Taxonomy annotation errors in 16S rRNA and fungal ITS sequence databases

Robert C. Edgar[1]

[1] Independent investigator, Sonoma, CA, USA

Corresponding Author:

Robert Edgar

Sonoma, CA 95476, USA

Email address: robert@drive5.com

Abstract

Sequencing of the 16S ribosomal RNA (rRNA) gene and the fungal Internal Transcribed Spacer (ITS) region is widely used to survey microbial communities. Specialized ribosomal sequence databases have been developed to support this approach including Greengenes, SILVA and RDP. Most taxonomy annotations in these databases are predictions from sequence rather than authoritative assignments based on studies of type strains or isolates. Here, I investigate the error rates of taxonomy annotations in these databases. I found 253,485 sequences with conflicting annotations in SILVA v128 and Greengenes v13.5 at ranks up to phylum (9,644 conflicts), indicating that the annotation error rate in these databases is ~15%. I found that 34% of non-singleton genera have overlapping subtrees in the Greengenes tree from 2001 according to the RDP taxonomy, most of which are probably due to branching order errors in the Greengenes tree, which is therefore an unreliable guide to phylogeny. Using a blinded test, I estimated that the annotation error rate of the RDP database is ~10%.

Background

Next-generation sequencing the 16S ribosomal RNA (rRNA) gene has revolutionized the study of microbial communities in environments ranging from the human body (Cho and Blaser, 2012; Pflughoeft and Versalovic, 2012) to oceans (Moran, 2015) and soils (Hartmann *et al.*, 2014). Specialized 16S rRNA sequence databases providing taxonomy annotations include and Greengenes (DeSantis *et al.*, 2006), SILVA (Pruesse *et al.*, 2007) and RDP (Maidak *et al.*, 2001). Most taxonomies in the RDP database were predicted by the Naive Bayesian Classifier (Wang *et al.*, 2017), while most taxonomies in Greengenes and SILVA were annotated by a combination of database-specific computational prediction methods and manual curation based on predicted phylogenetic trees (McDonald *et al.*, 2012; Yilmaz *et al.*, 2014).

Microbial taxonomy and phylogeny

The goal of taxonomy has been described as classifying living organisms into a hierarchy of categories (taxa) that are useful (based on biologically informative traits) and monophyletic (consistent with the true phylogenetic tree) (Hennig, 1965), though this point of view is not universally accepted (e.g., (Benton, 2000)). With microbial markers such as 16S rRNA, most organisms are known only from environmental sequencing, and in these cases taxonomy must necessarily be predicted from sequence alone. If taxonomy is based on traits, then a prediction of taxonomy is a prediction that characteristic traits are present, or, more precisely, whether a taxonomist would decide that sufficiently many characteristic traits are present, noting that these may be subject to revision when new species are studied. For organisms known only from sequence, the best evidence for phylogeny is usually a tree inferred from the sequences. It is therefore useful to consider

the problem of taxonomy prediction given a sequence-based tree. This is the approach used by SILVA and Greengenes, while RDP annotations for environmental sequences are predictions made by an algorithm, the Naive Bayesian Classifier, which does not explicitly incorporate phylogeny.

Taxonomy prediction from a sequence-based tree

Consider a binary phylogenetic tree with authoritative taxonomies for a small subset of its leaves. The goal is to predict taxonomies for the remaining leaves which represent organisms known only from sequence. The *authoritative tree* is the implied binary tree for the leaves having authoritative taxonomies, i.e. classifications based on direct observations of characteristic traits. Note that its annotations are considered to be authoritative, but not necessarily its branching order. The *lowest common ancestor* (*LCA*) of a taxon is the lowest node above all its leaves. If the branching order is correct and the taxonomy is consistent with the tree, then the subtree under the LCA for a taxon is *pure*, i.e. contains no other taxa at the same rank. The *lowest pure taxon* (*LPT*) for a node is the taxon at lowest rank for which all the leaves in its subtree belong to the same taxon. For example, the LPT for node n is family F if, and only if, the leaves in the subtree under n belong to two or more genera in F. If F contains exactly one genus G, then F cannot be the LPT of any node because if F is pure then G will also be pure and G is lower. A *pure edge* connects two nodes having the same LPT. For example, consider the tree shown in Fig. 1. Sequences s, g_1 and g_2 are type strains with authoritative taxonomy; w, x, y and z are sequences obtained from environmental samples. The authoritative tree (black edges) connects s, g_1 and g_2. Genus is the lowest rank; s belongs to a singleton genus S (i.e., it is the only authoritatively named sequence for that genus); g_1 and g_2 belong to the same genus G. S and G belong to the same

family, *F*. The LCA for *S* is *s* (a singleton leaf is always its own LCA), and the LCA for *G* is *g*. The LPTs of g_1, g_2 and *g* are *G*, the LPT of *s* is *S*, and the LPT of *f* is *F*. Pure edges are shown by continuous lines; impure edges are dashed lines. If the tree is consistent with the taxonomy, then all impure edges join an LCA node for some taxon *t* to a node which has an LPT at a higher rank than *t*, and conversely every LCA node has one impure edge which connects it to a higher node. Purity is closely related to monophylicity; however, monophylicity is generally understood to be with respect to the true but unknown phylogeny, while purity is defined with respect to a given predicted phylogeny which may have an incorrect branching order.

The clade extension problem

Adding a new sequence (*Q*) of unknown taxonomy introduces a new node into the tree, its *join node*, indicated by an arrowhead in Fig. 1. Assume the tree is correct and the taxonomy is consistent with the tree. Then, if the join node is below the LCA for taxon *t*, *Q* certainly belongs to *t*. In Fig. 1, *y* certainly belongs to *G* because its join node is below *g*, the LCA of *G*. Also, *w* certainly belongs to a novel genus because it joins a family edge (which is above two or more genera by definition), and a new genus must therefore be introduced because it is impossible for the join node or any node above it to be the LCA for any known genus. The environmental sequence *x* joins the tree at the impure edge joining *s* and *f*, so it is above the LCA for *S*, but the subtree under the join node contains only *S*. The tree therefore cannot resolve whether *x* belongs to *S* or to a novel genus because both possibilities are consistent with the data. If *x* belongs to *S*, its join node will be the new LCA for *S*, expanding the clade by moving its LCA higher in the tree. Otherwise, *x* belongs to a novel singleton

genus, call it *V*, and *x* will be the LCA of *V* if it is added to the tree. Thus, *s* is the only sequence which unambiguously belongs to *S*—if a new sequence *Q* differs from *s*, then the join node is above the LCA for *S* and *Q* could belong to a different genus. This ambiguity could be resolved by a sequence identity threshold. For example, if *x* is 97% identical to *s*, then it might be reasonable to assign *x* to *S*, but identity thresholds are only approximate guides and this prediction is uncertain even if the tree is assumed to be correct. A similar problem arises in non-singleton taxa. The join node for *z* is above the LCA for *G*, but the subtree below the join node contains only *G* and *z* could therefore belong to *G* (which would expand the clade by moving the LCA for *G* up to the join node for *z*), or to a novel genus, in which case *z* would be a new singleton LCA. In general, if a sequence *Q* joins the tree at an impure edge, then the tree is consistent with assigning *Q* to a new taxon or extending an existing clade to include *Q.*

The subtree impurity problem

Trees inferred from large sequence sets are unreliable guides to phylogeny due to incorrect branching orders (Huelsenbeck *et al.*, 2000; Philippe *et al.*, 2011). Naive assumptions that the tree is correct and the taxonomy is consistent with the tree are violated by branching order errors and also by taxa which are found to overlap by sequence such as *Escherichia* and *Shigella* (Escobar-Páramo *et al.*, 2003), which may be regarded as errors in the taxonomy. A conflict between the tree and the taxonomy occurs when taxa at the same rank overlap (Fig. 2). An *overlapped leaf* is below LCAs for two or more taxa at the same rank. Conversely, a taxon *t* is consistent with the tree if, and only if, the subtree under the LCA for *t* is pure. In Fig. 2, *i* is the LCA for *G* and *j* is the LCA for *H*; neither is pure because *i* is above h_1 and *j* is above all members of both *G* and *H*. If an environmental sequence *Q* joins the tree

at an edge where taxa overlap, i.e. the edge is above leaves belonging to two or more taxa at a given rank but below the LCAs for those taxa, the prediction at that rank is ambiguous. For example, consider cases where the join node for *Q* is in *p*-*i*, *q*-*i*, h_1-*q* or g_3-*q*. Here, *Q* could plausibly belong to *G*, *H* or a novel genus, depending on the underlying explanation of the tree conflict. If *Q* joins the tree below a node for which taxon *t* is pure, then it is plausible that *Q* belongs to *t*, though this *purity heuristic* is not reliable because the branching order may be incorrect. For example, suppose h_1 in Fig. 2 is an environmental sequence rather than a known strain, then the subtree under *i* is pure at genus rank and h_1 would be incorrectly assigned to *G* by the purity heuristic.

Methods

Database versions

I used Greengenes v13.5, RDP downloaded on 15th January 2017, and SILVA v128.

High-identity outliers

The *lowest common rank* (LCR) (Edgar, 2018) is the lowest rank shared by a pair of sequences. Note that an LCR is the name of a rank, not a taxon name at that rank. For example, if two sequences belong to different genera in the same family, then their LCR is genus. If a pair of sequences has an anomalously high identity considering its LCR, this indicates a possible annotation error, though it could also be explained by unusually high sequence conservation or by lateral transfer. For example, if two 16S rRNA sequences are 99% identical and are annotated as belonging to different orders in the same class, this is likely to be a problem in the annotations. This approach can be applied most effectively to a

reference used to make annotations in a larger database because an anomaly in the reference may propagate to multiple predictions, obscuring the underlying issue. I tested this method using RDP 16S rRNA training set v16, the BLAST 16S rRNA reference database (BLAST16S) (Sayers *et al.*, 2012) downloaded 17th January 2017, and the Warcup fungal ITS reference v2 (Deshpande *et al*., 2015).

Nomenclature conflicts

At least three prokaryotic taxonomy nomenclatures are widely used. The Greengenes nomenclature is based on the NCBI Taxonomy database (Sayers *et al.*, 2012), RDP on Bergey's Manual (Bergey, 2001), and SILVA on LSPN (Parte, 2014). These nomenclatures differ mainly in revisions to resolve conflicts with sequence and add new candidate groups identified by sequence. For example, *Escherichia* and *Shigella* overlap (see Background), which Greengenes resolves by leaving their sequences unclassified at ranks below *Enterobacteriaceae.* SILVA deals with this issue differently by retaining well-established species names such as *Escherichia coli* and introducing a combined genus named *Escherichia-Shigella*. An example of a candidate group is JS1 (Webster *et al.*, 2004), which is found in the SILVA nomenclature but not in the tested versions of RDP or Greengenes. Most well-established taxa are found in all nomenclatures, and in these cases a biologist would presumably expect a taxon name to be placed within the same group at a higher rank. To investigate this for a representative pair of large databases, I identified the *common nomenclature* as the set of names which appear in both SILVA and Greengenes. If a taxon name and its parent name were both found in the common nomenclature, I tested whether the databases agreed on the parent name.

Conflicting annotations in Greengenes and SILVA

I identified sequences found in both SILVA and Greengenes that are annotated by names in their common nomenclature in at least one of the databases, and listed cases where the annotations are different. A disagreement must be due a difference in the nomenclature hierarchy, a different definition of the taxon (e.g., different characteristic traits), or errors in one or both annotations.

Blinded test of RDP predictions

If a sequence that was previously known only from environmental samples is found in a newly-sequenced isolate strain, it can provide a blinded test of taxonomy prediction analogous to CASP (Moult, 2005), where protein structures predicted from sequence are compared to newly-solved structures determined by sequence-independent methods such as X-ray crystallography. In principle, predictions in an older version of Greengenes or SILVA could be assessed by comparison with isolates sequenced after annotations were generated, but this approach is stymied by a lack of data specifying which sequences were used as references and the deposition dates of new isolate sequences. With RDP, it is straightforward because some of the older training sets for the Naive Bayesian Classifier are deposited in public archives, and new sequences are readily identifiable as those not found in an older training set. If a training set is considered to be an authoritative reference, this enables predictions for new, authoritatively classified sequences to be assessed using an old training set as a reference, which is effectively equivalent to a blinded test. I implemented this test using versions 9 and 16 of the RDP reference data, which were the oldest and newest, respectively, available for download at the time this work was performed. Version 9 contains 10,049 sequences; version 16 has 13,212 and therefore

contains more than three thousand new sequences. I assessed predictions by the RDP Classifier on these new sequences using v9 as a reference and the v16 annotations as the truth standard.

Prediction performance metrics for the blinded test

I used the following metrics (Edgar, 2014, 2018) to measure prediction accuracy using a reference database (A) with a query set (S) containing sequences with known taxonomy. An *over-classification error* is a false positive where the taxon name of the query sequence is not present in the reference database. A *misclassification error* is a false positive where the taxon is known but the wrong name is predicted. An *under-classification error* is a false negative; i.e., case where no name is predicted but the taxon is known (present in A). Let N be the number of sequences in S, K be the number of query sequences with known names, i.e. names which are present in the reference database A, and $L = N - K$ be the number of novel test sequences, i.e. sequences in S with names that are not present in A. Let TP be the number of names which are correctly predicted, MC the number of misclassification errors, OC the number of over-classification errors, and UC the number under-classification errors. The rate for each type of error is defined as the number of errors divided by the number of opportunities to make that error: $OCR = OC/L$ (over-classification rate), $UCR = UC/K$ (under-classification rate) and $MCR = MC/K$ (misclassification rate). The true positive rate is $TPR = TP/K$, i.e. the number of correct names divided by the number of opportunities to correctly predict a name. *Accuracy* is calculated as $Acc = TP/(K + OC)$, i.e. the number of correct predictions divided by the number of predictions for which correctness can be determined. Metrics are calculated separately for each rank.

Conflicts between the RDP taxonomy and Greengenes tree

Greengenes and SILVA use sequence-based trees to guide predictions of taxonomy for environmental sequences. Conflicts between a phylogenetic tree and a taxonomic hierarchy arise when taxa at the same rank overlap in the tree (see Background). Such conflicts could readily be detected given the tree and the authoritative annotations used as the basis for prediction, but this data is not provided by Greengenes or SILVA and I therefore chose to assess the Greengenes tree from 2011 (http://greengenes.lbl.gov/Download/Trees/16S_all_gg_2011_1.tree.gz) using taxonomy annotations from RDP training set v16 as a reference. This assesses an old tree by comparison with a new authoritative reference, following the rationale of the previous section. Annotations were propagated from the RDP reference by identifying exactly-matching sequences in Greengenes. I extracted the binary tree implied by the Greengenes tree for leaves annotated by the RDP reference. For each taxon in the RDP hierarchy, I identified its LCA (i.e., the lowest node above all members of the taxon), and then the overlapped taxa and leaves at each rank. If taxonomy is required to be consistent with the true phylogenetic tree, then an overlap must be due to an error in the branching order of the tree and/or an error in the taxonomy.

Results

Annotation errors in the full RDP database

Annotations in the full RDP database are generated by the Naive Bayesian Classifier (NBC). Performance metrics for NBC predictions on new sequences in RDP training set v16 using v9 as a reference are shown in Table 1. The top-hit identity distribution (Edgar, 2018) of

the blinded test is shown in Fig. 3, together with the distribution of the full RDP database against training set v16 for comparison. The full distribution is strongly peaked at 99% identity, while the blinded distribution has a higher frequency of low-identity sequences. The blinded test is therefore more challenging than annotating the full database, and the average genus accuracy of the RDP annotations is probably higher than the ~85% observed in the blinded test. Given these results, a rough estimate is that ~10% sequences in RDP have incorrect genus annotations.

Conflicts between the RDP taxonomy and Greengenes tree

I transferred taxonomies from RDP training set v16 to the Greengenes tree from 2011 by finding identical sequences (see Methods), giving a binary tree with 20,952 leaves where the branching order is given by the Greengenes tree and taxonomy annotations at the leaves are given by the RDP training set. Results are shown in Table 2. Conflicts between the taxonomy and tree are identified as overlaps. At all ranks, a large majority of the leaves are overlapped; i.e., are found in the subtrees of two or more LCA nodes at the same rank. A taxon is defined to be overlapped if its LCA contains leaves belonging to other taxa at its rank. There are 1,454 genera in the authoritative tree; 586 of these are singletons which are necessarily pure. Of the 868 non-singleton genera, 294 (34%) are overlapped, and 4/28 phyla (14%) are overlapped. While some conflicts could be due to errors in the taxonomy, these results strongly suggest pervasive branching order errors in the Greengenes tree.

Nomenclature hierarchy disagreements between Greengenes and SILVA

I found that 46 taxon names were placed in different parent taxa by Greengenes v13.5 and SILVA v128, considering only names present in both databases. For example, family

Brevibacteriaceae is in order *Actinomycetales* according to Greengenes (e.g. accession 1109545), but this family is in order *Micrococcales* according to SILVA (JAJB01000072.80.1512). Order *Micrococcales* is also found in Greengenes (825086), so there is an unambiguous disagreement between the nomenclature hierarchies. A complete list is given in Supp. Table 1.

Annotation conflicts between Greengenes and SILVA

I found 253,485 sequences present in both Greengenes v13.5 and SILVA v128 with conflicting annotations. A rank was considered to disagree if the names at that rank were different and one or both were in the common nomenclature, i.e. was a name found in both databases. 61,744 (24%) of these conflicts were due to a rank which was blank (unspecified) in one database, which implies a false positive by one database or a false negative by the other. An example of a conflict where both names are in the common nomenclature is Greengenes 4366627 which has the same sequence as SILVA LOSM01000005.1106908.1108461. Greengenes assigns this sequence to family *Pseudoalteromonadaceae* while SILVA assigns it to family *Vibrionaceae*. *Pseudoalteromonadaceae* is found in SILVA (FJ889568.1.1501) and Greengenes contains *Vibrionaceae* (3059893). A small fraction of the conflicts (440, or 0.2%) reflected nomenclature hierarchy disagreements (see previous section); the rest were conflicts in taxon names for which both databases agreed on the parent group. Conflicts were found at all ranks: 9,644 at phylum, 31,851 class, 93,534 order, 48,945 family and 69,511 genus, counting only the highest conflicted rank. An example of a phylum conflict in a sequence classified to genus level by both databases is Greengenes 851746 which is assigned to genus *Streptococcus* in phylum *Firmicutes* and has the same sequence as SILVA

JZIA01000003.3571786.3573313, which is assigned to genus *Xanthomonas* in phylum *Proteobacteria*. If a given named taxon is required to have the same definition (e.g., the same characteristic traits) in all nomenclatures where it appears, then all conflicts are necessarily due to annotation errors in one or both databases. A conservative estimate of the total number of such errors is then the lower bound where all conflicts are due to an error in one database but not the other, i.e. a total of ~250k. Absent other evidence, it is reasonable to assume that the error rate is similar in both databases, giving an estimate of half the total, i.e. ~100k errors per database after rounding down. 596,648 (~600k) matching sequences had annotations from the common nomenclature in one or both databases. This is the number of cases where conflicts are identifiable, giving a conservative estimate of ~100k/~600k (~15%) for the annotation error rates of SILVA and Greengenes.

High-identity outliers

Representative outliers for BLAST16S are shown in Table 3, for Warcup in Table 4 and for the RDP 16S training set in Table 5. The ten pairs with highest identity at each rank for all tested databases are given in Supp. Table 2. As an example, sequences gi|645320505 and gi|265678369 in BLAST16S are 100% identical and have LCR=phylum, which probably indicates an annotation error. In Warcup, several cases are identified where the genus of the species binomial name is different from the genus annotation *per se*; for example, *Sporisorium destruens* is assigned to genus *Sphacelotheca* in accession AY344979. At least some of these cases reflect nomenclature revisions that were only partially implemented in Warcup. For example, *Serpula incrassata* is an older name for *Meruliporia incrassata* (Skrede *et al.*, 2011).

Discussion

The results presented here suggest that the annotation error rates of Greengenes and SILVA are ~15% and RDP is ~10%, though the uncertainties in these estimates are too great to support a conclusion that annotations in one of these databases are better or worse than another. However, while keeping this caveat in mind, it is striking that a pessimistic estimate for the RDP database of 15% on the blinded test is comparable to a conservative estimate for SILVA and Greengenes of a 15% error rate from disagreements between annotations. This suggests that the accuracy of automated annotations generated by the Naive Bayesian Classifier, which does not explicitly consider phylogeny, is comparable to or perhaps better than the accuracy of annotations in SILVA and Greengenes, which were generated by a combination of automated and manual methods guided by alignments and trees. Predictions generated by the RDP Classifier have several advantages: they are based on a documented reference, can easily be reproduced, and report a bootstrap confidence value. This allows a biologist to critically review the evidence underlying any annotation of interest. By contrast, annotations in SILVA and Greengenes do not report confidence and are not reproducible because there is a manual curation step and references are not documented.

References

Benton MJ. (2000). Stems, nodes, crown clades, and rank-free lists: Is Linnaeus dead? *Biol Rev* **75**: 633–648.

Bergey DH. (2001). Bergey's manual of systematic bacteriology. Wiley.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Cho I, Blaser MJ. (2012). The human microbiome: at the interface of health and disease. *Nat Rev Genet* **13**: 260–270.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–72.

Deshpande V, Wang Q, Greenfield P, Charleston M, Porras-Alfaro A, Kuske CR, Cole JR, Midgley DJ, Tran-Dinh N. (2015). Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia* 14-293-.

Escobar-Páramo P, Giudicelli C, Parsot C, Denamur E. (2003). The evolutionary history of Shigella and enteroinvasive Escherichia coli revised. *J Mol Evol* **57**: 140–148.

Edgar RC. (2014). Taxonomy benchmarks, *USEARCH manual v8.1*, https://drive5.com/usearch/manual8.1/tax_bench.html.

Edgar, RC. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences.

Hartmann M, Niklaus PA, Zimmermann S, Schmutz S, Kremer J, Abarenkov K, Lüscher P, Widmer F, Frey B. (2014). Resistance and resilience of the forest soil microbiome to logging-associated compaction. *ISME J* **8**: 226–44.

Hennig W. (1965). Phylogenetic Systematics. *Annu Rev Entomol* **10**: 97–116.

Huelsenbeck JP, Rannala B, Masly JP. (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science (80-)* **288**: 2349–2350.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610–618.

Moran MA. (2015). The global ocean microbiome. *Science (80-)* **347**: aac8455.

Moult J. (2005). A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* **15**: 285–289.

Nilsson RH, Tedersoo L, Ryberg M, Kristiansson E, Hartmann M, Unterseher M, Porter TM, Bengtsson-Palme J, Walker DM, De Sousa F, Gamper HA, Larsson E, K.-H. Larsson , Kõljalg U, Edgar RC, Abarenkov K. (2015). A comprehensive, automatically updated fungal ITS sequence dataset for reference-based chimera control in environmental sequencing efforts. *Microbes Environ* **30**. e-pub ahead of print, doi: 10.1264/jsme2.ME14121.

Parte AC. (2014). LPSN - List of prokaryotic names with standing in nomenclature. *Nucleic Acids Res* **42**. e-pub ahead of print, doi: 10.1093/nar/gkt1111.

Pflughoeft KJ, Versalovic J. (2012). Human microbiome in health and disease. *Annu Rev Pathol* **7**: 99–122.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol* **9**. e-pub ahead of print, doi: 10.1371/journal.pbio.1000602.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Krasnov S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J. (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **40**: D13-25.

Skrede, I, Engh, IB, Binder, M, Carlson, T, Kauserud, H, Bendiksby, M. (2011). Evolutionary history of Serpulaceae (Basidiomycota): molecular phylogeny, historical biogeography and evidence for a single transition of nutritional mode.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–7.

Webster G, Parkes RJ, Fry JC, Weightman AJ. (2004). Widespread occurrence of a novel division of bacteria identified by 16S rRNA gene sequences originally found in deep marine sediments. *Appl Environ Microbiol* **70**: 5708–5713.

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. (2014). The SILVA and 'all-species Living Tree Project (LTP)' taxonomic frameworks. *Nucleic Acids Res* **42**. e-pub ahead of print, doi: 10.1093/nar/gkt1209.

Figures

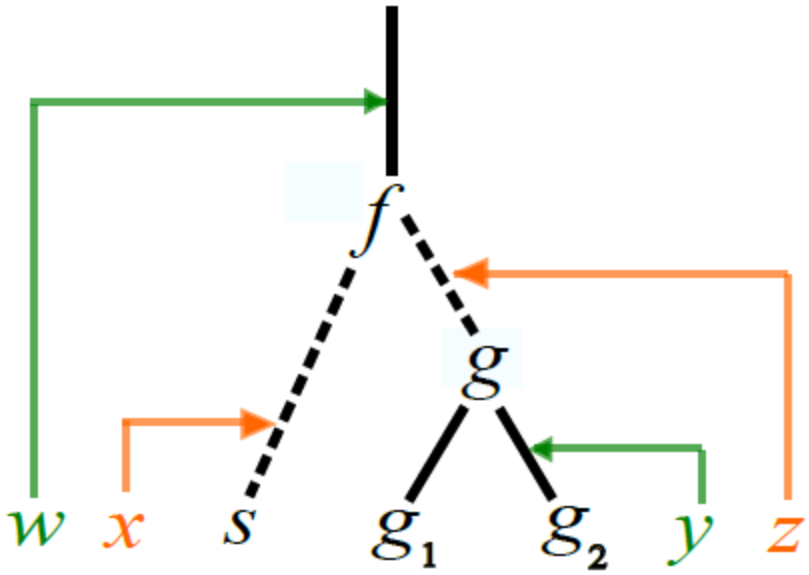

Figure 1. Example tree showing join nodes and impure edges.

Leaves s, g_1 and g_2 are sequences with authoritative taxonomies. Sequences w, x, y and z are environmental sequences. Sequence s belongs to genus S, which is a singleton (I.e., has only one authoritative sequence); sequences g_1 and g_2 belong to genus G. The LCA for G is g, the LCA for S is s. The inferred positions of w, x, y and z in the tree are indicated by arrows. An arrowhead is a join node, i.e. the new node created by adding a new sequence to the tree. Black edges are the authoritative tree, i.e. the binary tree for leaves with authoritative taxonomies. Dashed lines are impure edges; continuous lines are pure edges (see main text).

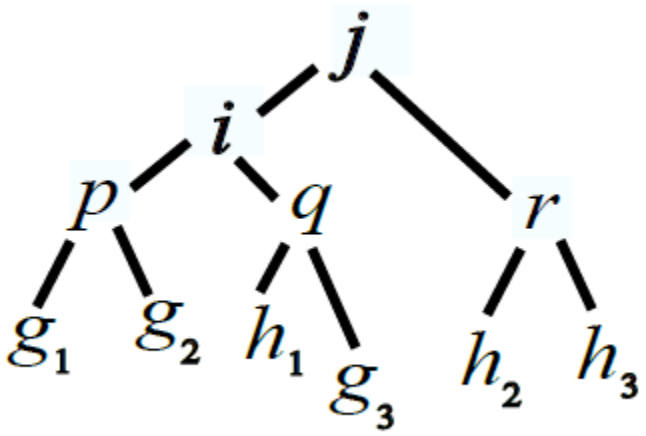

Figure 2. Example tree showing conflicts with taxonomy.

Leaves g_1, g_2 and g_3 belong to genus *G*, h_1, h_2 and h_3 belong to a different genus *H*. The tree is not consistent with the taxonomy because *G* and *H* overlap. The LCA for *G* is *i*, which is above h_1, and the LCA for *H* is *j*, which is above all members of both *G* and *H*. Overlaps are common in practice, where most are probably explained by errors in the branching order of the predicted phylogenetic tree (see main text).

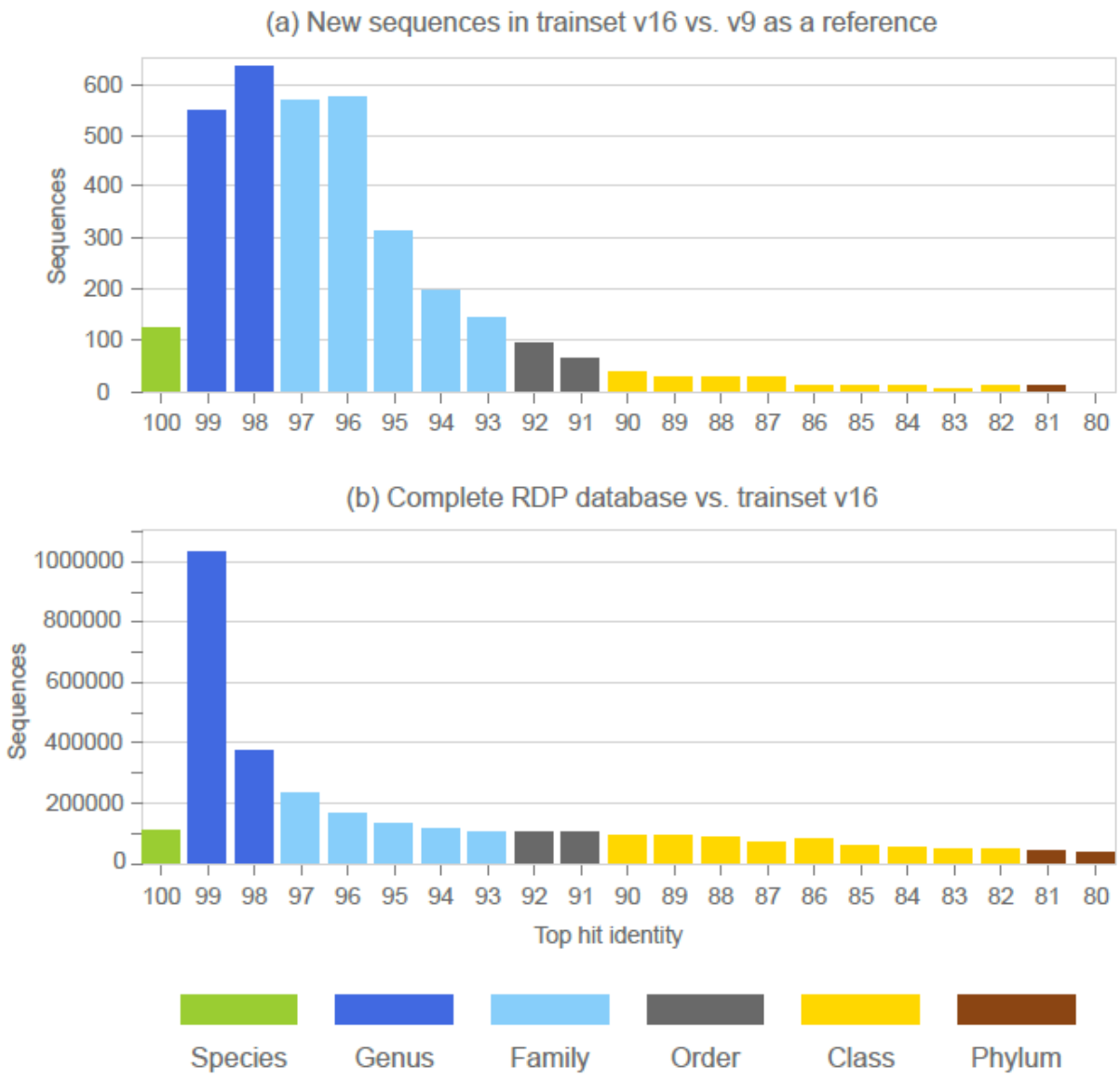

Figure 3. Top-hit identity distribution for the blinded RDP prediction test.

Panel (a) shows the top-hit identity distribution (THID (Edgar, 2018)) for new sequences in RDP trainset v16 compared to v9 as a reference. Panel (b) shows the THID for the full RDP database against trainset v16 as a reference. Histogram bars are colored according to the most probable lowest common rank (Edgar, 2018).

Tables

Rank	*TPR*	*MCR*	*OCR*	*UCR*	*Acc*
Phylum	99.5%	0.1%	41.1%	0.1%	98.8%
Class	99.1%	0.4%	90.2%	0.4%	83.5%
Order	97.8%	0.7%	84.2%	0.7%	92.9%
Family	96.9%	0.7%	71.7%	0.7%	92.8%
Genus	91.8%	2.6%	41.9%	2.6%	84.8%

Table 1. Accuracy metrics on blinded test of RDP annotation prediction accuracy.

Taxonomies for new sequences in RDP training set v16 were predicted using training set v9 as a reference and assessed using CVI metrics with v16 annotations as a truth standard.

Rank	*Pure taxa*	*Overlapped taxa*	*Pure leaves*	*Overlapped leaves*
Phylum	28	4	615	20,337
Class	50	11	615	20,337
Order	99	39	501	20,451
Family	205	106	455	20,497
Genus	1,454	294	7,633	13,319

Table 2. Conflicts between RDP taxonomy and Greengenes tree.

A pure taxon has no other taxa at the same rank under its LCA, otherwise it is overlapped. A pure leaf is under exactly one LCA at the given rank, if there are two or more it is overlapped.

%Id	*LCR*	*Accessions*	*Highest Distinct Taxa*	*Species*
100.0	Phylum	gi|645320505 gi|265678369	c:Betaproteobacteria c:Gammaproteobacteria	*Neisseria flava* *Moraxella caviae*
97.1	Phylum	gi|559795246 gi|645321781	c:Bacilli c:Clostridia	*Lactobacillus rogosae* *Lachnospira pectinoschiza*
99.9	Class	gi|558508609 gi|631251240	o:Pseudonocardiales o:Streptomycetales	*Prauserella marina* *Streptomyces parvus*
99.6	Class	gi|645319747 gi|343198728	o:Pseudomonadales o:Oceanospirillales	*Pseudomonas halophila* *Halovibrio denitrificans*
100.0	Order	gi|636559468 gi|219846253	f:Bacillaceae f:Sporolactobacillaceae	*Bacillus racemilacticus* *Sporolactobacillus laevolacticus*
100.0	Order	gi|645320480 gi|343200250	f:Rhodobiaceae f:Bradyrhizobiaceae	*Afifella marina* *Rhodopseudomonas julia*
100.0	Family	gi|636560543 gi|631251787	g:Obesumbacterium g:Hafnia	*Obesumbacterium proteus* *Hafnia alvei*
100.0	Family	gi|1212229262 gi|1137647790	g:Feifantangia g:Xanthomarina	*Feifantangia zhejiangensis* *Xanthomarina gelatinilytica*
100.0	Genus	gi|343198889 gi|343201672	s:aurantiaca s:erecta	*Stigmatella aurantiaca* *Stigmatella erecta*
100.0	Genus	gi|219857540 gi|219846097	s:florum s:grammopterae	*Mesoplasma florum* *Mesoplasma grammopterae*

Table 3. Outliers with high identity in the BLAST 16S rRNA database.

Columns are: *%id*, identity; *LCR*, lowest common rank; *Accessions*, database identifiers; *Highest Distinct Taxa*, names at the highest rank where the sequences are assigned to different taxa; *Species*, species names according to the database.

%Id	*LCR*	*Accessions*	*Highest Distinct Taxa*	*Species*	*Note*
96.5	Phylum	JX145398 HQ438030	c:Eurotiomycetes c:Leotiomycetes	*Phialophora mustea* *Cadophora malorum*	
96.0	Phylum	AY344979 AY344976	c:Ustilaginomycetes c:Microbotryomycetes	*Sporisorium formosanum* *Sporisorium destruens*	§1
98.5	Class	FJ897698 FJ230997	o:Onygenales o:Eurotiales	*Histoplasma capsulatum* *Penicillium corylophilum*	
98.4	Class	JN198395 JF934921	o:Botryosphaeriales o:Pleosporales	*Microdiplodia hawaiiensis* *Paraconiothyrium variabile*	
99.6	Order	GU237803 JN712495	f:Montagnulaceae f:Didymellaceae	*Microsphaeropsis olivacea* *Microsphaeropsis proteae*	§2
99.2	Order	AY616209 AY604527	f:Dothioraceae f:Dothideaceae	*Hormonema carpetanum* *Endoconidioma populi*	
100.0	Family	GU226846 EU918712	g:Botryosphaeria g:Dothiorella	*Botryosphaeria dothidea* *Dothiorella gregaria*	
99.8	Family	AF414349 HQ260622	g:Trichoderma g:Hypocrea	*Trichoderma parceramosum* *Trichoderma ghanense*	§3
100.0	Genus	EU429519 HM135647	(*ambiguous*)	*Meruliporia incrassata* *Serpula incrassata*	§4
100.0	Genus	HM037980 EF055200	s:olivacea s:lespedezae	*Pestalotiopsis olivacea* *Pestalotiopsis lespedezae*	

Table 4. Outliers with high identity in the Warcup ITS training set v2.

Columns are as for Table 3. Notes. §1. Species *Sporisorium destruens* is assigned to genus *Sphacelotheca* in accession AY344979. §2. Species *Microsphaeropsis proteae* is assigned to genus *Ascochyta* in accession JN712495. §3. Species *Trichoderma ghanense* is assigned to genus *Hypocrea* in accession HQ260622. §4 Species *Meruliporia incrassata* and *Serpula incrassata* are both assigned to genus *Serpula* (by accessions EU429519 and HM135647, respectively).

%Id	*LCR*	*Accessions*	*Highest Distinct Taxa*	*Genus*
98.1	Phylum	AB680539 Y10109	c:Betaproteobacteria c:Alphaproteobacteria	*Aquaspirillum* *Magnetospirillum*
99.9	Class	Y10755 AB021399	o:Xanthomonadales o:Pseudomonadales	*Xanthomonas* *Pseudomonas*
99.6	Class	AJ007800 AJ227812	o:Caulobacterales o:Sphingomonadales	*Asticcacaulis* *Sphingomonas*
99.8	Order	AY345990 FR733691	f:Entomoplasmataceae f:Spiroplasmataceae	*Entomoplasma* *Spiroplasma*
99.1	Order	AJ006086 EF114311	f:Planococcaceae f:Bacillaceae_1	*Solibacillus* *Bacillus*
100.0	Family	GU256437 KC109787	g:Nitrospirillum g:Azospirillum	*Nitrospirillum* *Azospirillum*
99.9	Family	FR733721 X79289	g:Nocardia g:Rhodococcus	*Nocardia* *Rhodococcus*

Table 5. Outliers with high identity in RDP 16S rRNA training set v16.

Columns as for Table 3.