TaxaSE: Exploiting evolutionary conservation within 16S rDNA sequences for enhanced taxonomic annotation

Ali Z. Ijaz[1], Thomas Jeffries[1], Christopher Quince[2], Kelly Hamonts[1], Brajesh K. Singh[1*]

1. Hawkesbury Institute for the Environment, Western Sydney University, Penrith, NSW, Australia

2. Warwick Medical School, University of Warwick, Coventry, United Kingdom

*Corresponding Author: b.singh@westernsydney.edu.au

Abstract

Amplicon based taxonomic analysis, which determines the presence of microbial taxa in different environments on the basis of marker gene annotations, often uses percentage identity as the main metric to determine sequence similarity against databases. These data are then used to study the distribution of biodiversity as well as response of microbial communities to environmental conditions. However the 16S rRNA gene displays varying degrees of sequence conservation along its length and percentage identity does not fully utilize this information. Additionally, the prevalent usage of Operational Taxonomic Unit, or OTUs is not without its own issues and may lead to a reduction in annotation capability of the system. Hence a novel approach to taxonomic annotation is needed. Here we introduce a new taxonomic annotation pipeline, TaxaSE, which utilizes Shannon entropy to quantify evolutionary conservation within 16S rDNA sequences for enhanced taxonomic annotations. Furthermore, the system is capable of annotation of individual sequences in order to improve fine grain taxonomic annotations. We present both *in-silico* comparison of the new similarity metric with percentage identity, as well as comparison with the popular QIIME pipeline. The results demonstrate the new similarity metric achieves better performance especially at lower taxa levels. Furthermore, the pipeline is able to extract more fine grain taxonomic annotations compared to QIIME. These exhibit not only the effectiveness of the new pipeline but also highlight the need to shift away from both percentage identity and OTU based approaches for ecological projects.

Introduction

Ecogenomics study of microbes is a rapidly growing field of research that aims at studying uncultured organisms via their nucleic acid sequences to determine the true

diversity of microbes, their function and distribution in a variety of environments (Huson et al. 2009). Many environments have been the focus of ecogenomics studies, including soil, the oral cavity, feces, and aquatic habitats (Riesenfeld et al. 2004). The field has been driven by the advent of high throughput sequencing where genomic information is acquired directly from the microbial communities in their natural environment, with a drastic reduction in the cost of sequencing (Morgan & Huttenhower 2014). As a consequence, bioinformatics pipelines aiming to characterize microbial community composition, have been developed alongside various 16S rDNA gene sequence databases, which serve as a reference set of sequences for microbial taxonomic analysis (Santamaria et al. 2012).

Sequencing of 16S rDNA amplicons primarily uses short reads, representing a specific region of a gene. Analysis requires a significant amount of time, typically a day or more for taxonomic annotation depending on computational resources and size of data. The underlying scoring scheme behind sequence similarity is currently percentage identity, a simple distance based approach which does not fully utilize the inherent variation in evolutionary conservation within 16S rDNA gene sequences, as every base is considered equal with respect to matches and mismatches and positions of these matches and mismatches are not essential (Fox et al. 1992; Stackebrandt & Goebel 1994). This is important in the context that certain regions of the 16S rDNA sequences are considerably variable while others are relatively conserved, and the degree of variability is not constant (Chakravorty et al. 2007; Stackebrandt & Goebel 1994). This distance based approach does not truly estimates the evolutionary distances between sequences as different nucleotide positions on sequences are changing at different rates (Woese 1987). Furthermore, fine-scale taxonomic annotation may not be resolved as well, especially at genus level (Fox et al. 1992). As

most taxonomic annotation pipelines, such as QIIME (Caporaso et al. 2010), MG-RAST (Aziz et al. 2008) and MEGAN (Huson et al. 2007) are dependent on percentage identity for sequence similarity measure, an improvement in this context would result in better downstream analysis. These represent the limitations of 16S rDNA gene sequence analysis primarily due to the selection of percentage identity as the determinant of sequence similarity.

Furthermore, the majority of taxonomic annotation systems use operational taxonomic unit (OTU), as the defining concept for determining community composition (He et al. 2015). Considered as a *de facto* standard approach to analysis, OTUs are formed by clustering sequences on the basis of a specified similarity threshold such as 97% (Drancourt et al. 2000; Tikhonov et al. 2015). Sequence based denoising approaches such as DADA2 (Callahan et al. 2016) and Deblur are also applied. Taxonomic annotation is performed on the representative sequence of each OTU, and all the sequences within the OTU are assigned the same taxonomy regardless of small-scale differences in base composition between them (Nguyen et al. 2016). This is a favorable technique as picking representative OTUs from a list of sequences drastically cuts down on computational requirements for analysis, giving the ability to quickly perform fast annotation, in addition to providing abundance information of how many reads form a cluster (He et al. 2015; Methé et al. 2012) and, therefore, allows for rapid analysis of large datasets (Nguyen et al. 2016).

However, OTU generation methods assume that all 16S rDNA genes evolve at the same rate (Schloss & Westcott 2011). Furthermore, OTUs made from short read sequences may not be as reliable in estimating species richness as the OTUs formed from near full-length sequences, primarily due to the 16S rRNA gene exhibiting different degrees of variability across its length and therefore region selection plays

an important role in accurately estimating microbial diversity (Kim et al. 2011). Additionally, OTU assignments may not be reliable and can differ on the basis of the algorithm used (Tikhonov et al. 2015), with common OTU creation approaches sometimes leading to inflation of species level diversity estimates (Edgar 2013; White et al. 2010). This is compounded by the fact that certain OTU construction techniques generate unstable OTUs where the membership of sequences changes significantly with the addition of new sequences or samples to the dataset and as a consequence, different sets of OTUs are observed with each clustering run (He et al. 2015). This has a significant impact on downstream diversity analysis including rarefaction curves, which determine how well sequencing depth captures diversity as well as identification of individual OTUs (He et al. 2015; Nguyen et al. 2016).

Our aim was to address these limitations by developing a new taxonomic annotation pipeline, defined here as Taxonomic Annotation *via* Shannon entropy (the TaxaSE system), which employs the novel Shannon entropy based sequence similarity measure, instead of percentage identity, to quantitatively assess variability across the whole of the 16S rDNA sequences within an aligned bacteria database, paving the way for a novel approach towards estimating sequence similarity and compared its performance against the most widely used QIIME pipeline (Caporaso et al. 2010). In fact, it was proposed determining the pattern of change at given positions in 16S rRNA gene may optimise analysis (Woese 1987). The technique has been utilized in other tools such as oligotyping, which looks at the variation within an individual OTU (Eren et al. 2013). Furthermore, the limitations associated with OTU generation and usages were resolved by following an OTU-independent approach where sequences are annotated individually. This resulted in the highest resolution annotation via a combination of an improved annotation algorithm as well as

extracting intra-OTU diversity, compared to the standard 97% OTU similarity approach, which obscures fine-scale variation. With the improvements in computational resources available to ecological projects, this approach is now practical to be used in determining microbial diversity.

To illustrate the effectiveness of our pipeline, *in-silico* comparison was performed between the underlying Shannon entropy based metric of the new pipeline against the percentage identity metric, to demonstrate the improvement in sequence similarity determination, while the pipeline itself was compared to QIIME on datasets from sugarcane habitat for both alpha diversity and beta diversity evaluation of the microbial community.

Materials & Methods

Shannon Entropy based sequence similarity scoring metric

SILVA (Quast et al. 2013) Release 123 aligned database of 16S rDNA sequences was used to quantitatively assess and calculate entropy across the whole 16S rDNA sequence. The database was taken as a matrix **M** of dimensions **m x n**, consisting of **m** rows and **n** columns. Each row was an aligned reference sequence and column denoted locations where a nucleotide, gap or dot occurred. As the database represented multiple sequence alignments of 16S rRNA, dots were used for padding before the start and after the end of a reference sequence depending on how the sequence was aligned against other sequences and therefore were not factored in any calculation, as they did not signify any information. To simplify calculations, ambiguous sequences that contained nucleotides other than A, T, C or G such as N were removed from the database. Shannon entropy was then calculated for every

column in the database, as given in pseudo code listed in supplementary material 1. USEARCH sequence aligner (Edgar 2010) was utilized for determining alignments between reference and query sequences. The system flowchart is illustrated in Figure 1, where USEARCH alignments (Edgar 2010) were used to reconstruct full alignments between query sequences and reference 16S rDNA gene sequences. This determined precisely where matches, mismatches and gaps occurred against a reference sequence. Relative entropy was then calculated using the vectors developed for each reference sequence and finally each query read was scored. The process is described as below:

1) Query sequences were aligned with the reference SILVA database. The resultant data contained complete information of alignment between the reference and query sequences as well as the location of alignments.
2) Alignments were then reconstructed where location of gaps, matches and mismatches were determined.
3) Shannon entropy for each query sequence and the matched reference sequence segment was calculated using the stored vectors in a separate database.
4) Finally, relative Shannon entropy score was calculated and query sequences were annotated with reference sequence taxonomic annotation.

Relative Shannon entropy for every query sequence was generated in the following manner:

1) Shannon entropy value on locations where a nucleotide mismatch occurred between the reference and query sequence was converted to a negative value for query sequence.

2) Next, for both reference sequence and query sequence, the maximum Shannon entropy value was added on each location. This enabled better segregation of sequences, which may contain mismatches.
3) Finally, the total entropy value for both reference sequence segment as well as query sequence was calculated by adding values at every location.
4) A relative entropy score was then calculated by dividing total Shannon entropy value of a query read by the total Shannon entropy value of the reference read segment. As every reference sequence had a taxonomic annotation associated with it, the matched input read was assigned this annotation.

Validation of Shannon entropy based scoring metric

Validation of the new scoring scheme was performed using an *in silico* approach. *MicroSim*: A motif-based next-generation read simulator developed by Schirmer *et. al.* was used to generate multiple datasets of 20,000 amplicon reads from reference sequences from SILVA release 123 database, simulating an Illumina MiSEQ Fusion Golay V4 Amplicon 250bp (DS78) platform. The following metrics were used in the validation process:

Recall: $\frac{TP}{TP+FN}$

Precision: $\frac{TP}{TP+FP}$

Accuracy: $\frac{TP+TN}{TP+FP+TN+FN}$

Here, TP denotes True Positives, FP as False Positives, TN as True Negatives, and FN as False Negatives. Thresholds were varied between 0 and 1 to determine recall, precision and accuracy for both percentage identity and the new Shannon entropy based scoring scheme. Lastly, for precision vs. recall curves, area under the curve was also calculated to determine if the new scoring metric is performing better than percentage identity.

The validation process consisted of removal of taxa approach, where 100 genera, 10 families and 1 class were randomly selected and removed. Sequences belonging to these removed taxa are effectively novel to the remaining sequences in the database and therefore should not closely match any of the taxa retained in the database. This approach can be useful in understanding how the system reacts to novel sequences that may present themselves in real datasets to which the database is naïve (Lanzen et al. 2012). Furthermore, application of MicroSim on these sequences ensured that the resultant mock community to be tested, would be much more representative of real datasets as compared to random cropping of sequences.

Real dataset analysis

For the real dataset analysis between TaxaSE and QIIME, samples from sugarcane environment were selected to elucidate the differences between both pipelines. Sugarcane leaf, stalk, root and rhizosphere soil samples were collected in November 2014 from eight sugarcane fields growing three sugarcane varieties (KQ228, MQ239 and Q240) near Ingham, Queensland, Australia. Bacterial 16S rRNA amplicon

sequencing was performed by the NGS facility at Western Sydney University using Illumina Miseq (2x 301 bp PE) and the 341F/805R primer set.

A total of 158 samples were used, with the breakdown from each sub-habitat listed in Table 1. To minimize noise artifacts and prevent occurrences of chimeras, the following preprocessing procedure was followed for all samples:

1) Read trimming:
    a. Sequences were trimmed on both R1 and R2 reads removing low quality regions with Phred (Ewing et al. 1998) score of less than 25 (Q25). This was performed using "seqtk" tool (Li).
2) Paired-end read merging:
    a. After quality trimming, both forward and reverse reads were merged using FLASH (Magoc & Salzberg 2011) with a maximum overlap set to 200.
3) Chimera removal:
    a. Finally, the merged reads were analyzed for the presence of chimeras. This was accomplished using VSEARCH, a sequence aligner and RDP (Cole et al. 2014) Gold database which contained 10,049 reference sequences. Subsequently, chimeras were removed from the samples.

Given that the new pipeline was developed to annotate on a per-sequence basis, comparison was based on the distinct number of annotations observed by each pipeline. OTUs were generated at 97% and 99% sequence similarity for QIIME. Following the annotation process via RDP classifier, OTUs, which had the same taxonomic annotations, were combined together to form pseudo-OTUs. Furthermore, OTUs belonging to Eukaryota and Archaea were removed from QIIME results as the

primary comparison between both systems was based on bacterial taxonomic annotations. Lastly, given that the new pipeline was using a completely new sequence similarity-scoring scheme, hence a new set of thresholds was selected. Primarily, three comparison approaches were followed and analysis were done *via* tools provided in QIIME:

- Alpha diversity comparison:
    - Implemented using QIIME's inbuilt *alpha_rarefaction.py* script
    - Distinct number of taxonomic annotations
    - Shannon diversity
- Beta diversity comparison:
    - Accomplished by using QIIME's *beta_diversity_through_plots.py* script. Bray Curtis was taken as the distance metric and plots were generated using the Emperor package (Yoshiki Vázquez-Baeza 2013).
- ADONIS and ANOSIM
    - *compare_categories.py* script was used for this purpose.

Results

Scoring metric comparison

The precision vs. recall curve of both Shannon entropy and percentage identity approaches closely match each other for the removal of genera based dataset (Figure 2-a). Precision started at less than 0.5, diminishing as recall improved for both approaches. For removal of families based validation, the precision vs. recall curve for Shannon entropy stayed above the precision vs. recall curve for percentage identity, illustrating better precision at the same recall (Figure 2-b). Precision for both

curves began at 0.4 and stayed below this until full recall was achieved. Finally, the precision vs. recall curves for removal of class-based validation approach is shown in Figure 2-c. Precision was low for both approaches, staying below 0.4.

The area under the curve illustrates the differences between the classification capabilities of both scoring metrics (Table 2). The new scoring scheme performs better at removal of families and class based datasets, while showing comparable performance to percentage identity for removal of genera.

Pipeline comparison

Alpha Diversity

Distinct number of taxonomic annotations comparison

For rhizosphere environment, TaxaSE produced the highest number of distinct taxonomic annotations at 807, while QIIME at 99% OTU similarity produced 578 distinct taxonomic annotations and QIIME at 97% OTU similarity coming up last at about 515 (Figure 3-a). Welch's t-test showed a very significant difference between QIIME at 97% OTU similarity and QIIME at 99% OTU similarity ($p=0.0059$). Furthermore, Welch's t-test also reported statistically very significant difference between QIIME at 97% OTU similarity and TaxaSE ($p=0.0001$) as well as between QIIME at 99% OTU similarity and TaxaSE ($p=0.0001$). All three approaches were therefore statistically different from each other, with the highest OTUs for TaxaSE pipeline.

For the root environment, here as well TaxaSE produced the largest number of distinct taxonomic annotations at 890, followed by QIIME at 99% OTU similarity with 593 distinct annotations and lastly QIIME at 97% OTU similarity at 522 (Figure

3-b). Welch's t-test illustrated a similar picture here as well, with a statistically significant difference between QIIME at 97% OTU similarity and QIIME at 99% OTU similarity (p=0.0018), a statistically very significant difference between QIIME at 97% OTU similarity and TaxaSE (p=0.0001) and lastly an extremely statistically significant between QIIME at 99% OTU similarity and TaxaSE as well (p=0.0001).

Soil showed similar pattern as with previous environments, with TaxaSE generating higher number of distinct taxonomic annotations reaching 907, while QIIME at 99% OTU similarity followed it at 697 annotations and QIIME at 97% OTU similarity coming up last at 574 distinct annotations (Figure 3-c). A very statistically significant difference was observed *via* Welch's t-test between QIIME at 97% OTU similarity and QIIME at 99% OTU similarity (p=0.003). An extremely statistically significant difference was observed between QIIME at 97% OTU similarity and TaxaSE (p=0.0001) as well as between QIIME at 99% OTU similarity and TaxaSE (p=0.0001).

Stem was the least diverse of all habitats, and TaxaSE generated a highest number of distinct taxonomic annotations at 167 (Figure 3-d). QIIME at 99% OTU similarity generated 121 distinct annotations while QIIME at 97% OTU similarity produced 101 distinct annotations. The difference was not statistically significant, as found by Welch's t-test between QIIME at 97% OTU similarityand QIIME at 99% OTU similarity (p=0.1742). However, statistically significant difference was found between QIIME at 97% OTU similarity and TaxaSE (p=0.0017), as well as between QIIME at 99% OTU similarity and TaxaSE (p=0.0311).

Shannon diversity index comparison

Shannon diversity index comparison displayed a similar picture as illustrated for distinct taxonomic annotation results. For rhizsophere samples, TaxaSE produced the highest Shannon diversity index for distinct taxonomic annotation based comparison, with a value of 7.7, compared to QIIME at 99% OTU similarity at 7.1 and QIIME at 97% OTU similarity at 6.9, as shown in Figure 4-a. Welch's t-test produced a statistically significant difference between QIIME at 97% OTU and QIIME at 99% OTU similarity ($p = 0.045$). The difference was statistically very significant between both QIIME approaches and TaxaSE ($p = 0.0001$).

Samples from root environment showed similar Shannon diversity index results between the two QIIME methods (Figure 4-b), with TaxaSE leading with more than 7.6, followed by QIIME at 99% OTU similarity with 6.8 and lastly QIIME at 97% OTU similarity at 6.6. The difference was not statistically significant between QIIME at 97% OTU similarity and QIIME at 99% OTU similarity ($p = 0.1639$). However, similar to rhizosphere samples, the difference was statistically very significant between both QIIME approaches and TaxaSE ($p = 0.0001$).

TaxaSE also had higher Shannon diversity results for soil samples compared to QIIME at 97% and QIIME at 99% (Figure 4-c), where TaxaSE showed slightly more diversity index at 7.77 than both QIIME methods, with QIIME at 97% OTU similarity at 7.1 and QIIME at 99% OTU similarity at 7.3. Welch's t-test illustrated that the difference was not statistically significant between QIIME at 97% OTU and QIIME at 99% OTU similarity ($p = 0.0565$). However, the difference was statistically very significant between TaxaSE and both QIIME approaches ($p = 0.0001$).

Finally, Shannon diversity index results for all three methods for stem samples showed TaxaSE having an average Shannon diversity of 2.7 while QIIME at 99% OTU similarity produced 2.4 and finally QIIME at 97% OTU similarity produced the

lowest Shannon diversity at 1.7 (Figure 4-d). The difference was statistically significant very between QIIME at 97% OTU similarity and QIIME at 99% OTU similarity and also between QIIME at 97% OTU similarity and TaxaSE ($p = 0.0001$). However, the difference was not statistically significant between QIIME at 99% OTU and TaxaSE ($p = 0.0591$).

Beta Diversity comparison

The beta diversity plots were almost identical across all three approaches and illustrated the same separation pattern of samples. The beta diversity plot for QIIME at 97% OTU similarity is shown in Figure 5-a. Stem samples were segregated from the samples belonging to other environments. Furthermore, root and soil samples displayed some segregation as well. The first principle coordinate, PC1 explained a variance of 58.31% in the case of QIIME at 97% OTU similarity.

Beta diversity plot for QIIME at 99% OTU similarity, as illustrated in Figure 5-b, provided a similar pattern as was seen for QIIME at 97% OTU similarity (Figure 5-a). Stem samples were segregated from the other samples and the first principle coordinate explained a variance of 57%, slightly lower than what was observed for QIIME at 97% OTU similarity.

Finally, the beta diversity plot for TaxaSE system is shown in Figure 7-c and here as well, stem samples were well segregated from other samples. Furthermore, soil samples were more densely packed along the first axis for TaxaSE system compared to either of QIIME based methods. The first principle coordinate axis, PC1 explained 53.22% of variance, the lowest between all three methods.

ADONIS results for the three methods as listed in Table 3 show a slightly different pattern, where the grouping of samples on the basis of environment was best explained by QIIME at 97% OTU similarity with a R^2 value of 0.6797, followed by QIIME at 99% OTU similarity with a R^2 value of 0.671 and lastly TaxaSE, with a R^2 value of 0.622. Overall, the ADONIS results were similar between all three methods.

The ANOSIM results illustrated that for all of the methods, the grouping of samples by environments is statistically significant, with p-value of 0.001 (Table 4). All three methods generated an R-value of more than 0.8, however TaxaSE produced a slightly lower, but still strong ANOSIM result compared to the other two methods.

Discussion

Shannon entropy based sequence similarity metric

The new Shannon entropy based sequence similarity metric can be used as a replacement of the current standard percentage identity. The new approach showed comparative performance for the whole SILVA dataset and slightly lower for removal of genus validation dataset. However it improved upon percentage identity for removal of families and classes datasets.

For removal of genus dataset, sequences were checked at family level. Both approaches generated almost the exact same result in this case, with percentage identity slightly leading over Shannon entropy approach. However, the Shannon entropy based approach showed improved performance compared to Percentage Identity based approach, with higher area under the curve in the case of removal of families dataset. For removal of class dataset, sequences were checked at phylum

level and while both approaches were similar in their capability, Shannon entropy based approach demonstrates slightly improved performance.

This translates into better annotation of novel sequences at the order level as well as phylum level compared to the percentage Identity based approach and is therefore much more effective at taxonomic annotation as novel sequences can be annotated better in the case of the new approach.

Unlike percentage identity, the new Shannon entropy based approach effectively captures evolutionary conservation from the 16S rDNA sequences as every location's degree of variability is directly determined and used in the new scoring scheme. This represents an advance towards better similarity measurements, which are in accordance with the evolution of sequences (Woese 1987). The results illustrate better annotation capability at class and families level while being comparative to percentage identity at other taxa levels.

Given that the vast majority of microbes are uncultivated (Huson et al. 2007; Marcy et al. 2007), there is a higher likelihood that in many ecological studies unknown sequences will be detected. The best possible annotation of these sequences will give insight into the inner workings of the environment, even if the exact taxonomic annotation cannot be determined at finer taxonomic levels (Huson et al. 2007). For this reason, new approaches should be able to handle these sequences in an improved fashion and here the new Shannon entropy based approach provides improved performance over the industry standard Percentage Identity.

TaxaSE performance evaluation

TaxaSE represents an advancement in taxonomic annotation compared to current approaches, with the utilization of a more evolutionary correct sequence similarity

measure and its application in a microbial taxonomic annotation pipeline. Given that the true number of species is unknown for a real dataset, a comparison cannot be made solely on the basis of number of species identified. Nonetheless, the real potential of the pipeline is illustrated when an OTU independent, per sequence annotation is performed. Given that TaxaSE produced better or similar patterns with respect to alpha diversity results, the new pipeline is as applicable as other pipelines in assessing alpha diversity in ecological studies.

The microbial community was observed to be more diverse in the case of soil, rhizosphere and root habitats, which are expected to have a high degree of diversity (Kirk et al. 2004; Pinton et al. 2001). However samples from the stem environment were far less diverse. This was primarily due to different species inhabiting plant stem, which may include endophytic microbes that are beneficial to the growth (Gouda et al. 2016) and health of the plant (Miguel et al. 2016) as well as pathogenic bacteria, however a single plant species may play as a host for only a limited number of microbes (Imam et al. 2016). Furthermore, the niche endophyte population is dependent on various factors such as host species and environmental conditions (Gouda et al. 2016).

As for beta diversity analysis, ADONIS results showed that QIIME at 97% OTU similarity explained the most variance, followed closely by QIIME at 99% OTU similarity, with TaxaSE explaining the least. The results correlate inversely with the number of distinct taxonomic annotations, where QIIME at 97% OTU similarity produced the least number of distinct annotations and explained the most variance and TaxaSE system produced the most number of distinct annotations but with low explanation of variance. Therefore, given that the ADONIS test described how much variation is explained by grouping on the basis of location, less variation is being

explained by approaches with a higher number of taxonomic annotations. This may be because some taxonomic annotations were common across different habitats and approaches such as QIIME at 99% and TaxaSE were able to extract these annotations more in comparison to QIIME at 97%. Beta Diversity plots illustrated similar patterns across all approaches, where QIIME at 97% OTU similarity, QIIME at 99% OTU similarity and TaxaSE, displayed almost identical patterns and were able to differentiate between different habitats. Furthermore, similar to OTU comparison, here as well stem samples were distinctly separated from root, soil and rhizsophere for all three methods. Thus TaxaSE is well suited to identifying ecologically distinct microbial assemblages. In the case of TaxaSE, slightly less variability was accounted by the first axis, PC1 compared to QIIME at 97% OTU similarity and 99% OTU similarity. This may be because more common taxa were observed for TaxaSE system and therefore the ability of the system to explain variability on the basis of taxonomy fell as an increase in the number of variables leads to a reduction in the total variation explained (Nagelkerke 1991). A similar case was observed between QIIME at 97% OTU similarity and QIIME at 99% OTU similarity as the later's first axis explained slightly less variability at 57%, compared to former's 58.31%.

Conclusion

The novel Shannon entropy based approach demonstrated its effectiveness over percentage identity, where the evolutionary conservation information of 16S rRNA is directly exploited to provide a more accurate sequence similarity metric. Most popular approaches forgo the utilization of this inherent information contained within the 16S rRNA sequences, instead relying on a measure that only counts mismatches between sequences. Given the variability across the whole of 16S rRNA, not every

base may be equally important as variable locations are much more essential in differentiating between sequences compared to conserved regions (Chakravorty et al. 2007).

The approach is competitive that it can be used alongside commonly applied percentage identity scoring schemes. Its higher performance at higher taxa levels is especially important as majority of bacterial sequences are not annotated, and more and more novel sequences are being detected in almost all of the next-generation sequencing projects. It's likely that these new sequences may not be resolved at genera level and hence new approaches, which are better at taxonomic annotation at higher taxonomic levels than genera, would be more appropriate.

Building upon this novel approach to sequence similarity is the new TaxaSE pipeline. The OTU independent approach, central to TaxaSE, provides an alternative method to improving taxonomic annotation. While this comes at the expense of more computational time and requirement of higher resources, it can be used to delve deeply into finer level of taxa levels and improve annotation process as a result, which would otherwise go unnoticed with an OTU based method. Alpha diversity results also illustrate a similar picture where TaxaSE generated the highest number of annotations across all habitats in comparison to QIIME based methods. This highlights the benefit of following this new approach.

The results of applied environmental dataset analysis demonstrate the advantage of using TaxaSE over OTU based, industry standard pipelines such as QIIME while demonstrating comparable performance in distinct taxonomic annotation based approach. With the ability to annotate sequences at the highest resolution (e.g. species level) annotation at times as well as using a novel scoring

approach based on Shannon entropy, TaxaSE represents a step forward in taxonomic annotation of microbial DNA sequences.

Author contributions

Ali Z. Ijaz: Developed the TaxaSE pipeline and the underlying Shannon entropy based sequence similarity measure. Performed validation, real dataset analysis and comparison with QIIME pipeline. Wrote the majority of the manuscript.

Thomas Jeffries: Provided evaluation on real dataset analysis, comparison with QIIME. Also provided feedback on the manuscript.

Christopher Quince: Provided feedback and evaluation on the validation process, real dataset analysis and comparison with QIIME. Also provided feedback on the manuscript.

Kelly Hamonts: Performed sampling and sequencing of the sugarcane dataset.

Brajesh K. Singh: Supervised the overall project. Provided critical feedback and comments on the manuscript. Gave approval for the submission of the article.

References

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C,

537 Stevens R, Vassieva O, Vonstein V, Wilke A, and Zagnitko O. 2008. The
538 RAST Server: rapid annotations using subsystems technology. *BMC*
539 *Genomics* 9:75. 10.1186/1471-2164-9-75
540 Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, and Holmes SP. 2016.
541 DADA2: High resolution sample inference from Illumina amplicon data.
542 *Nature Methods* 13:581-583. 10.1038/nmeth.3869
543 Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK,
544 Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights
545 D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M,
546 Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko
547 T, Zaneveld J, and Knight R. 2010. QIIME allows analysis of high-
548 throughput community sequencing data. *Nat Methods* 7:335-336.
549 10.1038/nmeth.f.303
550 Chakravorty S, Helb D, Burday M, Connell N, and Alland D. 2007. A detailed
551 analysis of 16S ribosomal RNA gene segments for the diagnosis of
552 pathogenic bacteria. *J Microbiol Methods* 69:330-339.
553 10.1016/j.mimet.2007.02.005
554 Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A,
555 Kuske CR, and Tiedje JM. 2014. Ribosomal Database Project: data and
556 tools for high throughput rRNA analysis. *Nucleic Acids Res* 42:D633-642.
557 10.1093/nar/gkt1244
558 Drancourt M, Bollet C, Carlioz A, Martelin R, Gayral JP, and Raoult D. 2000. 16S
559 ribosomal DNA sequence analysis of a large collection of environmental
560 and clinical unidentifiable bacterial isolates. *J Clin Microbiol* 38:3623-
561 3630.

562 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST.
563 *Bioinformatics* 26:2460-2461. 10.1093/bioinformatics/btq461
564 Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial
565 amplicon reads. *Nat Methods* 10:996-998. 10.1038/nmeth.2604
566 Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, and Sogin ML.
567 2013. Oligotyping: Differentiating between closely related microbial taxa
568 using 16S rRNA gene data. *Methods Ecol Evol* 4. 10.1111/2041-
569 210X.12114
570 Ewing B, Hillier L, Wendl MC, and Green P. 1998. Base-calling of automated
571 sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8:175-
572 185.
573 Fox GE, Wisotzkey JD, and Jurtshuk P, Jr. 1992. How close is close: 16S rRNA
574 sequence identity may not be sufficient to guarantee species identity. *Int J*
575 *Syst Bacteriol* 42:166-170. 10.1099/00207713-42-1-166
576 Gouda S, Das G, Sen SK, Shin HS, and Patra JK. 2016. Endophytes: A Treasure
577 House of Bioactive Compounds of Medicinal Importance. *Front Microbiol*
578 7:1538. 10.3389/fmicb.2016.01538
579 He Y, Caporaso JG, Jiang X-T, Sheng H-F, Huse SM, Rideout JR, Edgar RC, Kopylova
580 E, Walters WA, Knight R, and Zhou H-W. 2015. Stability of operational
581 taxonomic units: an important but neglected property for analyzing
582 microbial diversity. *Microbiome* 3:20. 10.1186/s40168-015-0081-x
583 Huson DH, Auch AF, Qi J, and Schuster SC. 2007. MEGAN analysis of metagenomic
584 data. *Genome Res* 17:377-386. 10.1101/gr.5969107

585 Huson DH, Richter DC, Mitra S, Auch AF, and Schuster SC. 2009. Methods for
586 comparative metagenomics. *BMC Bioinformatics* 10 Suppl 1:S12.
587 10.1186/1471-2105-10-S1-S12
588 Imam J, Singh PK, and Shukla P. 2016. Plant Microbe Interactions in Post
589 Genomic Era: Perspectives and Applications. *Front Microbiol* 7:1488.
590 10.3389/fmicb.2016.01488
591 Kim M, Morrison M, and Yu Z. 2011. Evaluation of different partial 16S rRNA
592 gene sequence regions for phylogenetic analysis of microbiomes. *Journal*
593 *of Microbiological Methods* 84:81-87.
594 http://dx.doi.org/10.1016/j.mimet.2010.10.020
595 Kirk JL, Beaudette LA, Hart M, Moutoglis P, Klironomos JN, Lee H, and Trevors JT.
596 2004. Methods of studying soil microbial diversity. *Journal of*
597 *Microbiological Methods* 58:169-188.
598 http://dx.doi.org/10.1016/j.mimet.2004.04.006
599 Lanzen A, Jorgensen SL, Huson DH, Gorfer M, Grindhaug SH, Jonassen I, Ovreas L,
600 and Urich T. 2012. CREST--classification resources for environmental
601 sequence tags. *PLoS One* 7:e49334. 10.1371/journal.pone.0049334
602 Li H. Toolkit for processing sequences in FASTA/Q formats. *Available at*
603 https://github.com/lh3/seqtk.
604 Magoc T, and Salzberg SL. 2011. FLASH: fast length adjustment of short reads to
605 improve genome assemblies. *Bioinformatics* 27:2957-2963.
606 10.1093/bioinformatics/btr507
607 Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, Szeto E, Platt D,
608 Hugenholtz P, and Relman DA. 2007. Dissecting biological “dark matter”
609 with single-cell genetic analysis of rare and uncultivated TM7 microbes

610 from the human mouth. *Proceedings of the National Academy of Sciences*
611 104:11889-11894.
612 Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, Gevers D,
613 Petrosino JF, Abubucker S, Badger JH, Chinwalla AT, Earl AM, FitzGerald
614 MG, Fulton RS, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V,
615 Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM,
616 Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO,
617 Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum
618 E, Arachchi HM, Armitage G, Arze CA, Ayyaz T, Baker CC, Begg L, Belachew
619 T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Vivien Bonazzi J, Brooks P,
620 Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain
621 PS, Chen IMA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton
622 SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ,
623 Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan
624 SP, Dunne WM, Durkin AS, Edgar RC, Erlich RL, Farmer CN, Farrell RM,
625 Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney L, Foster L, Di
626 Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H,
627 Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J,
628 Gonzalez A, Griggs A, Gujja S, Haas BJ, Hamilton HA, Harris EL, Hepburn
629 TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM,
630 Izard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA,
631 Kelley ST, Kells C, Kinder-Haake S, King NB, Knight R, Knights D, Kong HH,
632 Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL,
633 Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu
634 Y, Lo C-C, Lozupone CA, Lunsford RD, Madden T, Mahurkar AA, Mannon

PJ, Mardis ER, Markowitz VM, Mavrommatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, O'Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers Y-H, Ross MC, Russ C, Sanka RK, Pamela Sankar J, Sathirapongsasuti F, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA, Singh I, Smillie CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ, Truty RM, Vishnivetskaya TA, Walker J, Wang L, Wang Z, Ward DV, Warren W, Watson MA, Wellington C, Wetterstrand KA, White JR, Wilczek-Boney K, Wu YQ, Wylie KM, Wylie T, Yandava C, Ye L, Ye Y, Yooseph S, Youmans BP, Zhang L, Zhou Y, Zhu Y, Zoloth L, Zucker JD, Birren BW, Gibbs RA, Highlander SK, Weinstock GM, Wilson RK, and White O. 2012. A framework for human microbiome research. *Nature* 486:215-221. 10.1038/nature11209

Miguel PSB, de Oliveira MNV, Delvaux JC, de Jesus GL, Borges AC, Tótola MR, Neves JCL, and Costa MD. 2016. Diversity and distribution of the endophytic bacterial community at different stages of Eucalyptus growth. *Antonie van Leeuwenhoek* 109:755-771. 10.1007/s10482-016-0676-7

Morgan XC, and Huttenhower C. 2014. Meta'omic analytic techniques for studying the intestinal microbiome. *Gastroenterology* 146:1437-1448 e1431. 10.1053/j.gastro.2014.01.049

660 Nagelkerke NJ. 1991. A note on a general definition of the coefficient of
661 determination. *Biometrika* 78:691-692.
662 Nguyen N-P, Warnow T, Pop M, and White B. 2016. A perspective on 16S rRNA
663 operational taxonomic unit clustering using sequence similarity. *Npj*
664 *Biofilms And Microbiomes* 2:16004. 10.1038/npjbiofilms.2016.4
665 Pinton R, Varanini Z, and Nannipieri P. 2001. The rhizosphere as a site of
666 biochemical interactions among soil components, plants, and
667 microorganisms.
668 Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, and
669 Glockner FO. 2013. The SILVA ribosomal RNA gene database project:
670 improved data processing and web-based tools. *Nucleic Acids Res*
671 41:D590-596. 10.1093/nar/gks1219
672 Riesenfeld CS, Schloss PD, and Handelsman J. 2004. Metagenomics: genomic
673 analysis of microbial communities. *Annu Rev Genet* 38:525-552.
674 10.1146/annurev.genet.38.072902.091216
675 Santamaria M, Fosso B, Consiglio A, De Caro G, Grillo G, Licciulli F, Liuni S,
676 Marzano M, Alonso-Alemany D, Valiente G, and Pesole G. 2012. Reference
677 databases for taxonomic assignment in metagenomics. *Brief Bioinform*
678 13:682-695. 10.1093/bib/bbs036
679 Schloss PD, and Westcott SL. 2011. Assessing and improving methods used in
680 operational taxonomic unit-based approaches for 16S rRNA gene
681 sequence analysis. *Appl Environ Microbiol* 77:3219-3226.
682 10.1128/aem.02810-10
683 Stackebrandt E, and Goebel B. 1994. Taxonomic note: a place for DNA-DNA
684 reassociation and 16S rRNA sequence analysis in the present species

definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* 44:846-849.

Tikhonov M, Leach RW, and Wingreen NS. 2015. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J* 9:68-80. 10.1038/ismej.2014.117

White JR, Navlakha S, Nagarajan N, Ghodsi M-R, Kingsford C, and Pop M. 2010. Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. *BMC Bioinformatics* 11:152. 10.1186/1471-2105-11-152

Woese CR. 1987. Bacterial evolution. *Microbiol Rev* 51:221-271.

Yoshiki Vázquez-Baeza MP, Antonio Gonzalez and Rob Knight. 2013. EMPeror: a tool for visualizing high-throughput microbial community data.

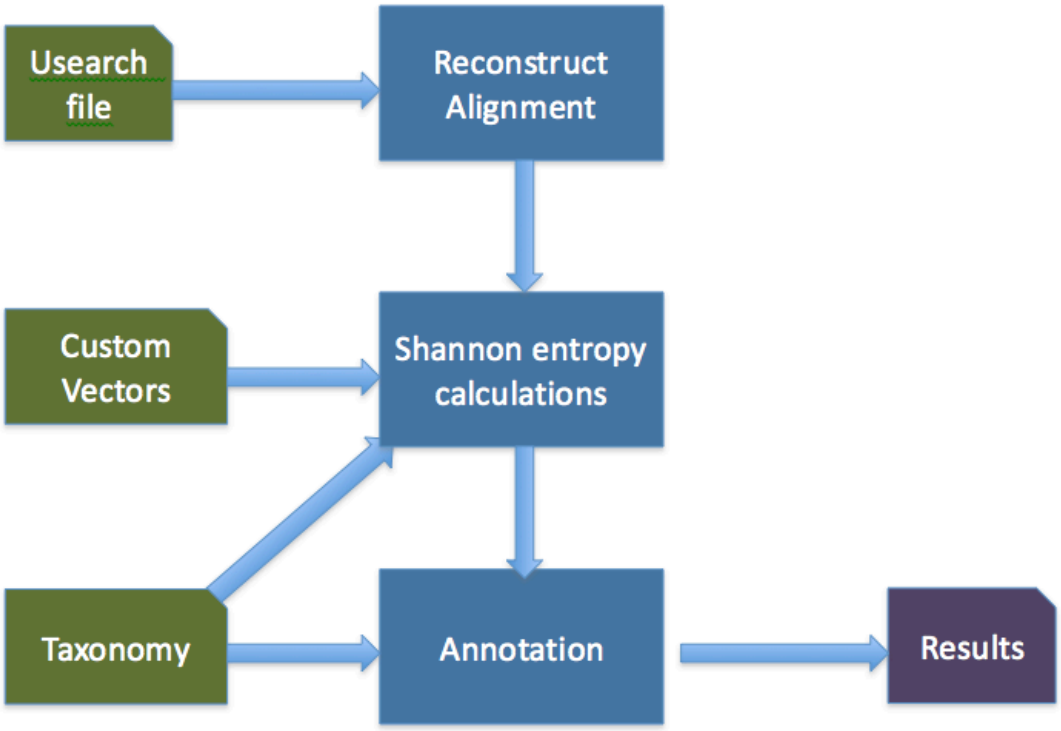
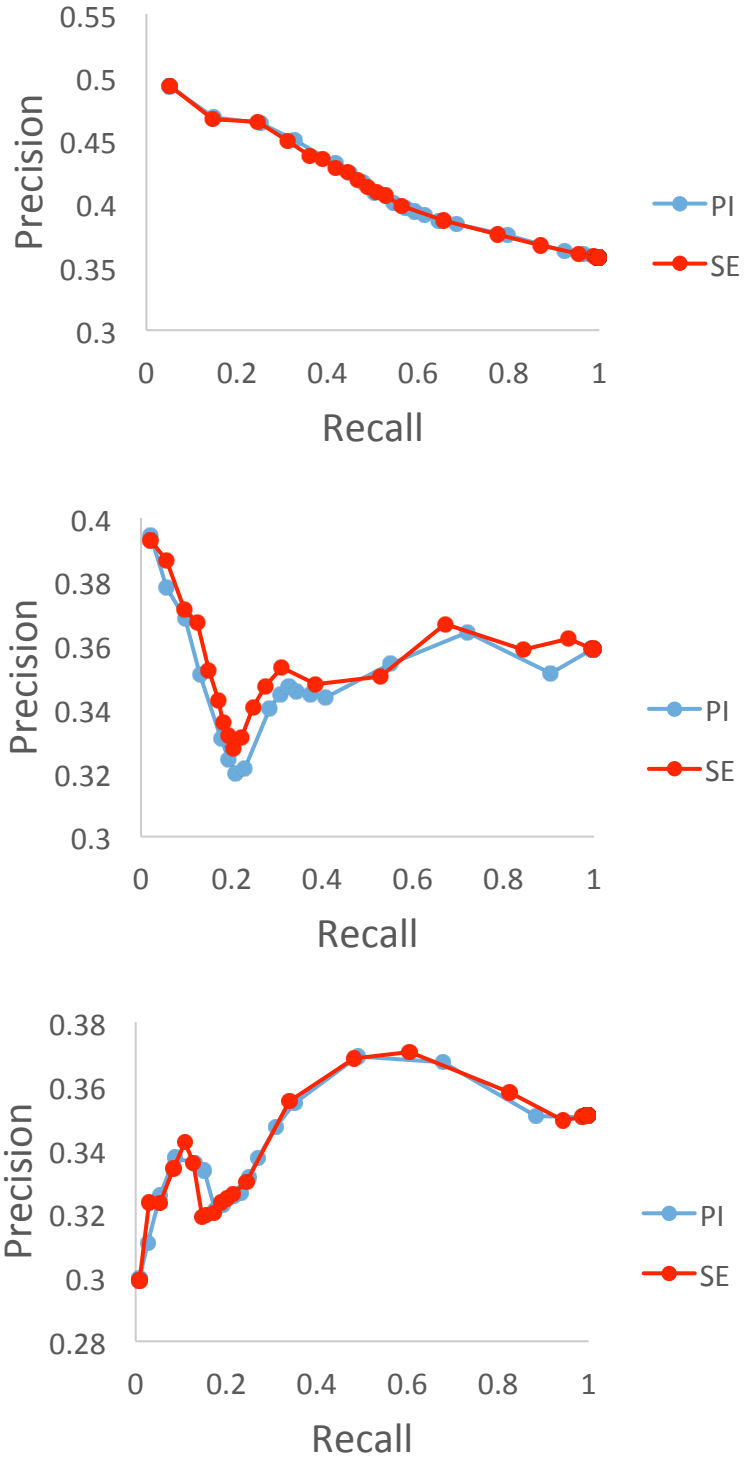
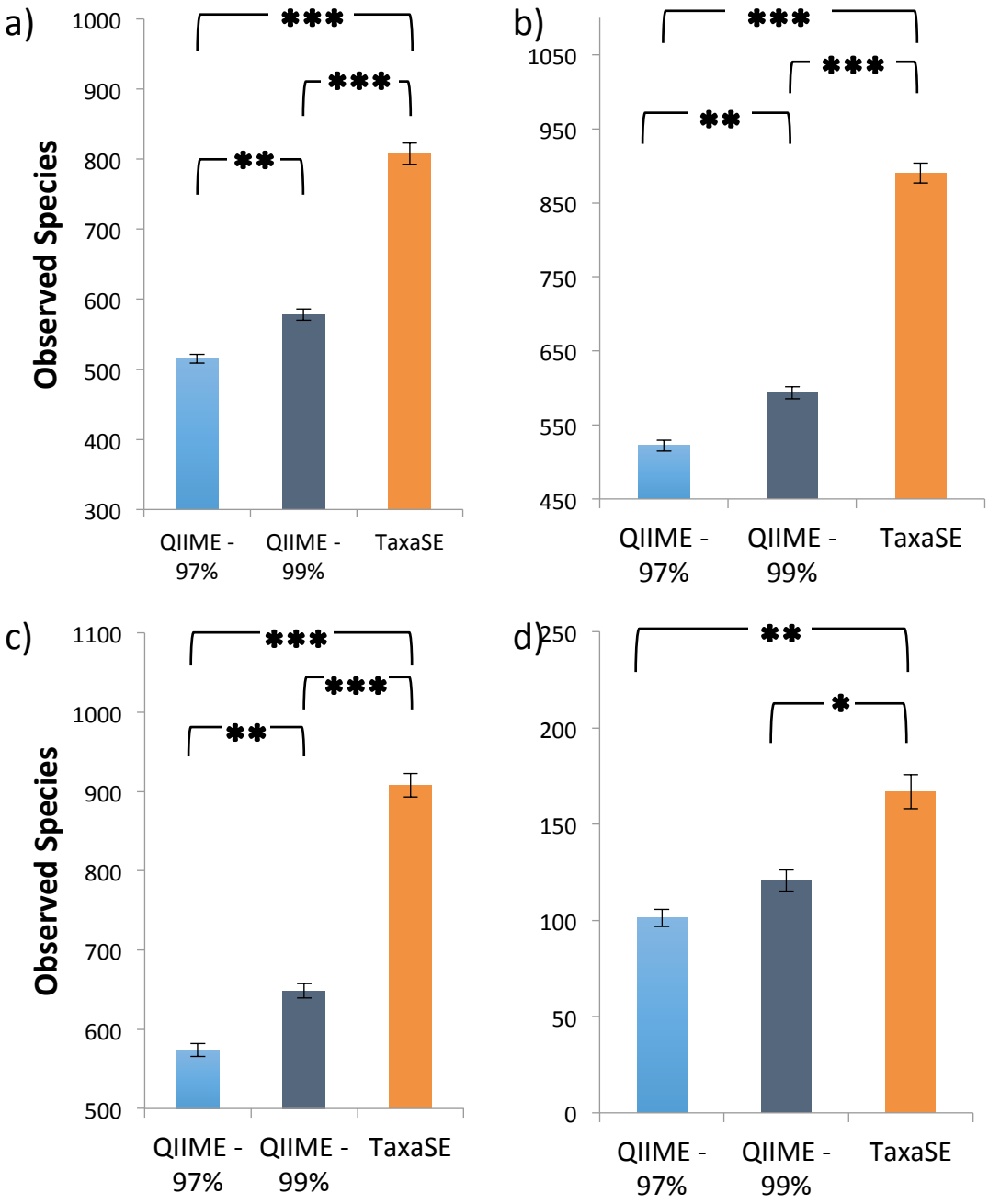


Figure 1: System Process Diagram where data files are shown in green, processing tasks in blue and results in purple.



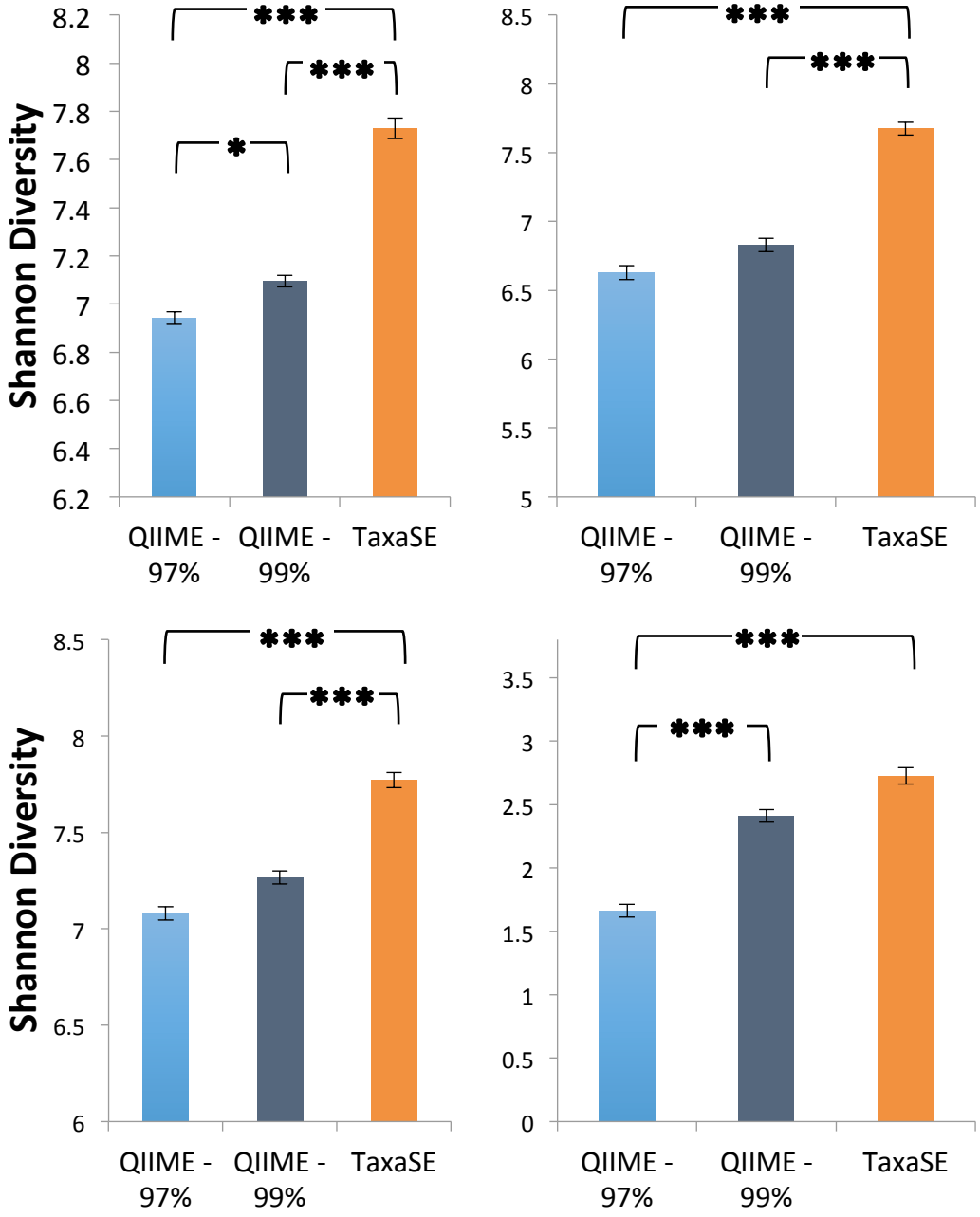
725
726 **Figure 2: Precision vs. recall graph for a) removal of genera dataset b) removal**
727 **of families dataset and c) removal of class dataset, with percentage identity in**
728 **blue and Shannon entropy approach in red.**

729
730



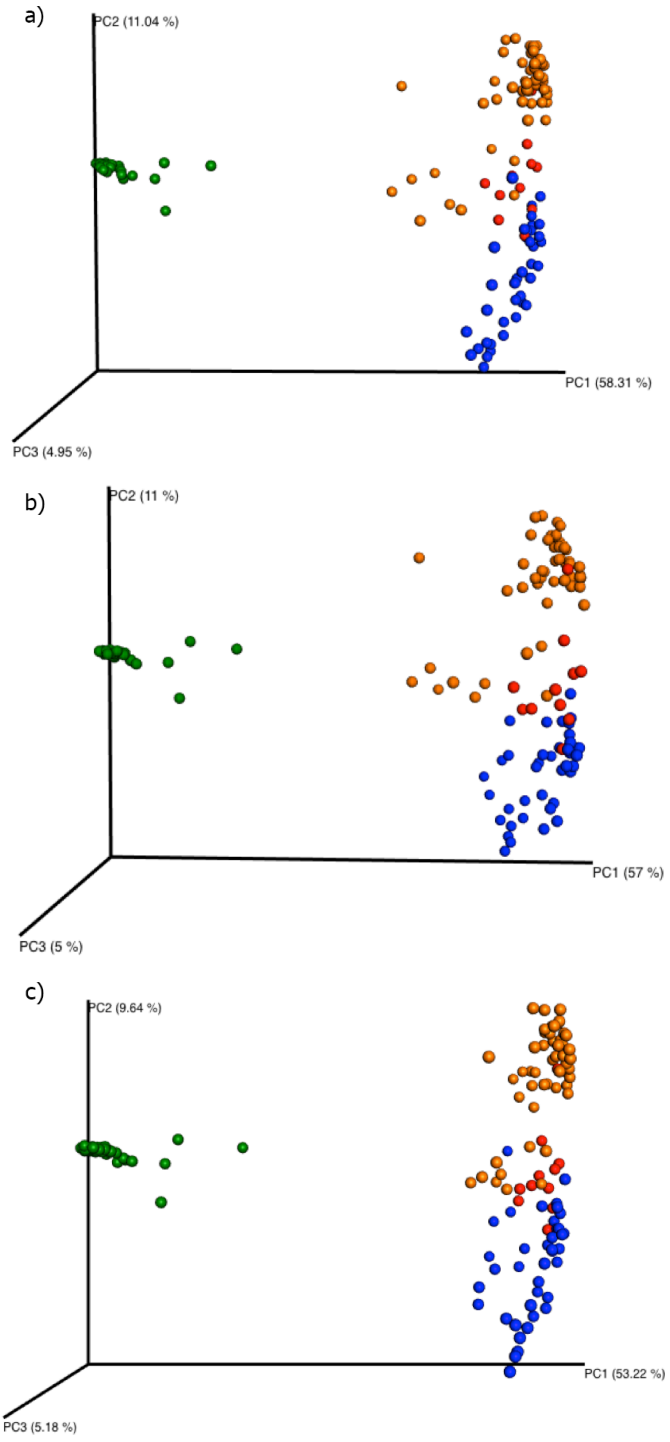
731

732 **Figure 3: Observed species for distinct taxonomic annotation comparison with a)**
733 **rhizosphere, b) root, c) soil and d) stem. QIIME at 97% OTU similarity is shown**
734 **in blue, QIIME at 99% OTU similarity in dark blue and TaxaSE in orange.**
735 **Error bars represent standard error. Significance levels are showed with**
736 **asterisks, where * represents $p < 0.05$, ** represents $p < 0.01$ and *** represents**
737 **$p < 0.001$.**



738

739 **Figure 4: Shannon diversity for distinct taxonomic annotation comparison with**
740 **a) rhizosphere, b) root, c) soil and d) stem. QIIME at 97% OTU similarity is**
741 **shown in blue, QIIME at 99% OTU similarity in dark blue and TaxaSE in**
742 **orange. Error bars represent standard error. Significance levels are shown with**
743 **asterisks, where * represents $p < 0.05$, ** represents $p < 0.01$ and *** represents**
744 **$p < 0.001$.**



745

746 **Figure 5: Beta diversity principle coordinate analysis plots for distinct taxonomic**
747 **annotation comparison of sugarcane dataset with a) QIIME at 97% OTU**
748 **similarity, b) QIIME at 99% OTU similarity and c) TaxaSE. Rhizosphere**
749 **samples are shown in red, root in blue, soil in orange and stem in green.**

Table 1: Environmental sample data used for comparative analysis

Sub-habitat	Number of Samples
Rhizosphere	12
Root	45
Soil	54
Stem	47
Total	158

Table 2: Area under the curve for removal of taxa validation

Area under the curve	Percentage Identity	Shannon Entropy
Genera	**0.393**	0.392
Families	0.345	**0.349**
Class	0.347	**0.348**

Table 3: ADONIS results for distinct taxonomic annotation comparison between QIIME at 97% OTU similarity, QIIME at 99% OTU similarity and TaxaSE.

QIIME at 97% OTU similarity						
	Degree of freedom	Sum of squares	Mean Squares	F-Model	R^2 value	p-value
Habitats	3	25.417	8.4725	99.008	0.67965	0.001
Residuals	140	11.980	0.0856		0.32035	
Total	143	37.398			1.00000	
QIIME at 99% OTU similarity						
	Degree of freedom	Sum of squares	Mean Squares	F-Model	R^2 value	p-value
Habitats	3	25.317	8.4391	95.371	0.67145	0.001
Residuals	140	12.388	0.0885		0.32855	
Total	143	37.706			1.00000	
TaxaSE						
	Degree of freedom	Sum of squares	Mean Squares	F-Model	R^2 value	p-value
Habitats	3	23.700	7.9000	76.743	0.62186	0.001
Residuals	140	14.412	0.1029		0.37814	
Total	143	38.112			1.00000	

Table 4: ANOSIM results for distinct taxonomic annotations comparison between QIIME at 97% OTU similarity, QIIME at 99% OTU similarity and TaxaSE.

ANOSIM		
Approach	p-value	R-value
QIIME at 97%	0.001	0.8528
QIIME at 99%	0.001	0.8558
TaxaSE	0.001	0.8238