Accounting for 16S rRNA copy number prediction uncertainty and its implications in bacterial diversity analyses

Yingnan Gao[1], yg5ap@virginia.edu

Martin Wu[1]*, mw4yv@virginia.edu

[1]Department of Biology, University of Virginia, 485 McCormick Road, Charlottesville, VA, 22904, USA

*Corresponding author

Abstract

16S rRNA gene copy number (16S GCN) varies among bacterial species and this variation introduces potential biases to microbial diversity analyses using 16S rRNA read counts. To correct the biases, methods have been developed to predict 16S GCN. A recent study suggests that the prediction uncertainty can be so great that copy number correction is not justified in practice. Here we develop RasperGade16S, a novel method and software to better model and capture the inherent uncertainty in 16S rRNA GCN prediction. RasperGade16S implements a maximum likelihood framework of pulsed evolution model and explicitly accounts for intraspecific GCN variation and heterogeneous GCN evolution rates among species. Using cross validation, we show that our method provides robust confidence estimates for the GCN predictions and outperforms other methods in both precision and recall. We have predicted GCN for 592605 OTUs in the SILVA database and tested 113842 bacterial communities that represent an exhaustive and diverse list of engineered and natural environments. We found that the prediction uncertainty is small enough for 99% of the communities that 16S GCN correction should improve their compositional and functional profiles estimated using 16S rRNA reads. On the other hand, we found that GCN variation has limited impacts on beta-diversity analyses such as PCoA, PERMANOVA and random forest test.

Introduction

The 16S ribosomal RNA (16S rRNA) gene is the gold standard for bacterial and archaeal diversity study and has been commonly used to estimate the composition of bacterial and archaeal communities through amplicon sequencing. Sequence reads are usually matched to reference databases like SILVA [1] and GreenGenes [2] to determine the presence of taxa and

their relative cell abundances. However, the 16S rRNA gene copy number (GCN) can vary from 1 to more than 15 [3,4] and this large copy number variation introduces bias in the relative cell abundance estimated using the gene read counts (thereafter referred to as gene abundance) [5], and consequently it can skew the community profiles, diversity measures and lead to qualitatively incorrect interpretations [5–8]. As a result, it has been argued that 16S rRNA GCN variations should be taken into account in 16S rRNA gene-based analyses [5].

The majority of bacteria species have not been cultured or sequenced and their 16S rRNA GCNs are unknown. Studies have shown that 16S rRNA GCN exhibits a strong phylogenetic signal [5,7], and therefore 16S rRNA GCN can be inferred from closely related reference bacteria. Based on this principle, software has been developed to predict the 16S rRNA GCN [5,7,9,10] in a process often referred to as hidden state prediction [11]. However, a recent study correctly points out that the accuracy of 16S rRNA GCN prediction deteriorates as the minimum phylogenetic distance between the query sequence and the reference sequences increases, and the prediction of 16S rRNA GCN is still an open question [12].

The increasing error of 16S rRNA GCN prediction with increasing phylogenetic distance roots from the stochastic nature of trait evolution, which leads to inherent uncertainty in the predicted trait values. One way of reducing the inherent uncertainty is to improve taxon sampling in the reference phylogeny to reduce the query's phylogenetic distance to the reference [13]. Another way of addressing the inherent uncertainty is to model the uncertainty directly and have a confidence estimate. By doing so, we will be able to determine how confident we should be about a GCN prediction and make meaningful interpretations. Unfortunately, few 16S rRNA

GCN prediction tools provide a confidence estimation for the predicted 16S rRNA GCN, and uncertainty is mostly ignored when interpreting the results of downstream analyses [5,7,10]. For example, PICRUST2 predicts functional profiles of bacterial and archaeal communities from 16S rRNA sequence data. It predicts 16S rRNA GCN for each operational taxonomic unit (OTU) in the community and uses the predicted values (point estimates) to estimate "corrected" relative cell abundances and metagenomes, without accounting for the uncertainty of the predictions. As a result, the impact of uncertainty in 16S rRNA GCN prediction on bacterial diversity analyses remains unknown and needs to be investigated.

Several points need to be considered to properly model the prediction uncertainty. First, because the uncertainty roots from the stochastic nature of trait evolution, we need to develop a good model for 16S rRNA GCN evolution. Previously the evolution of the 16S rRNA GCN trait has been modeled as gradual evolution using the Brownian motion (BM) model [5,7,10]. However, alternative models exist and need to be considered [14–16]. For example, the Ornstein-Uhlenbeck model assumes a centralizing trend towards an optimum [14,15]. As 16S rRNA GCN has been linked to the ecological strategy of bacterial species [17,18] and bacteria diversify across all types of environments, a consistent trend in the evolution of 16S rRNA GCN is unlikely. Thus, a model without any trend like the BM model is preferred. Pulsed evolution (PE) is another model that assumes no trend in evolution. Unlike the BM model, where small trait changes accumulate over time, the PE model postulates that traits evolve by jumps, followed by periods of stasis [14,19]. Previous studies have showed that pulsed evolution is prevalent in the evolution of mammalian body size [14,20]. It has been shown that 16S rRNA GCN of *Bacillus subtilis* can jump from 1 to 6 in a matter of days by gene amplification [21]. On the other hand, it

is well known that the 16S rRNA GCN of some bacterial clades such as the Rickettsiales order, a diverse group of obligate intracellular bacteria, has only one copy of 16S rRNA in their genomes, demonstrating stasis [22,23]. To develop a proper model for 16S rRNA GCN evolution, the tempo and mode of evolution need to be examined.

Secondly, 16S rRNA GCN can vary within the same species [24–27], which introduces uncertainty to GCN prediction that needs to be accounted for. It has been shown that modeling the intraspecific variation is essential for the analysis of comparative trait data and failing to account for this variation can result in model misspecification [14]. Because conspecific strains are usually separated by zero branch length in the phylogeny of the 16S rRNA gene, the intraspecific variation can be modelled as time-independent variation, which can also account for measurement errors [20].

Thirdly, there is notable rate heterogeneity in 16S rRNA GCN evolution. For example, the obligately intracellular bacteria and free-living bacteria with streamlined genomes (e.g., *Rickettsia* and *Pelagibacter*) have elevated molecular evolutionary rates [28,29] and therefore relatively long branches in the 16S rRNA gene phylogeny [30]. Nevertheless, they have only one copy of 16S rRNA in their genomes and the GCNs rarely change [23]. It is expected that the 16S rRNA GCN prediction for this group of bacteria should be accurate despite their large phylogenetic distances to the reference genomes. Such examples suggest that the rate heterogeneity of 16S rRNA GCN evolution should be systematically evaluated and modelled properly. However, no previous methods have evaluated and modeled such evolution rate heterogeneity, leading to potential model misspecification in 16S rRNA GCN predictions.

Here, we develop a novel tool *RasperGade16S* that employs a heterogeneous pulsed evolution model for 16S rRNA GCN prediction. Using simulation and cross-validation, we show that *RasperGade16S* outperforms other methods in terms of providing significantly improved confidence estimates. We show that even if we cannot eliminate the inherent uncertainty of 16S rRNA GCN prediction, having an accurate confidence estimate allows us to incorporate it in downstream analyses and therefore to make better inferences from the results with confidence intervals. We show that correcting 16S rRNA GCN improves the relative cell abundance estimates of the bacterial communities and is expected to be beneficial for more than 99% of 113842 environmental samples we have analyzed. We also show that GCN correction is unnecessary for beta-diversity analyses because it has limited impact on the results.

Methods

Preparing reference genomes and the 16S rRNA reference phylogeny

We downloaded annotated RNA gene sequences from 21245 complete bacterial genomes in the NCBI RefSeq database (Release 205) on April 9, 2021. For each genome, we counted the number of genes whose products are annotated as 16S rRNA genes. For genomes with multiple copies of 16S rRNA gene, we aligned the 16S rRNA sequences using MAFFT [31] (with parameters: --maxiterate 1000 --globalpair) and picked the 16S rRNA gene sequence that has the highest average similarity (calculated as the proportion of identical bases in the alignment) to other 16S rRNA gene sequences in the genome as the representative sequence. To remove potential errors introduced by mis-assembled genomes [32], we removed genomes whose 16S rRNA GCN differs from their 5S rRNA GCN (counted using the same strategy as 16S rRNA

GCN) by greater than 2 copies, genomes whose 16S rRNA sequence contains ambiguous bases, or genomes on the list of withheld genomes in the curated ribosomal RNA operon copy number database rrnDB [3]. The 17 genomes in the rrnDB withheld list are rejected from rrnDB because their 16S rRNA genes are missing, the 16S rRNA GCNs are too high, or the genomes have inconsistent meta data (https://rrndb.umms.med.umich.edu/withheld/). We aligned the remaining representative 16S rRNA gene sequences using HMMER version 3.2 [33] (hmmalign with parameters: --trim --dna –mapali) with the hidden Markov model (HMM) built from the GreenGenes 13.8 16S rRNA gene alignment (hmmbuild with default parameters), and trimmed the alignment with a mask from the GreenGenes database [2]. The HMM, profile alignment and the alignment mask are included in the R package RasperGade16S. After collapsing identical 16S rRNA alignments, 6408 representative sequences remained. They serve as the reference sequences and their taxonomies of are summarized in Table S1. We built a reference tree from the trimmed alignment using RAxML version 8.2 [34] with options -f d -m GTRGAMMA. We used the Deinococcus-Thermus group to root this reference phylogeny. To examine the effect of sequence alignment on model fitting, we also used the 16S rRNA HMM profile from the software Barrnap [35] to align the 16S rRNA genes (hmmalign with default parameters). We trimmed the alignment using a consensus posterior probability threshold of 0.95 (esl-alimask with parameters: -p --ppcons 0.95) and made a 16S rRNA phylogeny as described above.

Evaluating time-independent variation in 16S rRNA GCN

To evaluate the extent of 16S rRNA GCN intraspecific variation, we compared GCN between 5437 pairs of genomes with identical 16S rRNA gene alignments. To formally test whether accounting for time-independent variation is necessary, we modeled time-independent variation

as a normal white noise, and fitted the Brownian motion (BM) model to the evolution of 16S rRNA GCN in the 6408 reference genomes, with and without time-independent variation. We then calculated the likelihood and chose the best model using the Akaike Information Criterion (AIC).

Evaluating the rate heterogeneity of 16S rRNA GCN evolution

To estimate the degree of rate heterogeneity in 16S rRNA GCN evolution, we calculated the local average rate of evolution for each genus that contains at least 10 genomes in the reference phylogeny and examined the distribution of the average rates among genera. The average rate of a genus is calculated as the variance of phylogenetically independent contrasts (PICs) [36] of GCN within the genus.

Modeling 16S rRNA GCN evolution with homogeneous and heterogeneous pulsed evolution models

Using the R package *RasperGade* [37], we fitted one PE model to the entire reference phylogeny and calculated the likelihood of this homogeneous PE model. An analysis of the variance of the PICs associated with each genus indicated that there is a slowly-evolving group and a regularly-evolving group, with the average rate of the slowly-evolving group estimated to be at least 100-fold lower than that of the regularly-evolving group (Figure S1). To model the rate heterogeneity, we created two PE models: $PE_{regular}$ for the regularly-evolving group and PE_{slow} for the slowly-evolving group. We then use a two-step iterative binning procedure to estimate the parameters of $PE_{regular}$ and PE_{slow} (i.e., jump size and frequency). The $PE_{regular}$ model was initiated to take the parameter values of the homogeneous PE model. PE_{slow} was initiated to have a jump size equal

171 to that of $PE_{regular}$ but a jump frequency100-fold lower. In our first round of binning, from the
172 root to the tip of the reference phylogeny, we classified each node into the regularly- or slowly-
173 evolving group by testing which model ($PE_{regular}$ or PE_{slow}) provided a better fit. We merged
174 neighboring nodes belonging to the same group into one neighborhood and flipped neighborhood
175 assignment if the flip resulted in an improved overall AIC value. After the first round of binning,
176 we updated $PE_{regular}$ and PE_{slow} by fitting $PE_{regular}$ to nodes that were classified as regularly-
177 evolving and PE_{slow} to slowly-evolving nodes. We used the updated models to perform a second
178 round of binning to assign each node in the phylogeny to a group. Finally, we calculated r, the
179 rate of evolution in each group, as the process variance per unit branch length defined in a
180 previous study [14]. We then rescaled the reference tree by multiplying the branches in the
181 slowly-evolving group by the ratio $r_{slow}/r_{regular}$. To accommodate time-independent variation in
182 the tip trait values, we calculated a branch length over which the process variance of the fitted
183 pulsed evolution model is equal to the model's time-independent variation, and added this branch
184 length to each tip branch. We compared the homogeneous and heterogeneous PE models by AIC.
185
186 *Predicting 16S rRNA GCN*
187 We used the R package *RasperGade16S* to predict 16S rRNA GCN using the heterogeneous
188 pulsed evolution model. *RasperGade16S* first assigns the query sequence to either the regularly-
189 evolving or the slowly-evolving group based on where it is inserted in the reference phylogeny.
190 For a query sequence inserted into the slowly-evolving group, its insertion branch length is
191 scaled by the ratio $r_{slow}/r_{regular}$. For a query sequence inserted into the regularly-evolving group, a
192 small branch length is added to the insertion branch to represent the estimated time-independent
193 variation. *RasperGade16S* then predicts the GCN of the query using the rescaled reference

phylogeny. Because 16S rRNA GCN is an integer trait, the continuous prediction from hidden state prediction is rounded and a confidence (probability) that the prediction is equal to the truth is estimated by integrating the predicted uncertainty distribution. We marked the 16S rRNA GCN prediction with a confidence smaller than 95% as unreliable, and otherwise as reliable. As a comparison, we also predicted GCN using PICRUST2, which employs multiple hidden state prediction methods in the R package *castor* [38] for 16S rRNA GCN predictions. We selected three methods by which confidence can be estimated: the phylogenetically independent contrast (pic) method, the maximum parsimony (mp) method, and the empirical probability (emp) method. Otherwise, we run PICRUST2 using default options and the unscaled reference phylogeny. For the pic method, the confidence measure is not provided by the hidden state prediction function of the *castor* package, and thus we used a customized R script to reroot the tree at the query sequence and estimated the confidence of the prediction.

We did not test the tools CopyRighter [7] and PAPRICA [9] directly in this study because 1) neither provides the option of using a user-supplied reference data, and 2) neither provides uncertainty estimates (i.e., confidence intervals) of its predictions, which is the primary focus of this study. However, CopyRighter employs the pic method and we expect its performance to be highly similar to PICRUST2 running the pic method. As PAPRICA [9]employs the subtree average method, a continuous analogue to the emp method, we expect that its performance will be highly similar to that of PICRUST2 running the emp method.

Adjust NSTD and NSTI with rate heterogeneity

The adjusted nearest-sequenced-taxon-distances (NSTDs) [12] is calculated using the rescaled reference tree. The adjusted nearest-sequenced-taxon-index (NSTI) [10] is calculated as the weighted average of adjusted NSTDs of the community members.

Validating the quality of predicted 16S rRNA GCN and its confidence estimate

We used cross-validations to evaluate the quality of 16S rRNA GCN prediction and its confidence estimate, and how they vary with NSTD. We randomly selected 2% of the tips in the reference phylogeny as the test set and filtered the remaining reference set by removing tips with a NSTD to any test sequence smaller than a threshold. We then predicted the 16S rRNA GCN for each tip in the test set using the filtered reference set. We conducted cross-validation within 9 bins delineated by 10 NSTD thresholds: 0, 0.002, 0.005, 0.010, 0.022, 0.046, 0.100, 0.215, 0.464 and 1.000 substitutions/site, and for each bin we repeated the cross-validation 50 times with non-overlapping test sets. We evaluated the quality of the 16S rRNA GCN prediction by the coefficient of determination (R^2), the fraction of variance in the true copy numbers explained by the prediction. We evaluated the quality of confidence estimate by precision and recall. Precision is defined as the proportion of accurately predicted 16S rRNA GCN in predictions considered as reliable (with $\geq$ 95% confidence), and recall is defined as the proportion of reliable predictions in the accurately predicted 16S rRNA GCNs. We averaged the R^2, precision and recall for the 50 cross-validations in each bin.

Simulating 16S rRNA GCN variation under pulsed evolution model

To evaluate the performance of different prediction methods when the trait evolves under the pulsed evolution model, we simulated the evolution of 16S rRNA GCN using the fitted heterogeneous pulsed evolution model. Specifically, we first simulated the number of jump events for each branch in the reference phylogeny based on the rate group that branch belongs to. Then we simulated the continuous trait change for each branch using the corresponding number of jump events. We added up the continuous trait change from the root to the tips to get the tip trait values and rounded them to the nearest integers. This set of simulated 16S rRNA GCN is referred to as ST1.

Simulating bacterial communities with 16S rRNA GCN variation

To evaluate the effect of 16S rRNA GCN correction on bacterial diversity analyses, we simulated two sets of bacterial communities using the reference genomes: one set for relative cell abundance analyses (SC1) and the other set for beta-diversity analyses (SC2). We treated each reference genome as one OTU. For SC1, we simulated a total of 100 communities. For each simulated community, we randomly selected 2000 OTUs from the reference genomes, and assigned each OTU a cell abundance randomly drawn from a log-series species abundance distribution.

In SC2, we simulated communities in two environments to evaluate the effect of 16S rRNA GCN correction on beta diversity analyses. We simulated 10 communities per environmental type and 2000 OTUs per community. We controlled the community turnover rate by controlling the number of unique OTUs in each community. For example, at a turnover rate of 10%, a community would have 200 unique OTUs and 1800 core OTUs that are shared among all

communities. We varied the turnover rate from 10% to 90% at 10% intervals. To control for the effect size of environmental type, we assigned 5 (0.25%), 20 (1%) or 100 (5%) signature OTUs to each environmental type. These signature OTUs were shared between the two environmental types but were twice more likely to be placed in top ranks of the log-series distribution (i.e., to be more abundant) than the non-signature OTUs in their corresponding environment. The 16S rRNA GCN of each OTU was assigned randomly from the reference genomes' GCN. We simulated 50 batches of communities for each combination of turnover rate and signature OTU number, resulting in 27000 simulated communities in SC2.

Evaluating the effect of 16S rRNA GCN correction on relative cell abundance estimation

We evaluated the effect of 16S rRNA GCN correction on the simulated bacterial communities (SC1). To estimate the confidence interval (CI) of the corrected relative cell abundance of each OTU in a community, we randomly drew 1000 sets of 16S rRNA GCNs from their predicted uncertainty distribution. For each set of 16S rRNA GCNs, we divided the gene read count of OTUs by their corresponding 16S rRNA GCNs to get the corrected cell counts. The median of the corrected cell count for each OTU in the 1000 sets is used as the point estimate of the corrected cell count, and the OTU's relative cell abundance is calculated by normalizing the corrected cell count with the sum of corrected cell counts of all OTUs in the community. The 95% CI for each OTU's relative cell abundance is determined using the 2.5% and 97.5% quantiles of the 1000 sets of corrected relative cell abundances. The OTU with the highest corrected relative cell abundance is considered the most abundant taxon. The support value for the most abundant OTU is calculated as the empirical probability that the OTU has the highest cell abundance in the 1000 sets of corrected cell abundances. We calculated the coverage probability of the CI as the

283 empirical frequency that the relative gene abundance or true relative cell abundance is covered
284 by the estimated CI. We evaluated the effect of 16S rRNA GCN correction on relative cell
285 abundance estimation at different NSTD thresholds.
286
287 *Evaluating the effect of 16S rRNA GCN correction on beta-diversity analyses*
288 We used the Bray-Curtis dissimilarity and Aitchison distance for any beta-diversity analysis that
289 requires a dissimilarity or distance matrix and evaluated the effect of 16S rRNA GCN correction
290 on the simulated bacterial communities (SC2). To correct for 16S rRNA GCN variation in beta-
291 diversity analyses, we divided the gene abundance of each OTU by its predicted 16S rRNA GCN
292 and calculated the corrected relative cell abundance table and the corresponding
293 dissimilarity/distance matrix. We used the corrected cell abundance table to generate the
294 principal coordinates analysis (PCoA) plot and to conduct the permutational multivariate
295 analysis of variance (PERMANOVA) and the random forest test with the R package *vegan* and
296 *randomForest*, respectively.
297
298 *Predicting 16S rRNA GCN for SILVA OTUs*
299 We downloaded 592605 full-length representative bacterial 16S rRNA sequences of non-
300 redundant OTUs at 99% similarity (OTU99) in the SILVA release 132 [1]. We aligned and
301 trimmed the sequences using the method described above. We then inserted the OTUs into the
302 reference phylogeny using the evolutionary placement algorithm (EPA-ng) [39] with the model
303 parameters estimated by RAxML when building the reference phylogeny. We limited the
304 maximum number of placements per SILVA representative sequence to 1. We predicted the 16S

rRNA GCN for each SILVA OTU99 as described above using the heterogeneous pulsed evolution model and calculated adjusted NSTDs.

Evaluating the effect of GCN correction in HMP1 and EMP dataset

To check the effect of 16S rRNA GCN correction in empirical data, we analyzed the 16S rRNA V1-V3 amplicon sequence data of the first phase of Human Microbiome Project (HMP1) [40] and the sequence data processed by Deblur [41] in the first release of the Earth Microbiome Project (EMP) [42]. The 16S rRNA GCN for each OTU in the HMP1 and EMP datasets was predicted using *RasperGade16S*. We picked 2560 samples in the HMP1 dataset with complete metadata and used the 2000-sample subset of EMP, and determined the adjusted NSTI and relative cell abundance in each community as described above. For beta-diversity, we randomly picked 100 representative samples from each of the 5 body sites in the HMP1 dataset and analyzed their beta-diversity as described above. For the EMP dataset, we analyzed the beta-diversity within each level-2 EMP ontology (EMPO) category (around 400 to 600 samples per category).

Examining the adjusted NSTI of empirical bacterial communities

To check the predictability of 16S rRNA GCN in empirical data, we examined bacterial communities surveyed by 16S rRNA amplicon sequencing in the MGnify resource platform [43] that were processed with the latest two pipelines (4.1 and 5.0). The MGnify resource platform uses the SILVA database release 132 [1] for OTU-picking in their latest pipelines, and therefore predicted GCNs for SILVA OTUs can be used directly. We filtered the surveyed communities from the MGnify platform so that only communities with greater than 80% of their gene reads

mapped to the SILVA reference at a similarity of 97% or greater were included. This filtering yielded 113842 bacterial communities representing a broad range of environment types. We calculated the adjusted NSTI for each community and examined the adjusted NSTI distribution in various environmental types.

Results

Time-independent variation is present in 16S rRNA GCN evolution

To evaluate the extent of intraspecific variation in 16S rRNA GCN, we examined 5437 pairs of genomes with identical 16S rRNA gene alignments. The 16S rRNA GCN differs in 607 (11%) of them, suggesting the presence of significant intraspecific variation or time-independent variation. Using AIC, we found that incorporating time-independent variation with the BM model greatly improves the model fit (Table 1), indicating the necessity to take time-independent variation into account in 16S rRNA GCN prediction. In addition, we observed that the rate of evolution in the fitted BM model is inflated by 1670 folds when time-independent variation is not included in the model. Such inflation in the estimated rate of evolution will lead to overestimation of uncertainty in the 16S rRNA GCN prediction.

Table 1. The AICs of Brownian motion model and pulsed evolution model.

Model	BM	BM (with time-independent variation)	PE (with time-independent variation)
Homogenous model	34338	18028	-7925
Heterogeneous model	NA	NA	-15395

Pulsed evolution model explains the 16S rRNA GCN evolution better than the Brownian motion model

When predicting traits using phylogenetic methods, the BM model is commonly assumed to be the model of evolution. We have shown that PE model is a better model for explaining the evolution of [14,20] bacterial genome size [37], prompting us to test whether pulsed evolution can be applied to explain 16S rRNA GCN evolution as well. Using the R package *RasperGade* that implements the maximum likelihood framework of pulsed evolution described by Landis and Schraiber [14], we fitted the PE model with time-independent variation to the same dataset. Table 1 shows that the PE model provides a significantly better fit than the BM model, indicating that 16S rRNA GCN prediction should assume the PE model instead of the BM model. Fitted model parameters are not sensitive to the HMM profiles used for aligning the 16S rRNA sequences (Table S2).

Substantial rate heterogeneity exists in 16S rRNA GCN evolution

To systematically examine the rate heterogeneity of 16S rRNA GCN evolution in the reference genomes, we first used the variance of PICs as an approximate estimate of the local evolution rate of 16S rRNA GCN. We found that the rate of evolution varies greatly among genera (Figure S1), but can be roughly divided into two groups with high and low rates of evolution. Therefore, we developed a heterogeneous pulsed evolution model where all jumps are the same size but the frequency of jumps varies between two groups to accommodate the heterogeneity among different bacterial lineages. Using a likelihood framework and AIC, we classified 3049 and 3358 nodes and their descending branches into slowly-evolving and regularly-evolving groups respectively (Figure S2). The frequency of jumps in the regularly-evolving group is 145 folds of

the frequency in the slowly-evolving group (Table S3). The heterogeneous PE model provides the best fit among all models tested (Table 1), indicating that a heterogeneous PE model should be assumed in predicting 16S rRNA GCN.

Apart from the rate of pulsed evolution, we also observed heterogeneity in time-independent variation: for the slowly-evolving group, the fitted model parameters indicate no time-independent variation, while for the regularly-evolving group, the magnitude of time-independent variation is approximately 40% of a jump in pulsed evolution (Table S3). The presence of time-independent variation caps the confidence of prediction in the regularly-evolving group at 85%, which can only be achieved when the query has identical 16S rRNA gene alignment to one of the reference genomes.

The effect of NSTD on accuracy and uncertainty of 16S rRNA GCN predictions

Because of the stochastic nature of evolution, the inherent uncertainty in hidden state prediction accumulates over time, and consequently the accuracy of the prediction decreases as the phylogenetic distance to the reference increases [12]. To get a better understanding of the relationships between NSTD and metrics that measure the accuracy and uncertainty of the prediction, we performed a cross-validation experiment using simulated datasets. We simulated the evolution of 16S rRNA GCN under the fitted heterogeneous pulsed evolution model along the reference phylogeny (ST1), and predicted the simulated GCN of the tips using different methods: the pulsed evolution model (PE), the BM model (pic), maximum parsimony (mp) and empirical probability (emp). As expected, we found that the true uncertainty of the prediction, as predicted by the pulsed evolution model under which the simulated GCN evolves, increases with

the NSTD. The true confidence of the prediction, calculated as 1 - uncertainty, decreases with the NSTD (Figure 1A, purple line). The pic method predicts greatly inflated uncertainty (Figure 1A, red line), while the mp method predicts no uncertainty at all (Figure 1A, blue line). The emp method predicts intermediate uncertainty that is greater than the true uncertainty at small NSTDs, but smaller than the truth at large NSTDs (Figure 1A, green line). We used the coefficient of determination (R^2) of the predicted trait values to the truth to evaluate the accuracy of the prediction. Figure 1B shows the PE method performs the best. It is followed by the mp and the pic method. The emp method performs the worst. As observed in previous research [12], the accuracy decreases as the NSTD increases (Figure 1B).

Because we will use the uncertainty measure to evaluate the reliability of the prediction, we tested whether the uncertainty predicted by the various methods we compared here captures the true reliability of the prediction. Specifically, we calculated the precision and recall of predictions with a confidence of 95% or greater to examine the recovery of highly reliable predictions. We define precision as the proportion of accurately predicted 16S rRNA GCN in predictions with ≥ 95% confidence, and recall as the proportion of predictions with ≥ 95% confidence in the accurately predicted 16S rRNA GCNs. Ideally, the precision should be greater than 95% throughout the NSTD spectrum, while the recall should gradually drop as the NSTD and the uncertainty of prediction increase. We found that the PE method yields high recall at small NSTDs and it decreases as NSTD and uncertainty in the prediction increase (Figure 1C, purple line). In terms of precision, the PE method yields high precision throughout the spectrum of NSTD (Figure 1D, purple line). The pic method has the lowest recall rate at the smallest NSTD and no recovery beyond as it overestimates the uncertainty (Figure 1C and D, red lines).

On the contrary, the mp method yields the highest recall regardless of NSTD as it predicts no uncertainty at all (Figure 1C, blue line), but it suffers from lower precision than the PE method and the precision drops quickly as NSTD increases (Figure 1D, blue line). In essence, the uncertainty estimated using the pic method is so great that few predictions can be trusted. On the other hand, according to the mp method, there is no uncertainty in a prediction and every prediction is reliable. The emp method shows a similar trend in recall compared to the PE method, but suffers from lower precision than the PE method.

In summary, the uncertainty of 16S rRNA GCN prediction increases with the increase of NSTD, and as a result, the accuracy of prediction drops as the NSTD increases for all methods. The recall rate of highly reliable predictions also drops with the increasing NSTD and uncertainty, while the precision can remain high throughout the NSTD spectrum.

RasperGade16S improves confidence estimate for 16S rRNA GCN prediction in empirical data

Using 16S rRNA GCN from the 6408 complete genomes in the reference phylogeny for cross-validation, we compared the performance of various methods in accuracy and confidence estimates. In general, the trends of uncertainty, accuracy, precision and recall plotted against the NSTD (Figure 2) are very similar to those observed in the simulation study (Figure 1), indicating that *RasperGade16S* models the 16S rRNA GCN evolution reasonably well. As observed in the simulation, the pic and mp methods produce very large and zero uncertainty respectively (Figure 2A), leading to both poor precision and recall rates (Figure 2C and 2D). The emp method performs the worst in terms of accuracy. The PE method produces the best overall precision, achieving an average precision rate of 0.96 throughout the NSTD spectrum. Overall, the PE

method provides one of the best accuracies and the best confidence estimate for 16S rRNA GCN prediction over the full spectrum of NSTD, and should be preferred when predicting 16S rRNA GCN.

As NSTD of the 16S rRNA gene also depends on the sequence alignment and how it is trimmed, it can vary from study to study using the same set of reference sequences. Therefore, to put NSTD values of this study in a taxonomic context, we calculated the NSTDs between taxa at different taxonomical levels. For example, at the species level, we calculated the NSTD of a species to another species within the same genus. We found that the median NSTD between congeneric species is around 0.01 substitutions/site and the maximum NSTD threshold (0.464 substitutions/site) in our cross-validation experiment roughly correspond to a taxonomic distance somewhere between class and order (Figure S3).

Copy number correction improves relative cell abundance estimation

Because 16S rRNA GCN variation biases gene abundances disproportionately among the community members, it distorts the relative cell abundance estimated from the gene abundance [5]. From theoretical calculations, in general, community members with lower relative cell abundances suffer from greater impacts by 16S rRNA GCN variation, while those with higher relative cell abundances appear to be less affected by it (Figure 3A). The impact of 16S rRNA GCN variation also depends on the deviation of a member's GCN from the average GCN of the community members, with larger deviations resulting in larger impacts (Figure 3A). When a member's 16S rRNA GCN is greater than the average GCN of the community, its relative abundance will be overestimated. On the other hand, when a member's 16S rRNA GCN is

462 smaller than the average GCN, its relative abundance will be underestimated. In simulated
463 dataset SC1, we found that 16S rRNA GCN variation has a large detrimental effect on the
464 estimated relative cell abundance (Figure 3B). On average, the relative cell abundance estimated
465 using the gene abundance increased or decreased by 1.8-fold compared to the true relative cell
466 abundance, and the empirical probability of correctly identifying the most abundant OTU based
467 on the gene abundance is only around 13% (Figure 3C). Correcting for 16S rRNA GCN
468 improves the estimated relative cell abundance (Figure 3B). As expected, the improvement is
469 greatest when the adjusted NSTI is small (i.e., when there are closely related reference genomes),
470 and it gradually diminishes when the adjusted NSTI increases. At the smallest adjusted NSTI, the
471 average fold change of the estimated relative cell abundance decreases to 1.1-fold after 16S
472 rRNA GCN correction and the empirical probability of correctly identifying the most abundant
473 OTU increases to around 65% (Figure 3C).
474
475 Because we predict each OTU's 16S rRNA GCN with a confidence estimate, we can provide 95%
476 confidence intervals (95% CIs) for their relative cell abundance as well. Ideally, 95% of the true
477 relative cell abundances should be covered by the 95% CIs. Figure 3D shows that the average
478 coverage probability of the true relative cell abundance is about 98% across NSTD cutoffs,
479 indicating that our 95% CIs are slightly over-conservative. Similarly, we can also calculate the
480 coverage probability of our 95% CI to the relative gene abundance. As expected, when the
481 coverage probability to the relative gene abundance increases, the improvement by GCN
482 correction (quantified by the relative reduction in the difference between the estimated and true
483 cell abundances) decreases (Figure 3E), and that when this coverage probability is below 95%,
484 GCN correction always results in strong improvement in relative cell abundance estimates. In

empirical studies when the true abundance is unknown, we can use the coverage probability to the relative gene abundance as a conservative statistic to decide if GCN correction for a community will likely improve the relative abundance estimation or not. For the most abundant OTU in the community, we can calculate its support value from the 16S rRNA GCN's confidence estimates. We found that the calculated support value matches the empirical probability that the most abundant OTU is correctly identified (Figure 3C).

To demonstrate the effect of 16S rRNA GCN correction in empirical data, we analyzed the data from the first phase of the Human Microbiome Project (HMP1) and the 2000-sample subset of Earth Microbiome Project (EMP). We found that on average the relative cell abundance with and without 16S rRNA GCN correction changes around 1.3-fold in HMP1 and 1.6-fold in EMP. Since the true abundance of OTUs is unknown, we use the coverage probability of 95% CIs to the relative gene abundance described above to evaluate the effect of GCN correction. Our results indicate that a majority of HMP1 (over 82%) and EMP (over 90%) samples have a coverage probability below 95% (as shown in Figure 3F). Our simulations demonstrate that GCN correction improves the accuracy of relative cell abundance estimation in samples with coverage probability less than 95% (as demonstrated in Figure 3E), suggesting that GCN correction will likely improve relative cell abundance estimates in these HMP1 and EMP samples. In terms of the most abundant OTU, we found that the identity of the most abundant OTU changes after copy number correction in around 20% and 31% of the communities in HMP1 and EMP respectively. The support values for the most abundant OTUs are around 0.85 on average in both datasets, indicating high confidence in the identification of the most abundant OTUs.

Copy number correction provides limited improvements on beta-diversity analyses

Because 16S rRNA GCN variation affects the estimated relative cell abundances, it may also affect the beta-diversity analyses such as PCoA, PERMANOVA, and the random forest test that use the relative cell abundance information. To examine the effect of 16S rRNA GCN variation on these analyses, we simulated communities at different turnover rates in two types of environments where 0.25%, 1% or 5% of the OTUs are enriched in one environment compared to the other (the SC2 dataset). We performed beta-diversity analyses on the simulated data and generated the PCoA plots (an example with 0.25% enriched signature OTU is given in Figure 4). We found that when the relative gene abundance is used to calculate the Bray-Curtis dissimilarity or the Aitchison distance, the positions of the samples in the PCoA plot shift from their positions based on the true relative cell abundance (solid lines in Figure 4A and B), although this shift is much smaller if the Aitchison distance is used. Correcting for 16S rRNA GCN reduces about 56% of the shift in the Bray-Curtis dissimilarity space ($P<0.001$, paired t-test, Figure 4A) while it reduces about 85% of the shift in the Aitchison distance space ($P<0.001$, paired t-test, Figure 4B). Despite the shift in the PCoA plot, we found that the clustering of communities does not seem to be affected by the 16S rRNA GCN variation. The results with 1% and 5% enriched signature OTUs are similar to the examples shown in Figure 4.

In addition to the PCoA plot, we observed a limited effect of 16S rRNA GCN variation on other beta-diversity analyses. In PERMANOVA, depending on the metric used, the signature OTU numbers and turnover rates, the proportion of variance explained (PVE) by the environmental type using the true cell abundances ranges from 5.27 to 17.20% on average. Using gene abundance, the average PVE ranges from 5.27 to 17.22% and the change in PVE is not

statistically significant regardless the metric used, the signature OTU numbers, or the turnover rates ($P>0.002$, paired t-test with Bonferroni correction, $\alpha=9.26\times10^{-4}$, Table S4), indicating that PERMANOVA is not very sensitive to 16S rRNA GCN variation.

It is a common practice to compare the relative cell abundance of OTUs of interest between environments. We found that such comparison is also not sensitive to 16S rRNA GCN variation (Table S4), with the fold-change of relative cell abundance estimated using the gene abundance and the truth highly concordant ($R^2 > 0.99$). For the top OTUs with the highest fold-change in true cell abundance (i.e., signature OTUs), on average more than 98% of them are also the top OTUs with highest fold-change in gene abundance, indicating that abundance difference across environments is not sensitive to 16S rRNA GCN variation. Alternatively, we can use the random forest test to identify OTUs that are differentially abundant between environments by their importance scores (defined as the mean decrease in classification accuracy if removed from the data). We found that the top OTUs ranked by the importance score recovers from 20.0% to 89.0% of the signature OTUs when the true cell abundances were used (Table S4). When the gene abundances were used, this recovery rate changes to from 18.0% to 89.32% (Table S4), and the change is not statistically significant ($P > 0.032$, paired t-test with Bonferroni correction, $\alpha=1.85\times10^{-3}$). Correcting for 16S rRNA GCN changes the recovery rate to from 17.8% to 89.2% (Table S4), and the change is not significant either ($P>0.041$, paired t-test with Bonferroni correction, $\alpha=1.85\times10^{-3}$).

To examine the effect of 16S rRNA GCN variation correction on beta-diversity in empirical data, we analyzed the beta-diversity using the HMP1 and EMP datasets. Because we observed that the

554 effect of GCN correction is independent of the metric used in beta-diversity analyses, we only
555 used Bray-Curtis dissimilarity in HMP1 and EMP datasets. We found that correction of 16S
556 rRNA GCN does not seem to affect the clustering of communities by body sites in the HMP1
557 PCoA plot. Pairwise PERMANOVA shows that the mean PVE by the body site in HMP1 is 14.9%
558 before 16S rRNA GCN correction and decreases marginally to 14.6% after correction, and the
559 PVEs using the gene abundance and the corrected cell abundance are also highly concordant
560 (R^2>0.98). In EMP, within each level-2 environment (EMPO2), the average PVE by level-3
561 environment (EMPO3) remains at 7.7% before and after 16S GCN correction and the PVEs
562 using the gene abundance and the corrected cell abundance are highly concordant (R^2>0.99) as
563 well. On the other hand, pairwise random forest tests yield similar results before and after 16S
564 rRNA GCN correction, with around 9 out of the top 10 features identified by the random forest
565 test remaining unchanged before and after correction in HMP1 and around 8 out of the top 10
566 unchanged in EMP. In terms of the fold-change of relative cell abundances between body sites,
567 we found that copy number correction has little impact as the estimated fold-change before and
568 after correction are highly similar (R^2>0.95) in both datasets.
569
570 *Predicting 16S rRNA GCNs for SILVA OTUs*
571 Using *RasperGade16S*, we predicted the 16S rRNA GCN for 592605 bacterial OTUs (99%
572 identity) in the release 132 of the SILVA database. Overall, the median adjusted NSTD for all
573 bacterial OTUs is 0.070 substitutions/site, and 34.7% of the predictions have a high confidence
574 of 95% or greater, and 74.9% of the predictions have a moderate confidence of 50% or greater
575 (Table 2). This shows that for most OTUs in the SILVA database, the phylogenetic distance to a
576 reference 16S rRNA is small enough that we can have reasonable confidence in the predictions.

In comparison, randomly guessing has a null confidence of around 6.7% (1 out of 15 possible GCNs). Among major phyla with more than 10000 OTUs, the proportion of highly confident predictions varies greatly (Table 2), with Cyanobacteria having the lowest proportion of 19.1% and Acidobacteria having the highest proportion of 50.4%. Similarly, the proportion of moderately confident predictions varies from 58.3% to 89.5% among these phyla. Interestingly, the proportions of highly confident predictions closely match the proportions of slowly-evolving OTUs in each phylum (Table 2), suggesting a causal relationship between them.

Table 2. Summary of SILVA 16S rRNA GCN predictions.

Taxonomic group	Number of OTUs	Median adjusted NSTD (substitutions/site)	Proportion of OTUs in slowly-evolving group	Proportion of highly confident predictions	Proportion of moderately confident predictions
Bacteria	592605	0.070	34.9%	34.7%	74.9%
Proteobacteria	238929	0.062	43.4%	43.1%	85.4%
Firmicutes	149757	0.091	21.6%	21.5%	68.9%
Actinobacteria	60510	0.061	45.2%	45.2%	89.5%
Bacteroidetes	55663	0.117	29.9%	29.8%	58.3%
Acidobacteria	14534	0.006	50.4%	50.4%	82.7%
Cyanobacteria	13970	0.285	19.9%	19.1%	60.5%

Highly confident predictions are defined as predictions with a confidence of 95% or greater.
Moderately confident predictions are defined as predictions with a confidence of 50% or greater.

Vast majorities of bacterial community studies should benefit from copy number correction

To examine if analysis of real communities would benefit from 16S rRNA GCN correction, we calculated the adjusted NSTI for 113842 communities in the microbiome resource platform MGnify (formerly known as EBI Metagenomics) [43] that passed our quality control. These

593 microbiomes were sampled from various environments and include host-associated microbiomes
594 in animals and plants and free-living microbiomes in soil and aquatic environments (Table S5).
595 The adjusted NSTI varies greatly among samples and the median across all samples is 0.01
596 substitutions/site. For example, the smallest adjusted NSTI (8.2×10^{-7} substitutions/site) comes
597 from a human vaginal clinical sample that is dominated by one OTU (relative cell
598 abundance >0.98). This OTU is closely related to *Lactobacillus iners* in the slowly-evolving
599 group, which results in the extremely small adjusted NSTI. On the other hand, the largest
600 adjusted NSTI (0.6 substitutions/site) belongs to a sample from the rumen of dairy cows. The
601 majority of OTUs in this sample has an adjusted NSTD greater than 0.1 substitutions/site and
602 account for more than 90% of the total cell abundance. In the simulated communities, we
603 observed that GCN correction significantly improves the estimated relative cell abundances
604 ($P<0.001$, paired t-test) even when the adjusted NSTI reaches 0.3 substitutions/site. We found
605 that more than 99% of the communities from MGnify have an adjusted NSTI less than 0.3
606 substitutions/site, suggesting that they should benefit from 16S rRNA GCN correction when
607 estimating the relative cell abundances. The distribution of adjusted NSTI varies among different
608 environmental types (Figure 5), but the proportion of communities that will likely benefit from
609 16S rRNA GCN correction remains high, ranging from 98% to 100%.
610

611 **Discussion**

612 16S rRNA GCN variation skews bacterial community composition estimated from the 16S
613 rRNA read count. To correct for the bias introduced by the GCN variation, several methods have
614 been developed to predict GCN from reference genomes. A recent study has pointed out that the
615 GCN predictions come with inherent uncertainty, particularly for these taxa without closely

related genomes [12]. The concern is that inaccurate predictions can introduce noise to community compositions that can be worse than the original GCN-related biases, thereby raising doubt about the usefulness of the 16S rRNA GCN correction in bacterial diversity analyses [8,12].

We address the inherent uncertainty problem in 16S rRNA GCN prediction by directly measuring it with confidence estimates. Using simulations and cross-validation, we show that the PE method implemented in *RasperGade16S* outperforms other methods in both the precision and recall rates. This method's strength comes from three features of its modeling of the 16S rRNA GCN evolution: implementation of a pulsed evolution model and accounting for the rate heterogeneity and time-independent trait variation. Pulsed evolution model expects no trait changes to occur over a short branch as jumps are not likely to happen on that branch. This leads to a higher confidence to 16S rRNA GCN prediction with a short NSTD, and thus improves the recall of the accurate predictions. By incorporating rate heterogeneity, we can make predictions in the slowly-evolving groups with high confidence, even when their NSTDs are large, thereby further improving the overall precision and recall rates. In the reference phylogeny, 48% of branches were estimated to fall within this slowly-evolving group, whose evolution rate is 145 times slower compared to that of the regularly-evolving group. The third source of improvement for *RasperGade16S* comes from accounting for time-independent variation, which can result from measurement error and intraspecific variation. We show that failing to account for time-independent variation results in model misspecification (Table 1) and overestimated rate of evolution for the pic method.

639 Having confidence estimates is critical in the presence of inherent uncertainty because they
640 provide direct evaluation of the uncertainty associated with the predictions. Although such
641 uncertainty is positively correlated with the NSTD (Figure 2), use of NSTD as a measure of
642 uncertainty lacks a clear statistical meaning. Using cross-validation, we show that
643 *RasperGade16S* has high precision (around 0.96), which means for predictions with high
644 confidence (≥95%), 96% of the predictions are accurate. Therefore, we can use the confidence
645 score provided by *RasperGade16S* to select high-quality predictions if necessary, or we can draw
646 firm conclusions from the 16S rRNA data when the confidence is high. For example, 16S rRNA
647 GCN has been linked to the ecological strategy of bacterial species, with oligotrophs generally
648 having low GCNs and copiotrophs having higher GCNs [17,18]. To better understand the overall
649 ecological strategy of a bacterial community, we can predict its members' GCNs and classify the
650 community into either an oligotroph-dominant, copiotroph-dominant or a mixed community, and
651 we can do this with a measure of confidence.
652
653 The application of confidence estimation extends beyond the prediction of 16S rRNA GCN.
654 Because the uncertainty in the prediction is inherited by statistics derived from the predicted 16S
655 rRNA GCN, we can estimate the uncertainty and confidence intervals of important parameters in
656 downstream analyses, such as the relative cell abundance. With confidence intervals, we can
657 draw more meaningful and sound conclusions, such as identifying the most abundant OTU in the
658 community with a support value. Getting confidence estimates of the relative cell abundance is
659 also important for predicting the functional profile of a community based on 16S rRNA
660 sequences. Although PICRUST2 uses an extremely lenient NSTD cut-off to eliminate
661 problematic sequences, it does not provide an accurate confidence measurement of its

predictions. As shown in this study, the default maximum parsimony method used by PICRUST2 to predict 16S rRNA GCN essentially assumes there is no uncertainty in the predictions, which is unrealistic and leads to poor precision. Incorporation of a more meaningful confidence estimate of 16S rRNA GCN prediction in PICRUST2 should make its functional profile prediction more informative.

We predicted GCN for 592605 bacterial OTU99 in the SILVA database. Not surprisingly, we observed considerable uncertainty in the GCN predictions. This is because only a small fraction of bacterial diversity in the SILVA database has been captured by the fully sequenced genomes. In addition, 65.1% of OTUs in SILVA database belong to the regularly-evolving group and the confidence of predictions for these OTUs is capped at 85% because of the time-independent variation. However, we would like to point out that natural communities are not a random subsampling of the SILVA OTUs and the median NSTI (NSTD weighted by community members' relative abundance) of the 113842 bacterial communities we examined is 0.01 substitutions/site, much lower than the median NSTD of SILVA OTUs. Strikingly, 99% of 113842 bacterial communities we examined have an adjusted NSTI less than 0.3 substitutions/site, a range where we show that GCN correction improves the accuracy of the relative cell abundance estimation (Figure 3B). Because these communities represent a comprehensive and diverse list of natural and engineered environments, we recommend applying 16S rRNA GCN correction to practically any microbial community regardless of the environmental type if accurate estimates of relative cell abundance are critical to the study. Our results therefore affirm the conclusion of the previous studies based on analyses of a much smaller number of communities [5,7].

685

686 McLaren et al. have shown that statistics that are functions of individual taxon's relative
687 abundance will be sensitive to the systematic biases introduced in the sequencing and data
688 analysis pipeline (including bias introduced by 16S rRNA GCN variation) [44]. Contrary to
689 some popular belief, these biases will not cancel out when analyzing the differences between
690 samples that have been measured by the same protocol [44]. Nevertheless, few studies have
691 investigated to what extent the bias introduced by 16S rRNA GCN variation will have on the
692 microbiome beta diversity analyses. We show that the effect sizes of 16S rRNA bias on beta-
693 diversity analyses are small. Correcting 16S rRNA GCN provides limited improvement on the
694 beta-diversity analyses such as random forest analysis and PERMANOVA test. One possible
695 reason is that for an OTU, the fold change in the relative cell abundance between samples
696 remains more or less the same with or without correcting for the copy number. For example,
697 assuming the estimated relative cell abundances of an OTU in samples A and B are r_a and r_b
698 respectively without copy number correction. When correcting for the copy number, its relative
699 abundance is adjusted with the scaling factor ACN/GCN, where the GCN is the 16S rRNA copy
700 number of the OTU and the ACN is the average copy number of the sample. Assuming the ACN
701 does not vary much between samples, then the scaling factor for the OTU will be roughly the
702 same in samples A and B. So even with copy number correction, the relative abundance change
703 will still be close to r_a/r_b.

704

705 It should be noted that having a confidence associated with the 16S rRNA GCN prediction helps
706 to estimate the uncertainty of the prediction, but it does not improve the accuracy of the
707 prediction. Accuracy of the prediction is constrained by the inherent uncertainty, which can only

be improved by better sampling the reference genomes. However, as our current sampling is inadequate for accurate 16S rRNA GCN prediction of all environmental bacteria, we believe that incorporating confidence estimates is the best practice to control for the uncertainty in the 16S rRNA based bacterial diversity studies, as opposed to not correcting the GCN bias as previously suggested [8,12]. Based on bootstrapping, we have demonstrated that confidence intervals or support values can be calculated for key statistics in downstream analyses such as relative cell abundance and beta diversity. With the uncertainty incorporated into the statistical tests, users may decide if correcting for GCN variation is worthwhile on a case-by-case basis.

Conclusion

We have developed a robust model to estimate the confidence of 16S rRNA GCN predictions. As a rule of thumb, we recommend that, regardless of the environmental type, 16S rRNA GCN correction be applied to virtually all 16S rRNA bacterial communities when estimating their compositional and functional profiles. However, for commonly used bacterial beta-diversity analyses, the GCN correction does not appear to be necessary.

List of abbreviations

16S rRNA: 16S ribosomal RNA. GCN: gene copy number. 16S GCN: 16S rRNA gene copy number. OTU: operational taxonomic unit. BM: Brownian motion. PE: pulsed evolution. PIC: phylogenetically independent contrast. AIC: Akaike information criterion. NSTI: nearest-sequenced-taxon-index. NSTD: nearest-sequenced-taxon-distance. CI: confidence interval. PCoA: principal coordinates analysis. PERMANOVA: permutational multivariate analysis of variance

Declarations

Availability of data and material

The NCBI accession numbers of the reference genomes, the representative 16S rRNA sequences and alignments, the reference phylogeny, the predicted GCN for OTU99 in the SILVA database, and the simulated bacterial community data generated during the current study are available in the Dryad repository (https://datadryad.org/stash/share/OaS9BjM_kIVdJ3WkZRT7KO8fDr8D4k8jy3LsOtlYELM). The R package *RasperGade16S* can be downloaded from https://github.com/wu-lab-uva/RasperGade16S. The original scripts to conduct the analyses in this study are available in the GitHub repository (https://github.com/wu-lab-uva/16S-rRNA-GCN-Predcition).

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

YG developed the R package *RasperGade16S*, conducted statistical analyses in the manuscript and was a major contributor in writing the manuscript. MW conceptualized the rate heterogeneity model and was a major contributor in writing the manuscript. Both authors read and approved the final manuscript.

References

1. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research. 2012;41:D590–6.
2. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology. 2006;72:5069–72.
3. Klappenbach JA. rrndb: the ribosomal RNA operon copy number database. Nucleic Acids Research. 2001;29:181–4.
4. Větrovský T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS ONE. 2013;8:e57923.
5. Kembel SW, Wu M, Eisen JA, Green JL. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Computational Biology. 2012;8:16–8.
6. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. Elife. 2019;8.
7. Angly FE, Dennis PG, Skarshewski A, Vanwonterghem I, Hugenholtz P, Tyson GW. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome. 2014;2:11.
8. Starke R, Pylro VS, Morais DK. 16S rRNA gene copy number normalization does not provide more reliable conclusions in metataxonomic surveys. Microbial Ecology. 2021;81.
9. Bowman JS, Ducklow HW. Microbial communities can be described by metabolic structure: A general framework and application to a seasonally variable, depth-stratified microbial community from the coastal west Antarctic peninsula. PLOS ONE. 2015;10:e0135868.
10. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology. 2013;31:814–21.
11. Zaneveld JRR, Thurber RL V. Hidden state prediction: a modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Frontiers in Microbiology. 2014;5:431.
12. Louca S, Doebeli M, Parfrey LW. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome. 2018;6:41.
13. Ané C. Analysis of comparative data with hierarchical autocorrelation. The Annals of Applied Statistics. 2008;2:1078–102.
14. Landis MJ, Schraiber JG. Pulsed evolution shaped modern vertebrate body sizes. Proceedings of the National Academy of Sciences. 2017;114:13224–9.
15. Joy JB, Liang RH, McCloskey RM, Nguyen T, Poon AFY. Ancestral reconstruction. PLOS Computational Biology. 2016;12:e1004763.

16. Elliot MG, Mooers AØ. Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. BMC Evolutionary Biology. 2014;14:226.
17. Roller BRK, Stoddard SF, Schmidt TM. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nature Microbiology. 2016;1:1–7.
18. Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, et al. The genomic basis of trophic strategy in marine bacteria. Proceedings of the National Academy of Sciences. 2009;106:15527–33.
19. Eldredge N, Gould SJ. Punctuated equilibria - an alternative to phyletic gradualism. Models in Paleobiology. 1972. p. 82–115.
20. Uyeda JC, Hansen TF, Mcpeek A. The million-year wait for macroevolutionary bursts. Proceedings of the National Academy of Sciences. 2011;108:15908–13.
21. Yano K, Masuda K, Akanuma G, Wada T, Matsumoto T, Shiwa Y, et al. Growth and sporulation defects in Bacillus subtilis mutants with a single rrn operon can be suppressed by amplification of the rrn operon. Microbiology. 2016;162:35–45.
22. Rastogi R, Wu M, DasGupta I, Fox GE. Visualization of ribosomal RNA operon copy number distribution. BMC Microbiology. 2009;9:208.
23. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Research. 2015;43.
24. Sadeghifard N, Gu□rtler V, Beer M, Seviour RJ. The mosaic nature of intergenic 16S-23S rRNA spacer regions suggests rRNA operon copy number variation in Clostridium difficile strains. Applied and Environmental Microbiology. 2006;72:7311–23.
25. Lee CM, Sieo CC, Abdullah N, Ho YW. Estimation of 16S rRNA gene copy number in several probiotic Lactobacillus strains isolated from the gastrointestinal tract of chicken. FEMS Microbiology Letters. 2008;287:136–41.
26. Bodilis J, Nsigue-Meilo S, Besaury L, Quillet L. Variable copy number, intra-genomic heterogeneities and lateral transfers of the 16S rRNA gene in Pseudomonas. PLoS ONE. 2012;7:e35647.
27. Lavrinienko A, Jernfors T, Koskimäki JJ, Pirttilä AM, Watts PC. Does intraspecific variation in rDNA copy number affect analysis of microbial communities? Trends in Microbiology. 2021;29:19–27.
28. Viklund J, Ettema TJG, Andersson SGE. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. Molecular Biology and Evolution. 2012;29.
29. Moran NA. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proceedings of the National Academy of Sciences. 1996;93:2873–8.
30. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature. 2009;462.
31. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution. 2013;30.

833 32. Perisin M, Vetter M, Gilbert JA, Bergelson J. 16Stimator: statistical estimation of ribosomal
834 gene copy numbers from draft genome assemblies. The ISME Journal. 2016;10:1020–4.
835 33. Eddy SR. Accelerated profile HMM searches. PLoS Computational Biology.
836 2011;7:e1002195.
837 34. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
838 phylogenies. Bioinformatics. 2014;30:1312–3.
839 35. Torsten Seemann. Barrnap [Internet]. 2018 [cited 2022 Mar 12]. Available from:
840 https://github.com/tseemann/barrnap
841 36. Felsenstein J. Phylogenies and the comparative method. American Naturalist. 1985;125:1–15.
842 37. Gao Y, Wu M. Modeling pulsed evolution and time-independent variation improves the
843 confidence level of ancestral and hidden state predictions. Systematic Biology. 2022;
844 38. Louca S, Doebeli M. Efficient comparative phylogenetics on large trees. Bioinformatics.
845 2018;34:1053–5.
846 39. Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, et al. EPA-ng: Massively
847 parallel evolutionary placement of genetic sequences. Systematic Biology. 2019;68:365–9.
848 40. The Human Microbiome Project Consortium. A framework for human microbiome research.
849 Nature. 2012;486:215–21.
850 41. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur
851 rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017;2.
852 42. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal
853 catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551:457–63.
854 43. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify:
855 the microbiome analysis resource in 2020. Nucleic Acids Research. 2019;48:D570–8.
856 44. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic
857 sequencing experiments. Elife. 2019;8.
858

859 **Figure legends**

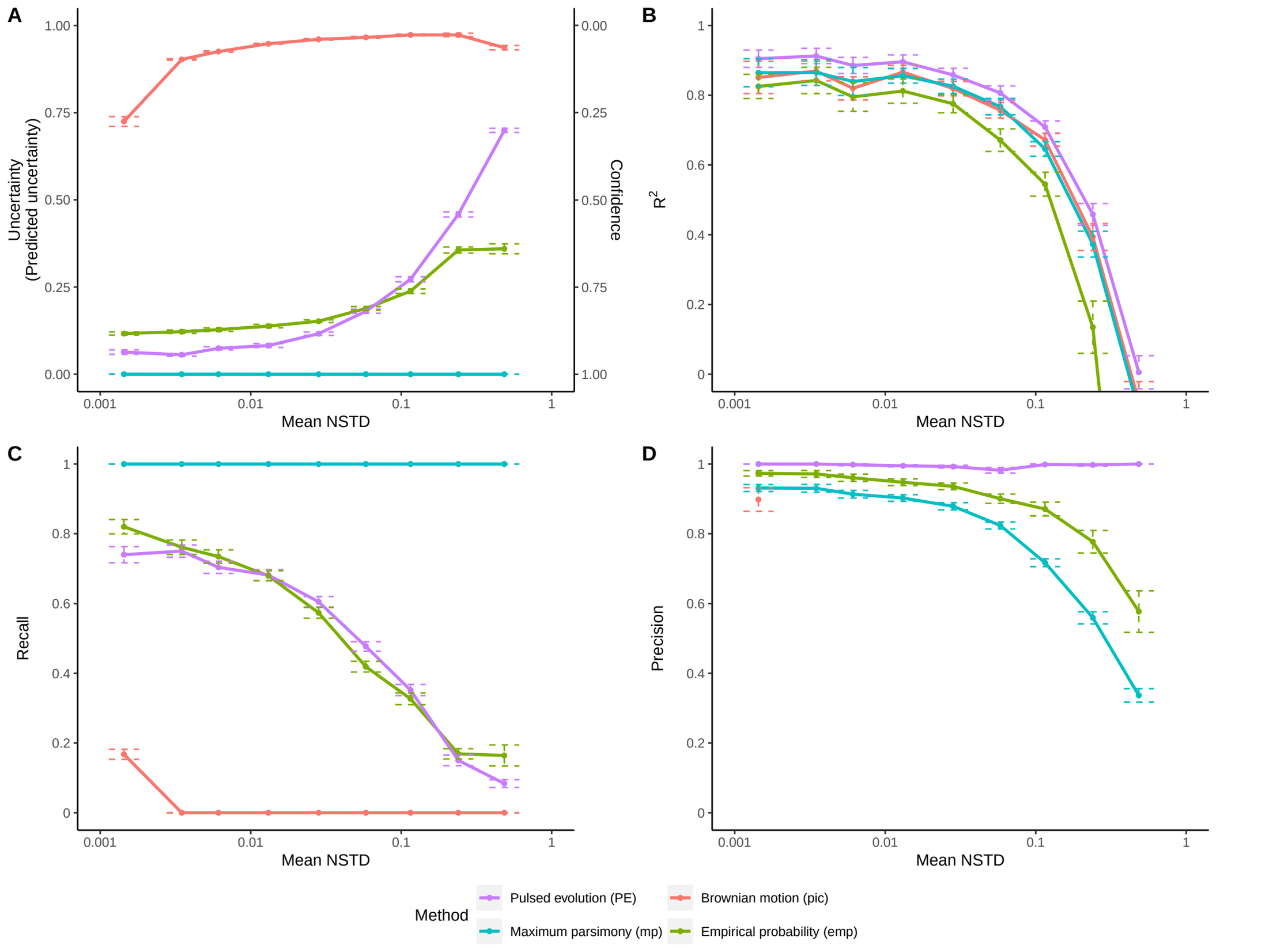
860 **Figure 1. The performance of prediction on simulated 16S rRNA GCN.** Using cross-
861 validation of simulated data, the mean estimated uncertainty and confidence of predictions (A),
862 the mean coefficient of determination R^2 of the predictions (B), and the recall (C) and precision
863 (D) of classification of predictions by their associated confidence estimate, plotted against the
864 mean NSTD. The red line is missing in D because no predictions under the BM model have ≥ 95%
865 confidence when the mean NSTD is greater than 0.002 substitutions/site. The error bars
866 represent the 95% CI of the mean.

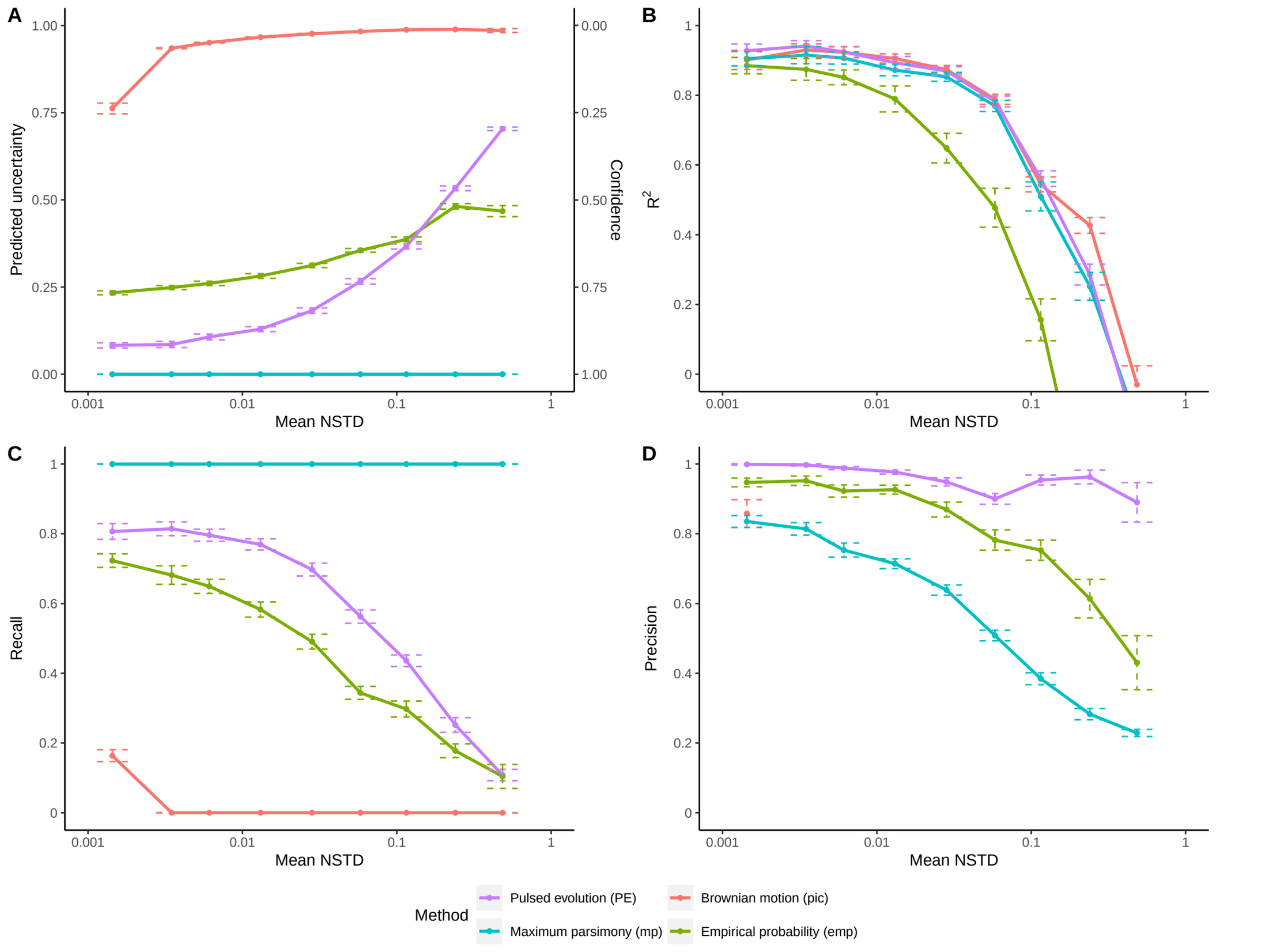
867

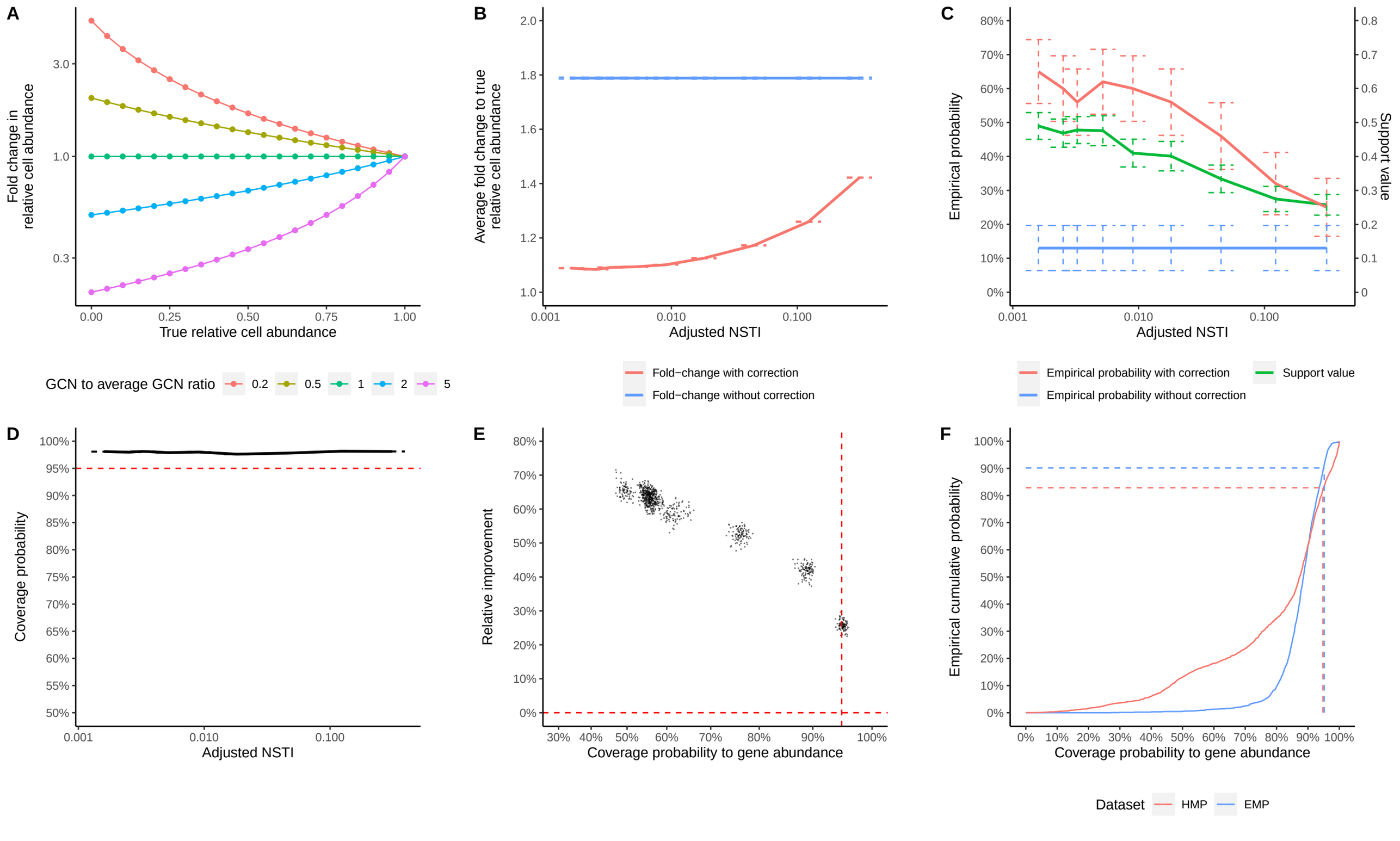
868 **Figure 2. The performance of prediction on empirical 16S rRNA GCN**. Using cross-
869 validation of empirical data, the mean estimated uncertainty and confidence of predictions (A),
870 the mean coefficient of determination R^2 of the predictions (B), and the recall (C) and precision
871 (D) of classification of predictions by their associated confidence estimate, plotted against the
872 mean NSTD. The error bars represent the 95% CI of the mean. The empirical 16S rRNA GCN
873 analyzed here are from the 6408 complete genomes in the reference phylogeny.
874
875 **Figure 3. The impact of 16S rRNA GCN variation on estimated relative cell abundances.** (A)
876 The impact of GCN variation on estimated relative cell abundance based on theoretical
877 calculations. The color of the lines denotes the ratio of an OTU's of GCN to the average GCN of
878 the community. (B) The average fold-change to the true relative cell abundance. (C) The
879 empirical probability of correctly identifying the most abundant OTU in the community and the
880 support value for the most abundant OTU. (D) The coverage probability of relative cell
881 abundances' estimated 95% CIs to the true relative cell abundance. Accurate confidence
882 estimates (95% CIs) should produce a coverage probability of 95% regardless of the adjusted
883 NSTI (dashed red line). (E) The correlation between the coverage probability to the relative gene
884 abundance and the improvement by GCN correction. The horizontal red dashed line represents
885 no improvement in relative cell abundance estimates; the vertical red dashed line represents 95%
886 coverage probability to the relative gene abundance. The improvement is quantified by the
887 relative reduction in the difference between the estimated and true cell abundances. (F) The
888 empirical cumulative distribution of the coverage probability to the relative gene abundance in
889 2560 samples from the HMP1 dataset and 1856 samples from the EMP dataset. All error bars
890 represent 95% CI of the mean.

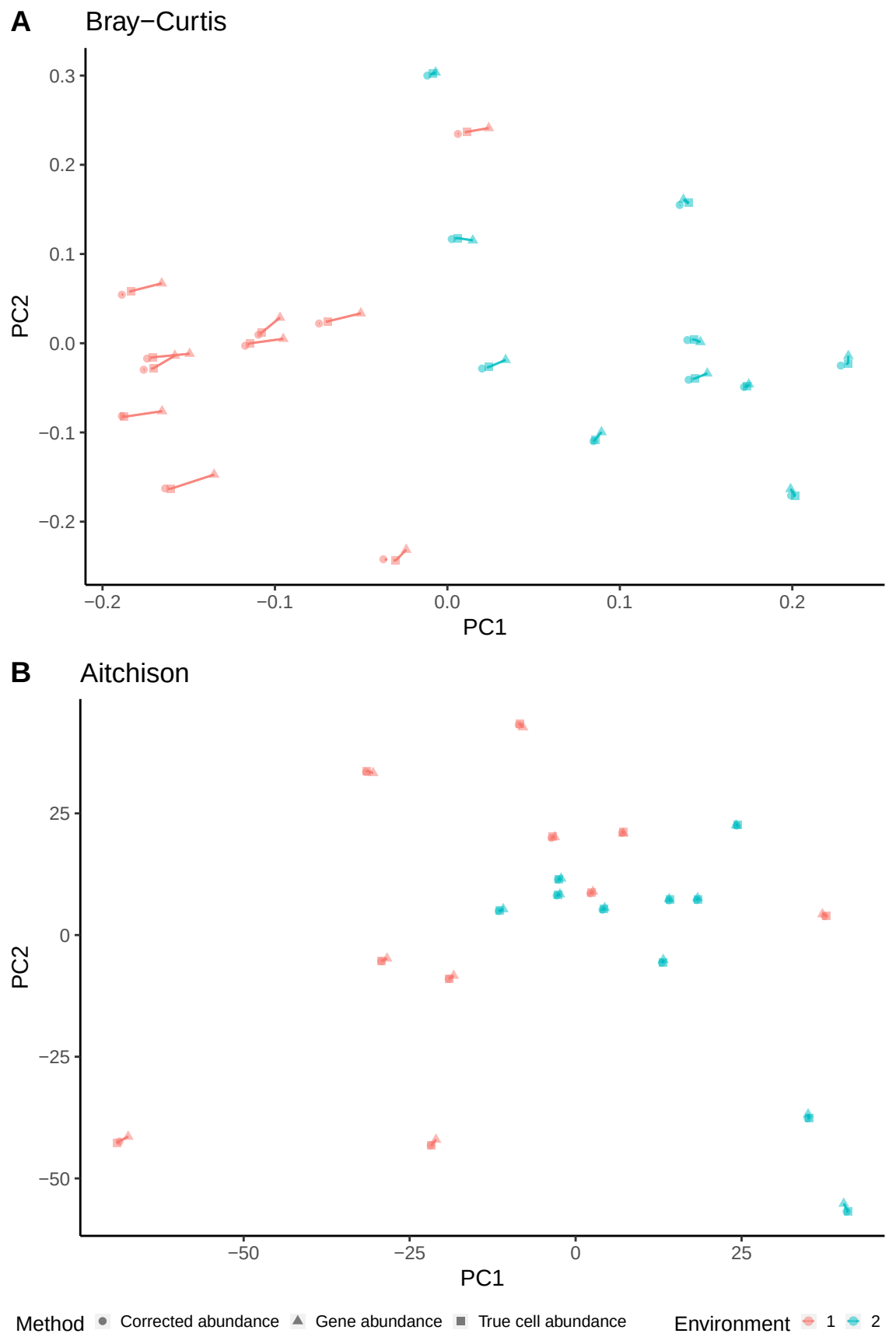
Figure 4. The impact of 16S rRNA GCN variation on beta-diversity. Examples of shift in the Bray-Curtis dissimilarity (A) and the Aitchison distance (B) matrices due to 16S rRNA GCN variation. The shift for each metric is visualized in a PCoA plot comparing 20 simulated samples from two hypothetical environments with 5 signature OTUs (0.25%) in each environment and a turnover rate of 20%. Solid lines represent the shift of a sample from its true location when using the gene abundance.

Figure 5. The distribution of adjusted NSTI in empirical data. The distribution of adjusted NSTI of 113842 communities in the MGnify database representing various environmental types. The red dashed line marks the adjusted NSTI of 0.3 substitutions/site.


A
B
C
D
Uncertainty (Predicted uncertainty)
Confidence
Mean NSTD
R2
Recall
Precision
Method
Pulsed evolution (PE)
Brownian motion (pic)
Maximum parsimony (mp)
Empirical probability (emp)


A
B
C
D
Predicted uncertainty
Confidence
Mean NSTD
R^2
Recall
Precision
0.001
0.01
0.1
1
0.00
0.25
0.50
0.75
1.00
0
0.2
0.4
0.6
0.8
Method
Pulsed evolution (PE)
Brownian motion (pic)
Maximum parsimony (mp)
Empirical probability (emp)


A
Fold change in relative cell abundance
3.0
1.0
0.3
0.00
0.25
0.50
0.75
1.00
True relative cell abundance
GCN to average GCN ratio
0.2
0.5
1
2
5
B
Average fold change to true relative cell abundance
2.0
1.8
1.6
1.4
1.2
1.0
0.001
0.010
0.100
Adjusted NSTI
Fold−change with correction
Fold−change without correction
C
Empirical probability
80%
70%
60%
50%
40%
30%
20%
10%
0%
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Support value
0.001
0.010
0.100
Adjusted NSTI
Empirical probability with correction
Empirical probability without correction
Support value
D
Coverage probability
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
0.001
0.010
0.100
Adjusted NSTI
E
Relative improvement
80%
70%
60%
50%
40%
30%
20%
10%
0%
30%
40%
50%
60%
70%
80%
90%
100%
Coverage probability to gene abundance
F
Empirical cumulative probability
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Coverage probability to gene abundance
Dataset
HMP
EMP


A
Bray−Curtis
0.3
0.2
0.1
0.0
−0.1
−0.2
PC2
−0.2
−0.1
0.0
0.1
0.2
PC1
B
Aitchison
25
0
−25
−50
PC2
−50
−25
0
25
PC1
Method
Corrected abundance
Gene abundance
True cell abundance
Environment
1
2

Engineered:Bioreactor:Unassigned
Engineered:Built environment:Unassigned
Engineered:Food production:Fermented beverages
Engineered:Food production:Fermented vegetables
Engineered:Lab enrichment:Unassigned
Engineered:Lab Synthesis:Unassigned
Engineered:Modeled:Simulated communities (microbial mixture)
Engineered:Wastewater:Water and sludge
Environmental:Aquatic:Freshwater
Environmental:Aquatic:Marine
Environmental:Aquatic:Sediment
Environmental:Terrestrial:Deep subsurface
Environmental:Terrestrial:Soil
Host−associated:Animal:Digestive system
Host−associated:Birds:Digestive system
Host−associated:Birds:Unassigned
Host−associated:Fish:Digestive system
Host−associated:Human:Digestive system
Host−associated:Human:Reproductive system
Host−associated:Human:Respiratory system
Host−associated:Human:Skin
Host−associated:Human:Unassigned
Host−associated:Insecta:Digestive system
Host−associated:Insecta:Unassigned
Host−associated:Mammals:Digestive system
Host−associated:Mammals:Gastrointestinal tract
Host−associated:Mammals:Skin
Host−associated:Mammals:Unassigned
Host−associated:Plants:Phylloplane
Host−associated:Plants:Rhizoplane
Host−associated:Plants:Rhizosphere
Host−associated:Plants:Root
Host−associated:Plants:Unassigned
Host−associated:Unassigned
Mixed:Unassigned

Count

Adjusted NSTI

Engineered
Environmental
Host−associated
Mixed