtained based on shared gene content, and researchers often infer properties of uncultured organisms from cultured relatives. For example, the genome of a Bacteroides spp. might reasonably be inferred to contain many genes encoding glycoside hydrolase activity, based on the commonality of these activities in sequenced Bacteroides isolates. This association is in turn closely related to the pan- and core-genomes of each phylogenetic subtree, in that larger and more uniquely conserved core genomes result in more confident linkages of genes with clades. Conversely, a clade’s core genome consists of genes its genomes can be expected to carry with high probability. The degree of correlation between phylogeny and functional attributes depends on factors including the complexity of the trait, but the overall degree of correlation suggests that a predictive approach may be fruitful.”

*It was good that the authors included HMP synthetic community dataset to help validate their approach. In this way they can compare 'truth' with what they know about the synthetic community, the shotgun metagenomes and the results from PICRUSt.*

Many thanks, and we have aimed to include pairings in all of our validations that are similar in spirit. The HMP mock community is perhaps the next most controlled setting to a purely synthetic dataset, to which we’ve then added the complexity of genomes with known annotations compared with PICRUSt inferences, metagenomes with independently analyzed annotations, and finally the newly added algal and vaginal datasets lacking paired functional information but with biologically plausible PICRUSt results.

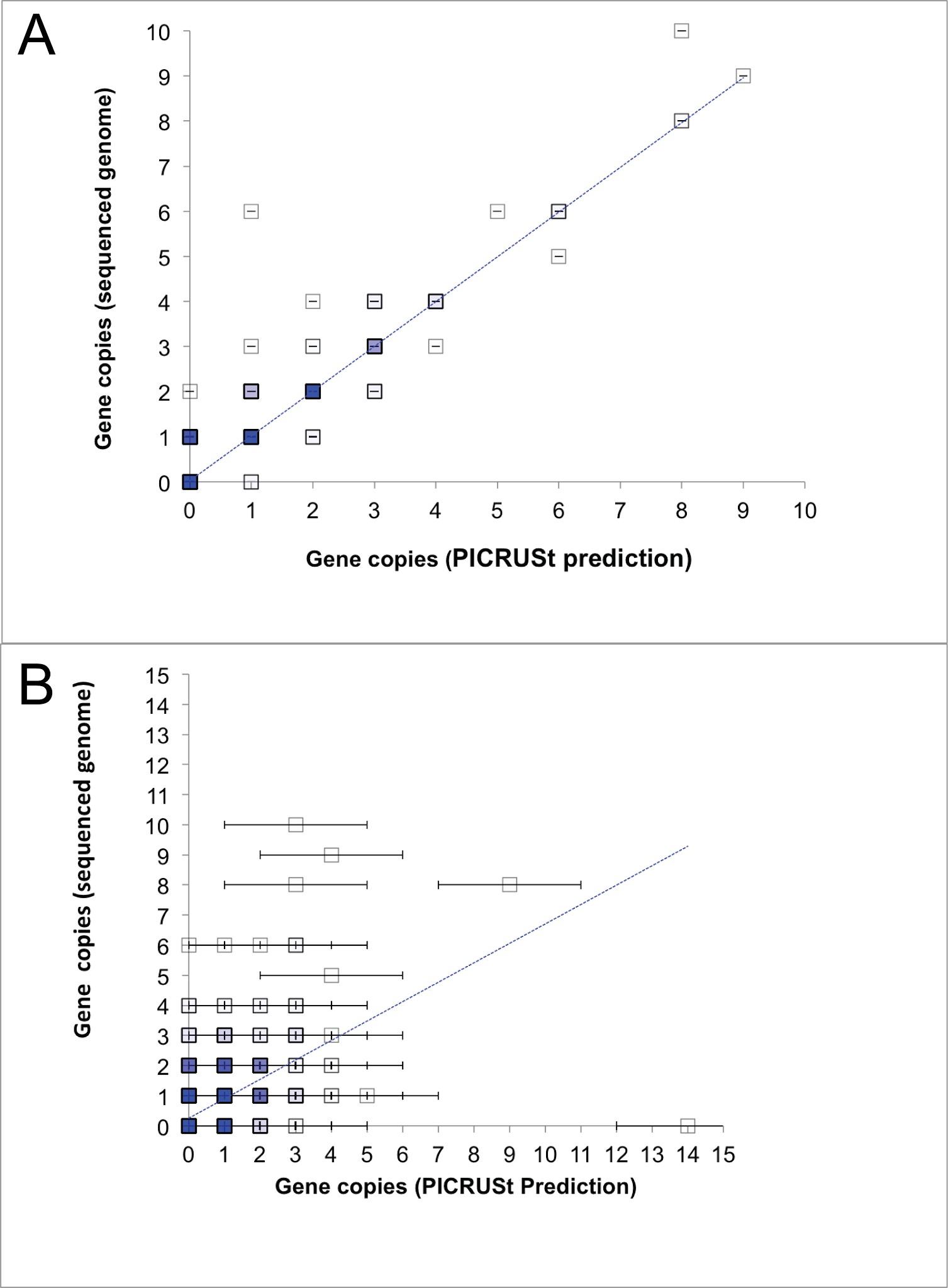
*It was important to consider the distance of the environmental genomes from the reference genomes. The NSTI that the authors describe is a good approach. Also, it was important to validate that the accuracy of PICTRUSt decreases with increasing NSTI, as expected. It's important with any tool to be able to know when it will be accurate and when it will not. While there is no specific NSTI cutoff below which researchers should not use the tool, Figure 3 does a good job of showing the range so a user can make their own determinations. Also, this metric is generally useful for assessing how novel a community is and could help guide what environments should be targeted for more sequencing of reference genomes.*

We thank the reviewer, and agree that providing accurate quality assurance information (both positive and, sometimes, negative) to users is essential. For the revision, we have included an update to this information as a result of feedback from other reviewers. Specifically, we have added a demonstration of PICRUSt’s built-in confidence intervals for per-genome prediction. These more formal confidence intervals supplement the NSTI score. We have added a figure demonstrating their performance under both favorable and very unfavorable availability of reference genomes.

Using the E.coli K12 MG1655 genome as our example, we performed PICRUSt predictions either using all other sequenced genomes in the database (very favorable genome coverage) or excluding all relatives within 0.20 16S rRNA substitutions/site on the greengenes phylogeny (equivalent to NSTI >= 0.20). We then characterized the relationship of PICRUSt predicted confidence intervals to the number of gene families that fell above or below the 95% confidence limit empirically. In the favorable case, all confidence intervals were below one gene copy, while in the unfavorable case many CIs extended to multiple gene copies. Although prediction accuracy fell dramatically when all reasonably related references were excluded (from Pearson r2 of 0.93 to 0.31), the proportion of genes within the confidence intervals stayed consistent. In either case, the CIs were slightly conservative: 97.7% of genes fell within the confidence intervals in the favorable prediction case, while 99.5% fell within the CI for the unfavorable case. Characterization of empirical CI values for the other two genomes shown in the tree randomization figure (including SAR11, which has a reduced genome) produced similar values (effective CIs ranged from 97.4 to 99.9). Separating calculation of upper and lower 95% confidence bounds for each of these genomes produced similar values, and in all cases empirical CIs for gene count data were >= PICRUSt predicted CIs (although lower bounds were more conservative than upper bounds because genes cannot be present in fewer than 0 copies). We therefore expect that these quantified confidence intervals will be useful for users interested in particular genes (e.g. high confidence genes associated with interesting biological patterns might be explored first, or low confidence genes of interest might be targeted for experimental confirmation).

We have added the following text to the manuscript (as well as references to the new supplemental figure where relevant):

“Ancestral state reconstruction 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 due to a variety of factors (see Discussion). Characterization under both favorable and unfavorable availability of reference genomes suggest that these confidence intervals are slightly conservative,but otherwise accurately capture uncertainty in per-gene predictions (Supplemental Fig. 7).”



**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 r2 = 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 r2 = 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). Separate testing of upper and lower confidence bounds found that in all cases the CIs were 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.

We also agree with the reviewer that the NSTI metric may well be a useful way of targeting reference genome sequencing. One direction for future development is addition of scripts that automatically select target reference genomes based on NSTI values for a sample collection. This would allow PICRUSt to be used as a tool for helping to fill in gaps in our understanding microbial genome contents in a systematic way, driven by 16S rRNA collections pertaining to a particular condition or environment of interest.

*It was also nice to see the result that PICRUSt could be more accurate in determining community functional potential than a shallowly sampled shotgun metagenome. It makes sense from a coverage perspective, so it's great to see that data. That conclusion might also help guide a researcher with limited resources to design their experiment differently, potentially even deciding to use only 16S and PICTRUSt, instead of trying to analyze undersampled shotgun metaenomes.*

We agree that undersampled metagenomes are an important consideration in experimental design; this is in some ways most critical for “deep” analyses like assembly, but we’ve shown here the extent to which it can also affect shallower functional profiling tasks. Ideally, PICRUSt will be used as part of larger workflows that integrate both large-scale computational prediction and smaller-scale, targeted follow-up for validation.

We have also updated the figure and text for this analysis to estimate the total number of reads (rather than the much lower number of annotated reads) as this was confusing to some readers.

*Software*

*I was very impressed with the software. First, it is key that the software has been released open source. Additionally, there is good documentation on installation and use of the software. That the development version is in git suggests good versioning programming practices. Reading the authors on the different scripts, multiple developers contributed to the code base, but seemed to conform to PEP 8 standards. There is reasonable code commenting. Additionally developing in Python makes the code easier to follow. It's great that they include scripts for parallelizing jobs. Once dependencies were installed, installation was straightforward. I didn't see anything about how the developers were continuing development. If researchers are starting to use it, I would hope that there was testing in place to ensure that future changes to the code base were not affecting output or results. I did not work with PICRUSt in Galaxy, but appreciate that the authors have already integrated it into that framework. That could make it easier for researchers to use.*

We very much appreciate the thorough software evaluation, and please feel free to provide additional feature requests or issue reports on Github once the review process allows. PICRUSt is automatically tested using a Jenkins suite, which provides nightly unit testing for exactly this reason. A set of additional script-level tests are in progress for inclusion in our next release.

*Also I didn't realize that it was 'pi crust' until I saw it written in all lower case. I was pronouncing it 'pic r u s t' before. ☺ How does your team pronounce it?*

A good question! We use “pie-crust,” and have updated the software web site to reflect this.

**Reviewer #2**

*This manuscript introduces a method of predicting abundances of different KEGG Orthology groups (KOs) in metagenomic data sample from the abundance of different 16S rRNA genes in the sample.*

*The main idea of the method is:*

*For each 16S rRNA gene in a sample, search reference genomes in the database with similar 16S rRNA gene sequence and predict the KOs on the reference genome by estimating abundances of different KOs in the sample. Although the methodology looks simple, it seems to work well on some metagenomic data, especially the human gut data which has many similar reference genomes in the database.*

Thanks to the reviewer for this excellent feedback and summary, and we hope that the current revision has additionally addressed the points below.

*Comments:*

*1) In this paper, experiments are performed on existing metagenomic data where the reference genomes or similar genomes of many high-abundance species are known. As the abundance of the KOs on these high-abundance species will be dominant in the sample, it is not surprised that the suggested algorithm performs well. The authors should perform experiments on some data (simulated data might be considered) with no reference genomes or similar genomes of the high-abundance species in the database so as to evaluate the performance of the algorithm on this new metagenomic dataset.*

We agree with the reviewer that availability of appropriate reference genomes for an environment is critical for successful application of the method, which is why we tested the accuracy of PICRUSt on several increasingly difficult metagenomes. The very diverse hypersaline mat communities represent an example where there exists very poor reference genome representation. PICRUSt provides an assessment of such coverage (and thus one measure of expected accuracy) using the weighted NSTI measure, which describes the phylogenetic distance for each OTU in a sample to the nearest reference genome. These are weighted by the abundance of the OTUs so that more dominant species have greater impact on the NSTI value (just as they do on prediction accuracy). Indeed, we see a decrease in PICRUSt accuracy for these high NSTI metagenomes (shown in Figure 3), although even in many challenging cases (e.g. diverse mammals) it does not decrease outside the range of biological utility.

Additionally, we did a more controlled and explicit experiment on the accuracy of predictions for each individual genome, using 16S rRNA copy number for our readout, by artificially removing reference genomes within various phylogenetic distances from individual genomes for which our preprocessing step predicts gene content (Supplemental Figs. 15-16). Accuracy remained fairly constant even when the few closest reference genomes were removed, but eventually declined as too many informative genomes were lost. PICRUSt can both quantify these declines for the user (in terms of per-dataset NSTI and per-prediction confidence intervals) and still provide quite accurate predictions for many types of well-sequenced microbial communities (e.g. human microbiome samples).

*2) Due to the limitation of existing sequencing techniques, the read length is much shorter than the 16S rRNA genes (from 100 bp to 400bp). Thus, experiments should be performed when only part of the 16S rRNA genes is known instead of the whole sequence of 16S rRNA genes.*

The metagenomes included in the manuscript for validation of PICRUSt contain reads that from several sequencing technologies, most of which are not full length. This included reads from Illumina, 454, and Sanger platforms, illustrating that PICRUSt can be applied to any of these sequencing technologies.

**Reviewer #3**

*Langille et al. present a novel software package (PICRUSt) that attempts to infer the "functional profile" (or gene family abundance) based on 16S rDNA sequencing. This approach successfully distinguishes different body habitats and has confirmed the abundance of some of the individual metabolic modules highlighted in the recently published Human Microbiome Project. The authors also highlight the limitations of this algorithm, in terms of the available reference genomes, the novelty of the habitat in question, and the particular metabolic pathway analyzed.*

*The paper is well written, with clear figures and many useful supplementary materials. The confirmation of recently published methods for inferring 16S copy number seems particularly useful, and well supported.*

*My major concerns are (1) the reliance on KEGG orthologous groups for validation and (2) the lack of a demonstration that this method might uncover novel biology.*

*Point 1: The KEGG annotation scheme is limited to well characterized core metabolic pathways and is undoubtedly missing many of the habitat-specific genes that encode key functions of interest. Relevant to the human gut, there are poor annotations of plant polysaccharide metabolism, acetogenesis, and membrane transporters, among others. Furthermore, a number of studies have demonstrated that microbiomes from a given body habitat have a high degree of correlation between KO abundances, suggesting that the interesting biology requires a finer level of analysis. Validations using individual genes or at the very least alternative annotation schemes (e.g. SEED, COG, GO) are necessary to confirm that this method will be useful. A more rigorous comparison would be to test if the predictions are significantly stronger than the correlation between different microbiomes, e.g. are the paired predicted/observed datasets nearest neighbors in a clustering analysis?*

The reviewer’s summary of the method and comments here have been very helpful, and we have specifically aimed to clarify 1) our use of KEGG gene families as compared to pathway-based gene groupings and 2) PICRUSt’s suitability for other gene/pathway catalogs, particularly a demonstration using COG.

First, although we present results summarized into pathways particularly in Figures 2 and 6, the method predicts the abundance of each gene family in each genome. We have clarified this in the added text:

“The final output from the system is thus an annotated table of predicted gene-level counts (e.g. KO or COG ortholog groups) for each sample, which can optionally be collapsed into pathway-level categories...”

The resulting gene-level predictions are what underlie reported accuracy and Spearman correlation values (e.g. in Figure 3). We’ve thus emphasized the gene-level data already reported using KEGG Orthology families, and expanded our visualization of these data by including a heatmap of HMP samples by individual KOs in the newly added Supplemental Fig. 1.

Second, since PICRUSt is generalizable to any annotated gene family identifiers and pathway-like gene groupings, we have included an added analysis of 4,975 COG gene families as inferred by holding out each individual genome from our complete catalog. We have shown very similar accuracy for COGs as we previously found using KOs. This is shown at the genome level in the newly added Supplemental Fig. 8B and across the different functional groups in Supplemental Fig. 10 (mean accuracy of 95% +/- 4%). We have additionally created and included with the PICRUSt distribution the COG precalculated files so that users can easily predict KOs and/or COGs for their data.

Although as the reviewer pointed out, that there will likely be habitat-specific genes not included in the KEGG or COG databases (an issue with any current protein or genome database), these systems are among the most popular schemes for annotating and describing both genomes and metagenomes. Providing these two formats ensures interoperability with a wide range of other catalogs (e.g. ECs, MetaCyc, etc.), and PICRUSt is architected to allow easy addition of other gene annotation schemes or pathway groupings.

Nonetheless, we felt the reviewer’s comment about poorly annotated pathways was important, and have added an additional note on this point to the discussion (see above), using the example of acetogenesis genes:

“PICRUSt’s ability to detect patterns also depends on the input data used: the software cannot distinguish variation at the strain level if the marker gene sequence used is identical among strains, and cannot detect genes families (or summarize them into pathways) if those genes are not included in the input genomic data used, or if pathway annotations are currently poor (e.g. for acetogenesis genes). However, because PICRUSt can accept trees produced by alternative marker genes or gene/pathway annotations, users have the flexibility to customize the tool to meet the needs of their system.”

*The paper could benefit from picking a few representative pathways of interest as test cases. For example, methanogenesis should be easily predictable based on its strong association with a single species, whereas acetogenesis would be difficult to predict based on its widespread nature among a variety of Firmicutes lineages.*

We agree that these pathways (along with many others) could be studied in much more depth, and many would likely benefit from their own entire separate manuscripts. In this work, we have focused on validating PICRUSt as a method, primarily quantitatively and with some qualitative biological examples focusing on specific functions. One such is the HMP guts, where we show that glycosaminosglycan degradation is found to be accurately predicted, along with the novel finding that the citrate cycle is more prevalent in the supragingival plaque (Fig. 2). Additionally at the genome level, we show that functions likely to be horizontally transferred such as membrane-bound proteins (Fig. 6) or those involved in defense mechanisms are not surprisingly not as accurately predicted (new Supplemental Fig.10).

*Point 2: The biological insights in Figure 2 remain at a very high level. It seems unsurprising that PICRUSt would be able to recapitulate the major differences between body habitats (Figure 2A). A more relevant validation would be to ask if the PICRUSt profile is significantly more similar to the same sample's WGS data than to a random sample from the same location. Figure 2A suggests that this is not the case - that although the general trends are found the algorithm has a consistent bias relative to the WGS data. Panels B-F are a nice validation, but it would have been really nice to see this algorithm applied to a new dataset (or existing 16S data) to demonstrate that novel biology can be revealed.*

We hope to have addressed this concern in several ways during revision, in part by noting that Figure 3 and Supplemental Figs. 2, 3, and 4 all assess PICRUSt accuracy by comparing inferences to a paired metagenome from the same originating sample, not a random sample from the same environment. To clarify Figure 2, we’ve further tested exactly this hypothesis, by assessing all HMP oral samples and comparing the accuracy (Spearman correlation coefficient) of PICRUSt to each sample compared with a random sample from the same body site. We found that PICRUSt predictions (mean r=0.950) were slightly but significantly greater with the sample of origin than with a random sample from the same site (mean r=0.928; Welch two-sample t-test p=2.57e-14). It is important to note that this small difference is to be expected since the difference between real metagenomic data for the same body site is so similar. It should also be noted that it is not surprising to see some clustering or “bias” in Figure 2A,since many methodological choices such as targeting different (informative) regions of the 16S rRNA also can produce small systematic shifts in PCoA plots, while still generally recapturing biological trends. In addition to our per-sample evaluations in the manuscript, this analysis shows that PICRUSt does have sample specific accuracy in addition to identifying the correct general signals from each body site or environment.

Second, we have further emphasized our novel analysis of the 6,431 16S-only HMP samples during this revision, including the finding that the citrate cycle is more abundant in supragingival plaque (not available from existing HMP metagenomes). We have moved these findings into a new section in the manuscript where they are highlighted in more detail. This section also includes the two entirely new analyses of the functional dynamics of the vaginal microbiome and of algal mucus microbes described above.

*Specific points:*

*"a human gut community rich in Bacteroides...inferred to contain many genes encoding glycoside hydrolase activity" - This seems like a bit of a stretch, given the widespread nature of these genes in many bacterial phyla (Firmicutes, Bacteroidetes, Verrucomicrobia). The authors might use a more specific case.*

We agree, and have adjusted this text to refer only to inference about particular genomes.

*"inference of metabolic functions is reliable" - The terminology is a bit loose in the paper. It should be clear that gene abundance is being inferred not metabolic function.*

We believe we have identified all occurrences of this language and corrected them accordingly.

*Figure 3 - what is the explanation for the strong accuracy for the soil samples despite a large distance to the nearest reference genome?*

Although NSTI does explain a great deal of the variation in PICRUSt accuracy across the different microbiomes we tested, there are other significant factors such as shotgun (Supplemental Fig. 3) and 16S (Supplemental Fig. 4) sequencing depth. For instance, the sequencing depth for the mammal metagenome was much lower (mean 55,000 seqs/sample) than that of the soil (mean 6,256,671 seqs/sample) or HMP (mean 49,913,014 seqs/sample) datasets (Fig. 2). As a result, the mean number of unique KOs (i.e. the number of KOs with abundance >1) from mammal metagenomes was very low (1,190), as compared to 3,463 unique KOs predicted by PICRUSt from paired 16S data. The degree to which shallow metagenomic sequencing undersamples gene content is apparent when considering that the average number of unique KOs within a reference genome is greater (e.g. 1,269) than the total detected in the described mammal metagenomic sequencing (e.g. 1,190), and that shotgun-based unique KOs for the HMP and soil datasets are 5,220 and 4,328, respectively. Overall, this would suggest that the shallowness of shotgun sequencing for the mammal metagenomes is responsible for inadequate capture of many of the functions within the true metagenome.

During the investigation of this explanation, we realized that we had perhaps previously under-emphasized the degree to which sequencing depth can influence metagenomic coverage (and thus both metagenome-based functional profile content and its comparison with PICRUSt inferences). We previously showed that subsampling metagenome reads will lead to PICRUSt outperforming shotgun sequencing at approximately 15,000 annotated reads. Considering the original paper describing the soil samples indicated that on average 17.3% of the reads could be annotated (Fierer et al, 2012), it would suggest that at least 88,235 raw reads (15,000/0.17) would be needed before metagenomic sequencing would do any better than using PICRUSt predictions based on adequately sampled 16S reads. Adjusting for success rates in closed-reference OTU picking for 16S rRNA data (to allow a fair comparison) produces a number of 72,650 total reads, below which PICRUSt outperforms shallow metagenomic sequencing (at least in the analyzed soil dataset). Although the mammal metagenomes do reflect different underlying community compositions, this strongly suggests that 55,000 sequences/sample is not enough to fully sample the true metagenome. This in turn causes PICRUSt to compare less favorably with the paired metagenomes since the shotgun sequence itself is an insufficient representation of the underlying community’s genomic content. We have updated Figure 4 to reflect this and added the following text to the manuscript to clarify that PICRUSt is likely performing better than is indicated due to shallow metagenomic sequencing:

“It should be noted that both the mammal and hypersaline metagenomes were shallowly sequenced at a depth expected to be insufficient to fully sample the underlying community’s genomic composition, thus likely causing PICRUSt accuracy to appear artificially lower for these communities (see below).”

*Figure 6 - The authors claim perfect accuracy for glycan metabolism. This seems hard to believe given the many limitations of KEGG for this set of enzymes - justifying the development of specialized annotation schemes (e.g.*[*www.CAZy.org*](http://www.cazy.org/)*) to parse these diverse gene families. Are the authors claiming that the abundance of GHs in general is accurate, or that they can accurately predict the specific GH families?*

This is another case where we hope to accurately balance descriptions of what PICRUSt can provide versus functional information too specific to be computationally inferred. We claim only that we recapture the same information on glycan metabolism as present in the KEGG annotations for the sequenced genomes, not that KEGG is the ideal resource to study glycan metabolism. This is thus indeed at a lower level of specificity than specialized annotations like CAZy, which we definitely agree is an excellent resource for detailed predictions of GH activity (we have also found it more useful/appropriate than KEGG for that particular task). Our pipeline is fully extensible to cover other annotation schemes, but we chose to focus our *precalculated* results on general annotations such as KO or COG families first. Extension of our precalculated files to CAZy is a potential direction for future development. Until then, interested users can perform such an analysis now if they compile a table of CAZy annotations across IMG genomes, precompute the resulting ancestral state reconstruction tree using our provided scripts, and use this (instead of the KO or COG table) to perform metagenome reconstructions.

*Figure S1 - These correlations look pretty poor by eye, with many KOs off the diagonal. In fact, they appear to be a weaker correlation than published data from natural microbial communities.*

Since most points agree so well in this scatter plot that they overlap, the Spearman correlation coefficients are actually quite high (r=0.9, p-value<0.0001). This value was previously included only in the manuscript itself, and we have now also included it in the figure legend.

*Figure S4 - The authors argue that ancestral state reconstruction is necessary, but this data seems to indicate that using the nearest genome (a far simpler method) is just as powerful.*

Although not obviously visible in the figure, the difference in accuracy between “ACE PIC” and “Nearest” is significant (mean ACE=0.955; mean “Nearest”=0.940; Wilcoxon rank sum test p< 2.2e-16). Although the absolute difference in accuracy appears modest, it actually implies a quite substantial reduction in the error rate, because both Nearest Neighbor or ASR-based methods perform quite well. Specifically, error rates of ~6% (Nearest Neighbor) are reduced to 4.5% (ACE PIC)- a 33.3% reduction. The other important reason we chose to emphasize an ASR method over simply using the “Nearest” approach is because it produces confidence intervals, which provide an important component of PICRUSt’s quality assurance output for users. Lastly, although ACE PIC was chosen as the default, all of the prediction methods tested in Supplemental Fig. 5 are included as current options in the software package (so users who prefer the simpler nearest-neighbor approach can certainly use it). PICRUSt was designed so that any ASR (or other) prediction method can be easily incorporated in the future. We added the following text to the caption:

“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 that is slightly more significant than just using the nearest approach (mean “ACE PIC”=0.955; mean “Nearest”=0.940; Wilcoxon rank sum test p< 2.2e-16). Note that all methods are available as options in the PICRUSt software.“

We have also update the results text to emphasize these points regarding error rates and confidence intervals:

“Ancestral state reconstruction(ASR) based methods (mean=0.955 +/- 0.04 s.d.) significantly (Wilcoxon rank sum test p< 2.2e-16) outperformed simply taking the functional content of the nearest neighbor among sequenced genomes (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- decreasing overall error rates by ~33%. Ancestral state reconstruction 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 due to a variety of factors (see Discussion). Characterization under both favorable and unfavorable availability of reference genomes suggest that these confidence intervals are slightly conservative, but otherwise accurately capture uncertainty in per-gene predictions (Supplemental Fig. 7). Given these results, the phylogenetically independent contrasts (PIC) method of Felsenstein[32](#h.32hioqz) 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[15.](#h.3j2qqm3)”

**Reviewer #4**

*This approach extrapolates from reference genomes the gene content of a metagenomic sample based on phylogenetic markers. While its utility is limited to cases where reference genomes exist, the authors indicate that their NSTI score allows the end-user to determine when the method is useful and when it is not.*

First, many thanks for the reviewer’s input, and we hope to begin our response with a clarification: because PICRUSt builds ancestral state reconstructions using the phylogeny and annotations of all sequenced genomes, *some* reference genomes always exist. The NSTI is designed to provide a quantitative assessment of how closely the available reference genomes apply to organisms in a sample of interest. We additionally include confidence intervals derived from the evolutionary model for each gene family to provide a more detailed picture of the variance in the prediction based on both reference genome coverage and the intrinsic rate of evolution for that gene class across all sequenced genomes. This approach stands in contrast to consideration only of reference genomes specifically for organisms present in the dataset (ignoring most OTUs), or the nearest-neighbor reference genome (ignoring the rest of the tree of life and confidence of ancestral reconstruction).

*Its major utility seems to be in analyzing metagenomic datasets that contain only 16S rRNA sequences, which they indicate "opens up new avenues for tiered, more cost-effective study designs and provides functional insights into the tens of thousands of existing samples for which only 16S data is available."*

*Major points:*

*• Since the early days of genome sequencing, it has been clear that strain-level variation can affect a huge proportion of the genome - such as a fourth to a third of differences between various E. coli - and that these variations are correlation with absolutely crucial metabolic processes such as the enabling of pathogenicity (often via integration of virulence factors into the host from phages and other mobile elements) vs. remaining benign.*

We are familiar with the and literature on pan-genome variation raised by the reviewer, and the observation from Welch *et al* that only 39.7% of protein families were shared among all of the first three strains of *E.coli* sequenced. In this context, it may be worth noting PICRUSt’s accuracy on these three strains. We calculated PICRUSt’s per-genome holdout prediction accuracy on *E.coli* strains K-12 MG1655, CFT073 and O157:H7 EDL933 by extracting these values specifically from our comprehensive single-genome holdout analysis. The balanced accuracy for prediction of gene presence/absence (identified as KEGG Orthology families) was 0.98, 0.97, and 0.92 for these three strains respectively. Thus even genomes in which there is a substantial history of gene transfer-driven strain-level variation can be predicted with great accuracy using phylogenetic methods, given enough genome sequencing.

Given Welch *et al*’s observations, our interpretation is that although much strain-level variation exists, the phylogenies inferred from 16S rRNA data are still relevant even amongst strains of the same species. For example, in Zaneveld *et al* (2010), we found that an additional 7-14% of genes were shared by genome pairs within phylogenetic distances of ~0.01 vs. ~0.03 16S rRNA substitutions/site on the phylogenetic tree. The idea that closely related genomes can help inform us about the function of microbes we find in the environment or in patient samples underlies most pathogen sequencing efforts, so none of these findings should perhaps be too surprising. Moreover for predictive purposes*, it doesn’t matter if some members of a given clade lack some gene of interest (e.g. the gene isn’t ‘core’), as long as the genomes most closely related to a sequence of interest have it.* Thus core genome concepts and functional prediction are related, but a gene does not need to be ‘core’ in order to be predicted. Given that even ‘challenging’ strains with reportedly rampant gene transfer can usually be accurately predicted from closely related neighbors (see above), the challenge shifts. Rather than asking whether, in general, gene content can be predicted from phylogeny, we should instead seek to quantify the relationship between variance (error) in the prediction of gene contents and the phylogenetic distance separating the strain of interest with available reference genomes. This line of reasoning led to our inclusion of NSTI and ASR-based confidence intervals in PICRUSt, and we have added an additional supplemental figure showing the per-gene confidence intervals resulting from “sufficient” or “insufficient” reference genome availability, using data held out from *E. coli* K12 MG1655 as an example (as discussed above).

It should also be noted that the enterobacteria have especially open pangenomes relative to other taxa, and that over the entire bacterial tree of life the correlations between 16S rRNA gene divergence and gene content are much better, as has been shown by multiple authors (see below). However, the information presented above to support the idea that PICRUSt works even in this worst case scenario for genome plasticity are, we hope, convincing.

*Regardless of the terminology used, I find it highly curious that this concept is almost completely absent from a paper that purports to identify the (entire?) gene content of an organism based only a single phylogenetic marker.*

Variation over evolutionary distance, both between strains and within broader clades, is a very central concept in PICRUSt. We agree with the reviewer that bacterial and archaeal genome plasticity, due to both gene transfer and genome reduction in intracellular endosymbionts, represents a crucial challenge in relating phylogeny to function. We understand that these challenges make it impossible to perfectly and deterministically predict entire genomes from 16S rRNA phylogeny and the set of all sequenced genomes - but we again wish to clarify that this is not the goal of PICRUSt. Instead, we seek to present the best possible summary of what the information that is available from all sequenced genomes tells us about the gene content in their unsequenced relatives (information that would otherwise be unavailable)- along with a quantification of uncertainty in that summary based both on the overall composition of a community (relative to available genomes) and on phylogeny-wide conservation of individual gene families.

Previous work by ourselves and many others provides a strong theoretical and empirical foundation for using maximum-likelihood methods to link phylogeny (derived from a single gene) and genome contents (for the entire genome) and to quantify the extent to which the former predicts the latter:

1) Gene content and 16S rRNA distance are strongly correlated (Goris *et al.*, 2007), even at extremely short phylogenetic distances and even when estimated using short markers from high-throughput sequencing (Zaneveld *et al.* 2010). The same connection has been described for gene order and 16S rRNA distance (Tamames *et al.,* 2001), KEGG pathway coverage and 16S rRNA distance (Chaffron *et a*l 2010), and the average nucleotide identity of gene pairs and 16S rRNA distance (Goris *et al*., 2007).

2) As we note in the text, whole-genome phylogenies largely mirror the results of 16S rRNA phylogenies (Segata 2011) further reinforcing that, while imperfect, the connection between phylogeny and function is more than strong enough to provide a great deal of information about genome content.

3) Even horizontally transferred genes tend to agree with central trends in organismal evolution described by ‘core’ genes. Specifically, when conserved gene transfer events are used as a phylogenetic marker the phylogenies tend on average to agree with putatively non-transferred core gene phylogenies (Huang and Gogarten, 2006).

4) Accounting for bacterial phylogeny improves the accuracy with which it is possible to predict the size of pan and core genomes from limited examples. These models outperform previous models for core genomes that do not incorporate the phylogeny and thus implicitly assume a star phylogeny (Collins and Higgs 2012).

Given these multiple lines of evidence that bacterial and archaeal gene content correlate well with phylogeny, we validated PICRUSt in such as a way as to quantify the degree to which these observations are biologically actionable. For one, we eschewed simulated datasets (which might suffer from incorrect models) and instead tested our computational prediction directly against complex real-world datasets relevant to practical analyses. Our control analyses were designed to determine under what circumstances the algorithm (and underlying evolutionary model) could accurately predict observed data not used in building the model: sequenced whole genomes and shotgun metagenomes. That is, we assessed the accuracy of the inference process relative to a non-computational, experimental method measuring roughly the same thing (whole-community gene content).

The purely phylogenetic model currently employed by PICRUSt thus provides much information that is of practical utility. Its open-source implementation further allows alternative evolutionary models to be directly compared against the current phylogenetic model. For example, one interesting direction for future development would be to test evolutionary models that explicitly incorporate biased within-taxon and within-environment HGT into PICRUSt’s estimates. Such a project would test whether these more intricate models do make a substantial difference in prediction accuracy, using the publicly available scripts for accuracy evaluation included here.

We hope that this clarifies our goals for PICRUSt, and that the method accurately quantifies the balance between gene families that are reliably phylogenetically distributed (and are thus accurately predicted) and those that are diffuse or frequently transferred (and thus are not included in predictions or are specifically annotated as low-confidence).

*The closest treatment of this issue is the "distance holdout" analysis in Supplementary Figure 9, whose mention is buried in the Methods section of the paper.*

This is exactly correct, in that both the single-genome holdout (in the main text) and the ‘distance holdout’ were in part inspired by the well-known divergent gene content results among *E.coli* strains mentioned above. Again, useful functional predictions (R2>0.9 in all cases) were made for the mentioned *E.coli* strains, and useful functional predictions could be provided even in a more challenging examples, such as when the nearest reference genome used for 16S rRNA copy number prediction was not in the same genus (Fig 10.; R2 =0.65).

*I believe that the issue of "core" vs. "pan-genome" needs to be presented much more prominently though, since it crucially affects the understanding of much of the manuscript.*

We agree that notions of “core” and “pan-genome” are useful models that may help to shed some light on the results. Indeed, there seems to be some convergence between recent advances in modelling core- and pan-genomes and our approach. Early core-genome concepts were operationalized by simply taking the intersection of all genes present in sequenced genomes in a species. This approach, although easy to calculate, implicitly assume an unstructured, star phylogeny amongst strains. More recent approaches instead employ explicit evolutionary models that incorporate phylogeny into their results. Empirically, these phylogenetic approaches appear to outperform previous non-phylogenetic approaches (e.g. Collins and Higgs 2012). These phylogenetically-aware approaches actually share some important features with PICRUSt, e.g. the quantification of uncertainty of gene presence/absence throughout the phylogeny based on branch length to the nearest observed genomes. We have modified the manuscript to note this important relationship between these two predictive problems:

“This association is in turn closely related to the pan- and core-genomes of each phylogenetic subtree, in that larger and more uniquely conserved core genomes result in more confident linkages of genes with clades. Conversely, a clade’s core genome consists of genes its genomes can be expected to carry with high probability.”

*For example, the claim that "phylogeny provides information about function", which of course applies only to the "core" part of the genome that is shared by all the strains of that organism, and only probabilistically to the "pan-genome" of an organism (but in ways that are unknown a priori, especially when there are only a few sequenced representatives).*

As the reviewer notes, pan-genome contents *are* statistically linked (albeit imperfectly) to phylogeny. However, imperfect does not mean unknown, in that the degree of linkage and uncertainty can be well-quantified from available genomes (and is during PICRUSt’s ancestral state reconstruction step). Therefore, phylogeny provides statistical information about both core and pan-genome contents. Our genome holdouts across the bacteria and archaea, for example, show that this is the case, as does the generally accurate prediction of metagenomic functional categories that are not part of any universal core. It is worth noting that ‘pan genome’ and ‘core genome’ are both terms that have meaning only relative to a particular clade: that is, the core genome of a species will contain genes that are in the pan genome of the genus.

Our approach intentionally quantifies this uncertainty, which is of course often small, as genes are distributed among genomes far from randomly. Specifically, ancestral state reconstruction estimates rates of evolution for each KO group (trait) across the entire bacterial and archaeal tree as captured by the full set of IMG genomes. This allows our confidence in any one trait’s presence or absence to draw on a larger pool of representatives than those present in a given species or genus. That is, if in many cases the frequency of horizontal gene transfer is more closely associated with the gene itself than with a particular “host” clade, then we expect copy number predictions for that gene family to have wider error bars than might be expected given host phylogeny. This can arise due to low connectedness in cellular complexes and/or pathways (Light *et al* 2005), incorporation in a mobile element (e.g. Zaneveld et al 2008; Leplae *et al.*, 2010), or because the gene itself facilitates mobility. Empirically, our summary of accuracy across functional categories argues that this is a productive approach to estimating error.

We have included a few additional results illustrating these confidence intervals for the *E. coli* K12 MG1655, using either the rest of the sequenced genomes, or only those at distances >0.20 16S rRNA subst. / site (NSTI >=20). As expected, the latter case produces wider error bars than the former. But regardless of the test data used PICRUSt’s confidence intervals are if anything somewhat conservative.

*Worse than that even, if there is only a single sequenced representative of an organism, it is impossible to distinguish between the "core" vs. "pan-genome", and this method would erroneously assume that all of the genes in that single sequenced representative are present in all of the strains of that organism (which would most decidedly NOT be the case - see for instance Rasko et al. J. Bact. 2008 and Perna et al. 2001 Nature for E. coli).*

If there is only one sequenced representative of a taxon, then the *point estimate* for the ancestral states (and therefore other organisms in that taxon)will be dominated by that organism (depending on the relationship in the tree to other sequenced taxa). However, the *confidence* in that estimate will be reduced based on the branch length separating that single genome from estimates tips in the tree. If the gene is relatively conserved in copy number over sequenced genomes across the tree (i.e. the Brownian motion Sigma parameter is low), then this effect will be small (e.g. for ribosomal proteins). However, for frequently transferred ‘auxiliary’ genes (some virulence factors, phage genes, transposases, etc.) the gene family will vary frequently across the tree, the Sigma value will be high, and therefore a very closely related representative may be needed for prediction. In some cases, the 95% confidence intervals may be wide even at 100% 16S identities, suggesting a need for fast-evolving markers to accurately predict *that particular gene family.* Thus the reconstruction method accounts for variable rates of evolution in gene copy number (including by HGT) across gene families, while also providing useful point predictions of gene copy number abundance.

*This concept can severely erode the utility of the method in some perhaps few but important respects (it cannot for instance distinguish between pathogenic and benign E. coli), especially when the authors seem unaware that it even exists (in the Abstract, Introduction, and in all but a single sentence in the Results and Discussion sections).*

We agree that fine-scale differences in clades with frequent HGT may require more genome coverage and/or phylogenetic coverage to allow high-confidence prediction, and we again A) alert the user to this issue quantitatively through confidence interval output and B) observe it quite infrequently in practice based on our whole-(meta)genome evaluations. As noted above, gene copy numbers for pathogenic vs. benign *E. coli* are not outliers in PICRUSt prediction accuracy. In fact, they are predicted with somewhat greater than average accuracy due to the dense sampling of reference genomes for this group. We do not directly address prediction of pathogenesis with PICRUSt, although the relationship between gene content and pathogenesis would be a fascinating, if challenging, subject for future study. Since identification of virulence factors and prediction of pathogenesis is not part of PICRUSt and space is limited, we do not discuss these topics in the main text.

*• The above point is just one example of the manuscript's inconsistency in addressing the limitations of the method. Some topics such as LGT are covered briefly, while others such as the "core" vs. "pan-genome" are present but highly buried, and in point of fact, most of the manuscript is dedicated to addressing the various limitations in the method's performance due to some factor or another. However, the abstract and other summary portions of the manuscript mostly ignore these issues and read as if they did not exist.*

*For instance the sentence in the abstract that reads: "PICRUSt...accurately predicts the abundance of gene families in host-associated and environmental communities." Obviously this information can only be predicted accurately when it is already completely known.*

Since this manuscript introduces PICRUSt as a methodology, we have endeavored to provide results and evaluations of the technique that can be validated. As with any machine-learning problem, this requires labeled test data (i.e. test data where the true result is known), which in our case takes the form of paired 16S and metagenomic datasets derived from the same underlying samples. These validations thus occupy the bulk of the manuscript in order to demonstrate that the method works in a variety of useful conditions, and also to identify circumstances under which it works less reliably. PICRUSt’s ability to generalize to new 16S datasets is hopefully not in question, given the basic premise of the method and our demonstration applications to >6,000 (unpaired) HMP 16S rRNA gene datasets, correlations between iron transport and menstruation in vaginal time course data, and the association of carbohydrate utilization and the degree of competition from algae in a coral microbiome.

As discussed above, we are of course aware of the impossibility of perfectly predicting function from phylogenetic information alone, especially without adequate reference genome collections. However, evaluations of paired metagenomes and cross-validation results for per-genome prediction strongly contradict the reviewer’s more general claim of inaccuracy. When a given genome is excluded from the analysis, its gene content can be predicted with 95% average accuracy (see Results and Fig 5). Similarly, when only 16S rRNA information is used, entire collections of shotgun metagenomes can be estimated, and key trends recaptured, both on a per-gene family or holistic basis (see e.g. Figures 2, 3, and 4). A more accurate statement would be that we can only perform a *quantitative* *evaluation* of predictionaccuracy when the answer is already known via a sequenced genome or metagenome, which is again the focus of much of this manuscript.

*Normally however, many or even most of the bacteria in a given sample may be relatively unknown, and while NSTI very usefully allows identification of the cases where this approach will work, no mention of that fact is given in the abstract, and in fact NSTI is not mentioned until page 5 in the manuscript.*

*Also, as mentioned above, at best it seems that PICRUSt can only hope to predict, accurately or otherwise, the "core" genome of an organism, and can make no claims at all about the "pan-genome".*

As discussed above, we should clarify that we by no means predict only the core genome for an organism; indeed, a “core genome” can only be defined in the context of a specific clade, at an arbitrarily chosen phylogenetic depth. PICRUSt will readily predict genes in a specific genome that would not be found in the core genome for e.g. that species, if the gene is present in closely related strains but absent in distant ones. Thus ‘pan genome’ as well as ‘core genome’ genes can be predicted (although core genes will of course be predicted with greater confidence). Given the correspondence between all previous observations of OTU identifiers and the contents of associated sequenced genomes, we infer with known confidence A) the potential contents of unobserved genomes and B) the resulting weighted combination of genome contents making up a community metagenome.

We thus predict the entire set of genes (represented as orthologous clusters) associated with each 16S rRNA sequence. At the same time, we quantify the uncertainty for each gene family, and fully expect that less conserved genes may have greater variance in presence/absence (and therefore require more closely related reference genomes) in order to predict accurately. We have added the following text to clarify this methodological point for other readers:

“Each of these factors can be modeled by ancestral state reconstruction, allowing environmental strains’ genome contents to be inferred with known uncertainty based on each gene family’s rate of change (encompassing propensity for lateral transfer, degree of “core” conservation within the associated clade, and distribution throughout all currently sequenced genomes).”

*To properly understand this method then, it would be nice if these issues were mentioned up-front, in the abstract and other summary portions of the text. For example, after "...accurately predicts the abundance of gene families in host-associated and environmental communities", the following could be added: "[...communities], when sufficient information from reference genomes is available." Or perhaps even "[...communities], when sufficient information from reference genomes (at close enough phylogenetic distance) is available."*

We completely agree with this suggestion, and have amended the text as suggested:

“Using 16S information, PICRUSt recaptures key findings from the Human Microbiome Project and accurately predicts the abundance of gene families in host-associated and environmental communities, with uncertainty quantified based on the availability of reference genomes at appropriate phylogenetic distances.”

*• The reasoning behind making predictions for tips in the reference tree that do not have corresponding genomes would benefit from being explained better. For instance, these are only partial predictions, and not even the full "core" genomes can be predicted for these, since any genes gained after divergence from the ancestral organism will not be known, but this issue and its impact are not mentioned, nor is it explained what the utility is of having these partial predictions, as compared to simply removing them.*

Generating gene content predictions for OTUs without associated genome sequences is important because the alternative strategy of considering only OTUs that map directly to sequenced genomes would ignore abundance values for many OTUs (except perhaps in extremely simple communities with heavy genome coverage). As a simple thought-experiment, consider a community in which all OTUs are related at roughly the genus level (~94% 16S rRNA identity) to known genomes. Such a community would have a fairly low NSTI value (0.06), but the direct mapping strategy mentioned by the reviewer would fail to predict anything about the metagenome at all (since no OTUs directly map to species). In contrast, PICRUSt can provide much useful information for real communities that are very similar to this hypothetical example, both in terms of the genomes of each member (see the holdout vs. distance analysis), and the complete metagenome (see the plot of metagenome accuracy vs. NSTI). Additionally, simple mapping of reference genomes (like nearest neighbor approaches) cannot provide a confidence interval for copy number predictions. Therefore we do not use this approach.

The reviewer suggests that PICRUSt predicts only a subset of ‘core’ genes. This is not the case, and predicting only core genes (at the species or genus level, for example) would miss a large proportion of genes (although exactly how many depends on which of various published core genome definitions was used). For example, Lefebure and Stanhope, (2007) report that only 11% of genes in genomes from *Streptococcus* species were present in 26/26 genomes analyzed. Using only e.g. genus-level core genes would therefore likely reduce accuracy substantially by introducing many false negatives, and we therefore do not use this approach.

However, the reviewer is quite correct that genomic changes on the branch separating a tip from its last common ancestor with a sequenced genome are not captured by PICRUSt. We address this uncertainty in PICRUSt estimates using confidence intervals from the model, and NSTI scores. We agree that this could be described in more detail, and we’ve expanded this explanation with the following Methods text:

“It is important to note that the prediction of gene content for tips in the trees without reference genomes is an estimate only, and that while our method does model gene gain and loss, some instances of gain/loss or laterally transferred genes will be poorly predicted (with broad confidence intervals as a result). This is rare in practice, however, as validated at the genome and metagenome level by comparing our predictions with the known gene contents from actual sequencing (see below).”

*• Why does the plot in Figure 3 show a mostly horizontal as opposed to diagonal trend for certain environments? For instance while Hypersaline shows a downward diagonal indicating that as 16S distance to the nearest reference genome increases, PICRUSt accuracy decreases, Soil and Mammal samples for the most part show that increasing 16S distance does not lead to a corresponding decrease in PICRUSt accuracy.*

Although the correlation between accuracy and NSTI is the strongest for the hypersaline samples, we do see significant negative correlations for each of the samples when analyzed by themselves:

Hypersaline (Spearman r=-0.818, p=0.007)

Soil (Spearman r=-0.556,p=0.042)

Mammal (Spearman r=-0.471, p=0.003)

HMP (Spearman r=-0.255, p= 2.931e-9)

All Samples (as reported in manuscript) (Spearman r=-0.409, p=2.2e-16)

This behavior is of course also dependent on sequencing depth as noted above. We have amended the previous result to include this extra information:

“Also as expected, the accuracy of PICRUSt in general decreased with increasing NSTI across all samples (Spearman r=-0.4, p< 0.001) and also within each microbiome type (Spearman r=-0.25 to -0.82, p<0.05).”

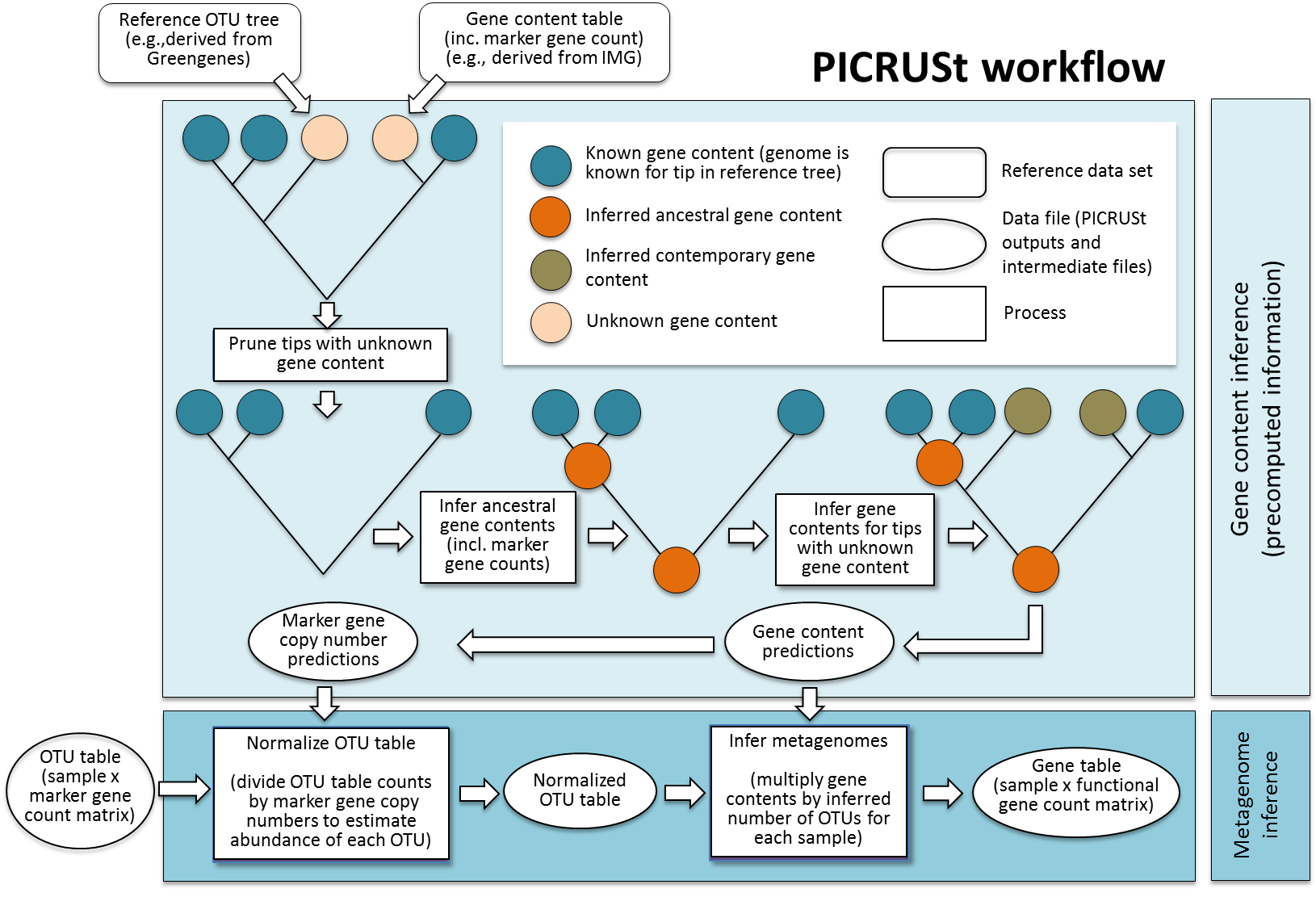
*Minor points:*

*• In the abstract, you can simply say "reference genomes" instead of "phylogenetically referenced genomes".*

We have made this change as suggested.

*• Figure 1 is very crowded, busy, and somewhat confusing, as the plethora of colors and shapes makes it difficult to read. Aside from the legend apparently being a "reference dataset" itself (it is contained within a rounded rectangle shape), it also seems unnecessarily large, with at least the right half possibly unnecessary.*

We thank the reviewer for feedback on this figure. We have modified the layout of some text in the legend, modified the shape and outline of the legend box for clarity, and removed non-essential colors. We feel that everything remaining in this figure is essential, but if there are specific reasons for additional edits we are happy to oblige. The figure caption is unchanged:



“**Figure 1: The PICRUSt workflow.** PICRUSt is composed of two high-level workflows: gene content inference (top box) and metagenome inference (bottom box). Beginning with a reference OTU tree and a gene content table (i.e., counts of genes for reference OTUs with known gene content), the gene content inference workflow predicts gene content for each OTU with unknown gene content, including predictions of marker gene copy number. This information is precomputed for 16S based on Greengenes and IMG, but all functionality is accessible in PICRUSt for use with other marker genes and reference genomes. The metagenome inference workflow takes an OTU table (i.e., counts of OTUs on a per sample basis), where OTU identifiers correspond to tips in the reference OTU tree, as well as the copy number of the marker gene in each OTU and the gene content of each OTU (as generated by the gene content inference workflow) and outputs a metagenome table (i.e., counts of functional genes on a per-sample basis).”

*• Forgive me if I am missing something in the last paragraph of the Results section, but isn't the fact that PICRUSt finds housekeeping genes rather obvious? Something would be wrong indeed if it could not find ribosomal genes or proteins present in all known bacteria! (which brings up the issue of why accuracy is only 99% instead of 100%? And sometimes as low as 90% it seems? Even ribosome functions can be missed, as seen in Figure 6) Instead of presenting this as a "result" found by your method, it may be helpful to re-word that sentence instead to present it as a positive control that shows that your method works in a scenario when it is expected to work.*

We apologize for the confusion, as this was indeed the intended presentation of the result - as above, it is typical for a new predictive methodology to be validated using labeled examples and positive controls. We have specified this in text:

“Analysis of PICRUSt predictions across functional groups (Fig. 6) revealed that as a positive control, core or “housekeeping” functions... even these more challenging functional groups were well predicted by PICRUSt (min. accuracy=0.82), suggesting that our inference of gene abundance across various types of functions is reliable.”

*• The authors mention that there are tens of thousands of existing samples for which only 16S data is available. This statement would be far more meaningful though if it could be extended to include what biological insights the authors expect could potentially be found among these samples, that could not or would not be found from full-gene metagenomics samples? (even one example would be better than none)*

The initial manuscript contained 7,207 total samples permitting biological conclusions drawn from 16S datasets, and we have supplemented these in several ways during the revision. First, we have introduced the new coral and vaginal datasets, inferences, and analyses described above. Second, we have listed above at least three additional ongoing studies in which PICRUSt is being applied alongside other ‘omics techniques in order to understand microbial community function in vivo. Third, we have included the new Discussion text also quoted above indicating possible uses for PICRUSt within functional microbial community experimental designs. We trust that these examples demonstrate that there is a good expectation of addressing and generating biological hypotheses using the method.

The metagenome inference process opens up any existing or newly assayed community amenable to 16S to extremely cost effective metagenomic profiling, as indicated in part by our metagenomic rarefaction analysis.While *sufficiently deep* and *sufficiently broad* shotgun metagenomic analysis provides the same information, it takes several orders of magnitude more *shotgun* sequence to attain a reasonably accurate metagenome profile than it does *marker gene* sequences, due to the additional prior knowledge incorporated by PICRUSt during the ancestral state reconstruction step over all 2,884 available microbial genomes.

Note that PICRUSt predicts metagenomes for all 6,431 HMP samples at no additional expense beyond the initial 16S rRNA sequencing. Our rarefactions demonstrate that a comparable metagenomic sequencing effort would have required each sample to be metagenomically sequenced to >72,000 total sequences (~15,000 annotated sequences) per sample. Since we did not previously clearly distinguish between raw and annotated reads (in either 16S rRNA or metagenomic sequencing), we have also annotated this in Fig. 4; since MG-RAST uses pre-annotation read numbers, this also increases the number of existing studies that would benefit from PICRUSt to ~16%. For the HMP, however, as a rough estimate, Illumina shotgun metagenomic sequencing at this (extremely modest) depth would currently cost on the order of $250 per sample, necessitating some $1.6 million dollars for the analysis. This stands in contrast to the ~$482,000 cost of a 16S rRNA survey (at a cost of $75 per sample).

Thus the reviewer is completely correct that deep metagenomic sequencing is expected to yield similar insights. However, this has not taken place for the vast majority of already-collected microbial community samples, and we argue that such an approach would represent an inefficient way of expending resources for every sample in the future. Similar funds could instead finance a broad 16S-based survey coupled with a modest number of deep metagenomes and a large number of additional complete genomes, with a corresponding decrease in average NSTI and gain in functional insights. These insights would then, for the same cost, also *carry over* from one study to the next; a genome sequenced for one project also lowers NSTI, tightens confidence intervals, and improves prediction accuracy for every subsequent 16S rRNA analysis in similar environments. However, we of course still recommend that large projects perform deep metagenomic sequencing of some (but likely not all) samples in order to carefully check PICRUSt accuracy and also to gain insights into other groups (e.g. phage, fungi, other microbial eukaryotes) that aren’t predicted associated with 16S rRNA marker genes.

We have updated the text of the Discussion to explicitly lay out our recommendations for how large-scale projects might implement such a cost-efficient, rigorous approach to large-scale functional prediction in the context of a time when routine analysis of thousands or (soon) tens of thousands of 16S rRNA samples is commonplace:

“To best leverage the strengths both of (meta)genomic sequencing and of PICRUSt, we recommend its incorporation into marker gene studies using a deliberate, tiered approach. Because phylogenetic dissimilarity among environmental organisms and sequenced genomes (as captured by NSTI) affects PICRUSt accuracy, NSTI values can be calculated from preliminary 16S rRNA data to assess whether reference genome coverage is sufficiently dense to allow for accurate PICRUSt prediction. If adequate reference genomes are not available, additional genome sequences can be collected to fill in phylogenetic “gaps” in the reference database and allow for accurate prediction. This can be performed either through traditional culture-based techniques, single-cell genomic approaches, or deep metagenomic sequencing of samples targeted based on 16S data. If NSTI appears sufficient but additional controls are desired, a preliminary set of paired 16S rRNA and shotgun metagenomic samples can be compared using PICRUSt’s built-in tools to empirically test prediction accuracy on the sample types of interest. Based on such validations from select samples, PICRUSt can then be used to extend approximate functional information from a few costly metagenomes to much larger accompanying 16S rRNA gene sequence collections.”

Finally, we describe how the PICRUSt framework can be refined and elaborated in the future to provide a framework uniting marker gene studies, shotgun metagenomes, and genomic data in a common predictive framework:

“We anticipate several experimental and computational improvements that will further refine the predictive accuracy of PICRUSt. In addition to extending genome coverage and metagenome calibration as above, PICRUSt predictions could also likely be improved by including habitat information in a predictive model. This may provide additional predictive power in that some genes might correlate strongly with environmental parameters as well as phylogenetic similarity to reference organisms. Modification of prediction methods that incorporate information from partial genome sequences could expand the sensitivity of predictions in under-studied environments by including additional reference gene content information. Finally, as reference genome sequence databases continue to expand and incorporate isolates from ever more diverse environments, PICRUSt prediction accuracy will improve by default over time. Predictive metagenomics thus holds the promise of uniting completed genome sequences, 16S rRNA gene studies, and shotgun metagenomes into a single quantitative framework for assessment of community function.”

To conclude, we again thank the editor and the reviewers for this opportunity to clarify the points raised in these reviews. We feel these changes have helped to improve the quality of the manuscript, and we hope they've likewise addressed the referees' and editor’s questions and comments.

Many thanks again for your time and consideration,

Curtis Huttenhower