Published online 22 August 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 18* ***e120***
doi:10.1093/nar/gkn491

Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers

Zongzhi Liu[1], Todd Z. DeSantis[2], Gary L. Andersen[2] and Rob Knight[1,*]

[1]Department of Chemistry and Biochemistry, UCB 215, University of Colorado at Boulder, Boulder, CO 80309-0215 and [2]Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 70A-3317, Berkeley, CA 94720, USA

Received June 16, 2008; Revised July 15, 2008; Accepted July 16, 2008

ABSTRACT

The recent introduction of massively parallel pyrosequencers allows rapid, inexpensive analysis of microbial community composition using 16S ribosomal RNA (rRNA) sequences. However, a major challenge is to design a workflow so that taxonomic information can be accurately and rapidly assigned to each read, so that the composition of each community can be linked back to likely ecological roles played by members of each species, genus, family or phylum. Here, we use three large 16S rRNA datasets to test whether taxonomic information based on the full-length sequences can be recaptured by short reads that simulate the pyrosequencer outputs. We find that different taxonomic assignment methods vary radically in their ability to recapture the taxonomic information in full-length 16S rRNA sequences: most methods are sensitive to the region of the 16S rRNA gene that is targeted for sequencing, but many combinations of methods and rRNA regions produce consistent and accurate results. To process large datasets of partial 16S rRNA sequences obtained from surveys of various microbial communities, including those from human body habitats, we recommend the use of Greengenes or RDP classifier with fragments of at least 250 bases, starting from one of the primers R357, R534, R798, F343 or F517.

INTRODUCTION

Pyrosequencing (1) has the potential to revolutionize our ability to understand the microbial world. Because the vast majority of microbes have not been cultured, and perhaps cannot be cultured using existing techniques (2), culture-independent techniques that analyze small subunit ribosomal RNA sequences (16S rRNA in the case of bacteria and archaea), amplified by PCR directly from environmental samples, are critical for understanding the composition and dynamics of complex microbial communities (3). Pyrosequencing greatly increases the rate at which we can characterize microbial communities for several reasons. Because pyrosequencing is a single-molecule technique, the heterogeneous 16S rDNA PCR products can be characterized directly without cloning, thus saving an immense amount of labor. The cost is also much less than with traditional methods such as capillary (Sanger) sequencing. Finally, linking a unique sequence barcode to a 16S rDNA-directed PCR primers allows dozens to hundreds of uniquely tagged samples (4), e.g. 16S rRNA samples from different microbial communities (5,6), to be analyzed in a single run (multiplex barcoded pyrosequencing).

One limitation of pyrosequencing is the short read lengths: ~100 bases for the original GS 20 instrument, ~250 bases for the current GS FLX platform and an anticipated ~400 bases for the next-generation GS XLR instrument. We previously demonstrated that these short read lengths could be used for phylogenetic-based comparisons of communities based on UniFrac (a metric based on the extent of branch length that any two communities share on a phylogenetic tree constructed from all reads from all communities that are being analyzed) (7–10). However, these analyses rely solely on phylogenies, which are necessarily approximate on datasets of this scale, and taxonomic information is lost. Taxonomic information is critical for relating findings about each community to what is known about the lifestyles of each kind of microbe at varying levels of resolution. For example, at the species level, *Bacillus anthracis* is a pathogen, whereas *B. cereus* is not; at the phylum level, cyanobacteria tend to be photosynthetic primary producers whereas bacteroidetes are heterotrophs, and an increased firmicute:bacteroidetes ratio is associated with obesity (11–13). Consequently, testing whether the short read

[*] To whom correspondence should be addressed. Tel: +1 303 492 1984; Fax: +1 303 492 7744; Email: rob.knight@colorado.edu

lengths that are produced by pyrosequencing could be used for taxonomic assignment is an important task.

Several different methods for taxonomic assignment have been proposed. The vast datasets produced by pyrosequencing prevent the application of standard phylogenetic techniques such as likelihood- or parsimony-based tree reconstruction because there are too many sequences and too many possible trees to search, and thus fast but approximate techniques are required. Similarity searching (e.g. BLAST) can be used to find the closest matches to each sequence. The speed of this technique can be improved by examining reduced representations of the sequence (such as the abundance of each subsequence of a defined length), albeit with some cost in accuracy. Tree-based methods, in which a tree of reference sequences is constructed, and internal nodes are assigned to taxa based on known tip taxon assignments, might be expected to give more accurate assignments (14,15). In this study, we compare several methods: BLAST (16), the online Greengenes (17) and RDP (18) classifiers, and two tree-based methods. For each of the tree-based methods, we used several algorithms to assign the sequences using the tree, thereby allowing us to compare a broad range of taxonomic assignment methods against one another. The results have implications for designing, executing and interpreting large-scale surveys, such as those that will emanate from the recently launched International Human Microbiome Project (19).

MATERIALS AND METHODS

Datasets

The reference taxonomy and sequences were taken from Bergey's manual release 7.8 from RDP. These sequences were used for 'leave-one-out' evaluation of each method, in which each sequence is excluded from the dataset and classification is performed using the remaining sequences. The three bacterial 16S rRNA datasets (only near full-length sequences) used for the analysis of unclassified sequences from community samples were the same as those used in ref. (7) and included: (i) sequences from the distal gut (ceca) of 19 C57Bl/6J lean and obese (*ob*/*ob*) mice (total of 3732 sequences, 3453 unique, with an average length of 1161 bases) (11); (ii) 16S rRNA sequences from microbial communities occupying five locations along the length of the colon, plus stool, from three healthy, unrelated adult humans (total of 11 864 sequences, 7761 unique, with an average length of 1349 bases) (20) and (iii) sequences obtained from communities positioned at 10 different depths (0–60 mm) in the hypersaline Guerrero Negro microbial mat (11 738 sequences, 11 164 unique, with an average length of 1233 bases; Harris,J.K., Walker,J.J. and Pace,N.R., unpublished data; see ref. (21) for details about this mat located in Baja, Mexico).

Clipping

Unique near full-length 16S rRNA sequences were aligned using the NAST web tools (22) and the Greengenes Core Set (17) (12 January 2008 release). Parameters used were: minimum identity, 75%; minimum length, 50% of the length of the sequence (either full length or clipped). Clipped sequences were generated by extending forward from each primer (relative to the orientation of the primer) 100, 250 or 400 bases on the original, ungapped sequences. Primers used in this study were the same as those described in ref. (7). Sequences that had unknown bases at the primer starting point or were not long enough to clip to the desired size were excluded from the analysis. Sequences that were unique as full-length sequences but identical when clipped were consolidated into a single record, associated with the number of times they corresponded to full-length sequences.

Calculating lineage recovery and coverage

Each taxonomy assignment method produces lineage assignments at the levels of domain, phylum, class, order, family and genus, both for the original sequences and for the clipped sequences. Because the true taxonomy is unknown for most environmental samples, taxonomy assignments for the clipped sequences were compared to those obtained using the same method for the original near full-length sequences in order to assess accuracy. The procedure thus tests reproducibility of each method on the clipped sequences. If five different full-length sequences all produced the same clipped sequence, and that sequence was incorrectly assigned at the family level, we would count this as five incorrectly assigned sequences rather than one.

For each taxonomy rank, coverage was defined as the number of taxa assigned to both the original and clipped sequences, divided by the number of taxa assigned to the original sequences (in some cases, taxa at a given level were assigned to clipped but not full-length versions of the same sequence). Recovery was defined as the fraction of the covered clipped sequences assigned to the same taxon as the corresponding full-length sequences.

Assigning lineages to sequences

We used five different methods to assign taxonomy to each sequence (Figure 1). The first three of these were tree independent; the remaining two were 'tree-aware'. Each method provided for each sequence a single putative taxonomy assignment at some level of resolution, although the level often differed among the methods for a given sequence. For example, one method might give a genus-level assignment, whereas another might give only a phylum-level assignment for the same sequence.

All five methods required a tree representing the taxonomy in order to map an original lineage assignment to the levels of Domain, Phylum, Class, Order, Family and Genus. To build this taxonomic tree, we obtained ranks from Bergey's Manual release 7.8. In the tree, taxa at each level of the taxonomy were considered to branch from the level above as a polytomy; e.g. all the classes within the firmicutes, such as the bacilli and the clostridia, branched from the firmicutes. The taxonomic tree contained only the levels described above; intermediate levels such as suborder were excluded.

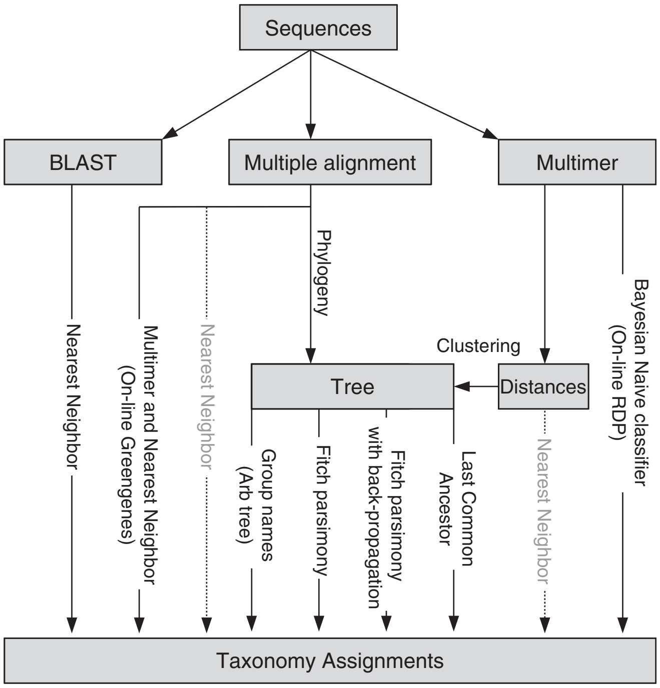

Figure 1. Overview of different methods for taxonomy assignment (see text for details).

Method 1—BLAST. In this procedure, we used BLAST (NCBI's blastall version 2.2.15) (16) to search each sequence against the set of sequences in version 7.8 of Bergey's manual as downloaded from RDP (note that this list only includes cultured microbes). First, we built a BLAST database from all the original full-length sequences. Then we BLASTed each clipped or full-length sequence against the database (using an *E*-value of 1e-50 for 'leave-one-out' evaluation, in which we excluded one sequence from the dataset and asked how well it could be classified based on the remaining data, and 1e-20 for real data). From the BLAST results, we determined the 'nearest neighbors' by highest bit score: 'more neighbors' were determined using a maximum permitted difference of 10% from the maximum bit score. The query sequence itself was excluded from the hit list for leave-one-out evaluation. Finally, from the selected hits, we identified the relevant lineages, and chose the 'common lineage' by starting at the highest level classification and working down until the classifications disagreed. Alternatively, we chose the 'major lineage' where two-thirds of the classifications agreed.

Method 2—online RDP classifier. In this procedure, we submitted the query sequences to the online RDP classifier at http://rdp.cme.msu.edu/classifier/classifier.jsp (18). We parsed the detailed classification results, using only taxon assignments with ≥50% bootstrap support: allowing less bootstrap support than this resulted in very poor recovery (data not shown). We then converted each assignment to the standard levels in the taxonomic tree.

Method 3—online Greengenes classifier based on NAST alignment. In this procedure, we submitted a NAST alignment of the query sequences to the online Greengenes classifier at http://greengenes.lbl.gov/Classify (17). At the time we performed the comparison, the last database update was 1 December 2008, and the database contained 188 073 aligned 16S rRNA records of at least 1250 bases. We parsed the taxonomy assignments in the Greengenes output using the RDP taxonomy, and converted each assignment to domain, phylum, class, order, family and genus using the taxonomic tree and the thresholds suggested in ref. (23).

Method 4—phylogenetic tree-based method: building tree from the NAST alignment followed by a tree-based algorithm. In this procedure, we aligned the query sequences using NAST (75% identity over ≥50% of the sequence length). We then applied the mask Lane MaskPH (24) to the alignment, and used Clearcut 1.0.7's relaxed neighbor joining method (25) to build the tree. Taxonomy was assigned using the nearest ancestral node in the tree that had a defined taxonomy assignment. We also performed a test of this procedure using the ARB parsimony insertion procedure (26) to build the phylogeny, which is a method traditionally used when performing these analyses manually. For this test, we inserted the fragments into the Greengenes core set tree (2006 release), mapped the sequences back to ancestral nodes in the Hugenholtz taxonomy, and converted the names to the RDP taxonomy for comparison. Because ARB cannot be easily automated, we focused on Clearcut for large-scale comparisons.

Method 5—multimer clustering tree-based method: building trees from multimer counts followed by a tree-based algorithm. Multimer clustering has been widely used for taxonomic assignment of metagenomic sequences: its great attraction is that it does not require a multiple sequence alignment, so can potentially avoid the computational cost of the alignment step (27,28). In this procedure, we first calculated the frequencies of each of the overlapping multimers (in this case, 5-mers) for all sequences. Any multimer containing an 'N' or other ambiguous base was excluded from the analysis. We then built a Bray–Curtis distance matrix (where distance is defined as the sum of the *k*-mer frequency differences divided by the sum of all *k*-mer frequencies) from the matrix of multimer counts. We then built a relaxed neighbor joining (NJ) tree from the distance matrix using Clearcut, and assigned the taxonomy based on the nearest ancestral node in the tree that had a defined taxonomy assignment. The tree-building algorithm was adapted from that described in ref. (27).

Three algorithms for inferring the lineage from the tree topology

For each of the tree-aware methods, we evaluated three algorithms for inferring the lineage using the tree. The input to each algorithm consisted of (i) a tree (either a phylogenetic tree built from the alignment using Clearcut or a UPGMA clustering of a distance matrix based on multimer counts) containing both the query sequences and the database sequences, (ii) a taxonomic tree containing all the taxon names as nodes and (iii) a mapping of each database sequence to a node in the taxonomic tree.

Algorithm 1—last common ancestor. In this method, we first assigned taxonomic information to each tip in the tree that represented a sequence in Bergey's taxonomic database. We then performed a postorder tree traversal (i.e. visiting each node after visiting all the nodes descending from that node), and assigned the taxonomic information for each internal node to the node in the taxonomic tree that was the last common ancestor (LCA) of all taxonomic tips that descend from that internal node. For instance, an internal node that had descendants from the taxonomic database from both the classes Bacilli and Clostridia within the phylum firmicutes would be assigned only to the phylum level (firmicutes), since that would be the LCA of these two sequences in the taxonomic tree. For each query sequence with unassigned taxonomy, we traced back to the nearest ancestor with a valid node assignment in the taxonomic tree (assigned from at least two children with taxonomic nodes assigned). We then assigned the lineage from this taxonomic node.

Algorithm 2—Fitch parsimony. We performed a postorder traversal of the phylogenetic tree, calculating the states of internal nodes using the Fitch parsimony algorithm (29). To assign states to query sequences, which were stored as leaves in the tree without assigned taxonomies, we traced back to the nearest ancestor that had a valid taxonomy assignment (assigned from at least two children). We then assigned taxonomic categories from phylum to genus, stopping at the level of detail at which the category became ambiguous.

Algorithm 3—Fitch parsimony with back-propagation. The approach was the same as Fitch above, except that after the postorder Fitch assignment, we performed a preorder traversal to set ambiguous status (if the parent state was unambiguous, and the state of one of the children is the same as the parent, we assigned the unambiguous status to the current node). When ties occurred (i.e. two assignments were equally good), we chose one of the assignments at random. This method allowed more nodes to be assigned unambiguously at the expense of some inaccurate assignments.

Analyses were implemented using the PyCogent toolkit (30).

RESULTS

The overall strategy of our analysis was as follows. First, we performed leave-one-out analysis using full-length sequences where the taxonomy is known to ask how well each method predicts the taxonomy of one sequence from the remaining sequences in the database. Second, we used large, empirical datasets to check the internal consistency of each method in assigning taxonomy to full-length and clipped sequences. Third, we examined the ability of each method to consistently determine the proportion of taxa of each type in a given sample using clipped sequences (i.e. to check whether the errors in assignment cancel out, so that the number of sequences of a given type is correct even when there are errors in which sequence is placed in which class). Finally, we collected information about the relative speed of the methods. In general, different regions of the 16S rRNA gene differed greatly in their ability to recapture the taxonomic information inferred from full-length gene sequences. Moreover, different methods varied greatly in stability and reliability. Because gold-standard taxonomic information is difficult to obtain, we primarily relied on internal consistency of the methods between clipped and full-length sequences.

Figure 2 shows the 'leave-one-out' evaluation. In this analysis, we withheld one full-length sequence from the database, and asked how effectively each of the methods is able to place that sequence. We use only sequences classified according to version 7.8 of Bergey's Manual. The points on each line in this figure are, from left to right, genus, family, order, class, phylum and domain (marked a spectrum of colors from red to blue). As expected, classification at the domain level is excellent (essentially 100% coverage and recovery for all methods, top-right cluster of dark blue points), whereas classification at the genus level is both much more variable and much less accurate (red points scattered at bottom and left). One important point is that excluding sequences that are the only representatives of their genera from the analysis has a large impact on the results (the gray arrows in Figure 2 show examples of the effects of excluding single-sequence genera from the analysis). For example, at the family level, accuracy increases from 96.0% to 97.8% using the BLAST method and the 'major lineage from more neighbors' criterion, and from 97.7% to 98.5% using the multimer clustering tree-based method with 5-mer fragments (see Methods section for descriptions of these approaches). Although these may seem like relatively small differences, because most of the methods perform well in leave-one-out analysis, they represent up to a doubling of the error rate and can change the conclusions about the relative accuracy of different methods.

Of the BLAST-based approaches, the 'common lineage from nearest neighbors' (blue line) approach always returned a result, but was always less accurate (lower recovery rate) than the other methods. The 'common lineage from more neighbors' (black line) approach always had a higher accuracy, but lower coverage (at all levels) than the 'major lineage from more neighbors' (green line) approach (see Methods section for descriptions of these algorithms). However, none of the BLAST-based approaches performed as well as the tree-based approaches (i.e. the black line on the right panel). The error rates of these methods are comparable to those previously reported for the RDP classifier, suggesting that the RDP classifier does not perform uniquely well. In the following analyses, single-sequence genera were always included.

Figure 3 shows the effects of different regions of the sequence on the 'leave-one-out' analysis using the different methods. Recovery (defined as the fraction of sequences given their correct assignment) and coverage (the fraction of sequences for which an assignment could be made) at the genus level are shown. Of the three algorithms used to assign taxonomy from a phylogenetic tree,

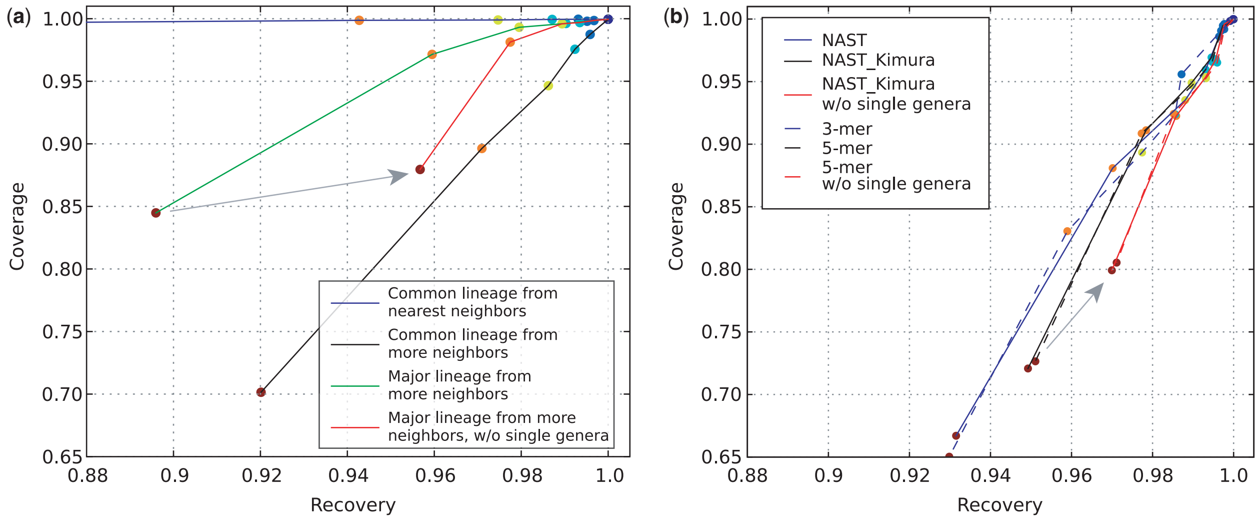

Figure 2. 'Leave-one-out' evaluations of full-length sequences from Bergey's Manual. The *x*-axis shows recovery (i.e. fraction of sequences given their correct assignment). The *y*-axis shows coverage (i.e. fraction of sequences for which an assignment could be made using each method). Each line represents the assignments of a chosen method at different ranks. Each colored point represents a rank (blue to red correspond to levels from domain to genus). Gray arrows indicate effect of including/excluding sequences that are the sole representative of their genera. (**a**) BLAST methods. See the text for 'nearest neighbors', 'more neighbors', 'common lineage' and 'major lineage'. (**b**) Tree-based methods followed by Fitch parsimony assignment. 'NAST', 'NAST_Kimura' are phylogenetic tree-based methods that build the relaxed NJ tree from NAST alignments. With 'NAST_Kimura', a Kimura adjustment was applied to the distance matrix before tree building. '3-mer', '5-mer': multimer clustering tree-based method that builds the relaxed NJ tree from a Bray–Curtis distance matrix obtained from the multimer (3-mer or 5-mer) count matrix.

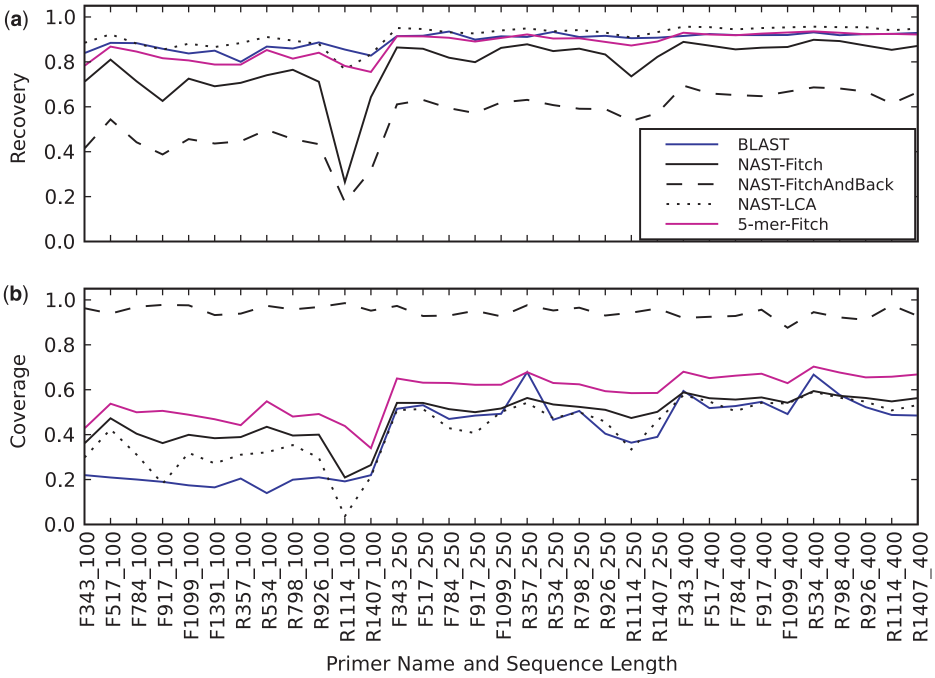

Figure 3. 'Leave-one-out' recoveries at the genus level for clipped sequences from Bergey's Manual. The *x*-axis shows the primer and the length of the read. The *y*-axis shows recovery (**a**) or coverage (**b**) for each method. Recovery and coverage are defined as in Figure 2. Each method is represented as a line. 'BLAST', BLAST method using 'common lineage for more neighbors'; 'NAST-Fitch', 'NAST-FitchAndBack', and 'NAST-LCA': these phylogenetic tree-based methods build trees from NAST alignments with Kimura correction, followed by Fitch parsimony, Fitch parsimony with back-propagation, and last common ancestor algorithm, respectively. '5-mer-Fitch': multimer clustering tree-based method using Fitch parsimony algorithm (the same with '5-mer' in Figure 2).

the Fitch parsimony with back-propagation algorithm essentially always produced a result, but the recovery was invariably worse; the last common ancestor algorithm is more conservative, increasing recovery rate at the expense of poor coverage; while the Fitch parsimony algorithm behaves in the middle. The BLAST method had consistently low coverage using 100-base sequences (19% of sequences classified on average). However, coverage increased markedly using 250- and 400-base sequences (average of 49% of sequences classified, essentially no difference between 250- and 400-base reads). Compared with the phylogenetic tree-based method with the tree built from the NAST alignment, the multimer clustering method using trees built using similarities in the counts of five-base oligomers had better coverage (average 63% for 250 bases), and averaged 91% recovery (250 bases) using the same assignment algorithm (Fitch parsimony). However, the multimer clustering tree-based approaches performed very poorly on real datasets (data not shown). Therefore, the NAST alignment-based tree methods were used in the following analyses. One important point is that most methods are sensitive to the region of the 16S rRNA gene being sequenced. For example, for the 250-base sequences, the R357 primer (which when used together with the 8F primer encompasses the V2 and V3 regions of the gene) is the best primer across all the methods.

Recovery and coverage of six methods is shown using three different datasets at the genus level (Figure 4a and b) and at the phylum level (Figure 4c and d). In each case, recovery and coverage were measured relative to the results of the full-length sequence, as the true classifications of these sequences are unknown. However, based on the misclassification rates presented in Figure 2 from the leave-one-out evaluation, we expect these classifications to be at least 95% accurate (meaning that 95% of the sequences should be placed in the correct phylum). In general, the region of the 16S rRNA gene targeted for sequencing had a larger effect on the taxonomic assignment than the method used for taxonomic classification, and regions that performed poorly often did so with different length reads. For 100-base reads, F517, R534 and R798 performed especially well at the genus level, with the best of the methods (>90% recovery and coverage). For 250-base reads and above, these regions were joined by R357 and F343. Interestingly, the V6 region as described in ref. (31) performs poorly relative to the other regions, as do the regions that overlap V6 such as R1114 (with 250-base reads) and F1099, reaching recovery levels below 10% on the gut datasets. Results at the phylum level mirrored results at the genus level, although the errors were not quite so pronounced.

The phylogenetic tree-based methods ('Fitch', 'FitchAndBack' and 'LCA'), which are based on trees built from NAST alignments and represented by the black lines in Figure 4, are much more sensitive to the region sequenced than other methods. In general, they performed poorly on the gut datasets (especially the mouse gut) but were among the best methods on the Guerrero Negro dataset (see, for example, the Fitch parsimony results with F517-100 and F343-250). The Greengenes classifier outperformed the other methods both in terms of coverage and accuracy under almost all conditions, and BLAST and the RDP classifier performed about equally well (although for specific regions and datasets, one of these methods could greatly outperform the other: e.g. with R534 and 250 base clips, the RDP classifier had an error rate of ~20% on the mouse dataset, whereas BLAST's accuracy was essentially 100%). One encouraging result is that with a good choice of primer, taxonomy assignment can be recovered as well by a short sequence as by a long sequence (for example, 100-base reads from F517).

Thus far, we have discussed the consistency of taxonomic assignments for full-length versus clipped sequences. An equally important question is how consistent the overall proportion of each phylogenetic group is after applying each of the procedures. Figure 5 shows the phylum-level compositions produced by the different methods on the same three datasets. As we would expect from previous analyses of full-length 16S rRNA sequences, the human and mouse gut datasets are dominated by the firmicutes and bacteroidetes phyla, whereas the Guerrero Negro dataset is more diverse and has more unclassified phyla [not in the present taxonomy as they were newly discovered in that environment (14)]. The Greengenes classifier produces essentially perfectly consistent ratios of the members of the different phyla no matter which sequence fragment is used, suggesting that the misclassifications seen in Figure 4 tend to cancel out. The RDP classifier similarly performs extremely consistently in the gut (a community that is characterized by a high level of species- and strain-level diversity), accuracy is somewhat diminished in regions that overlap V6. BLAST generally produces far more unclassified sequences, especially in the mouse gut sample, and is more sensitive to the details of the 100-base reads (however, it stabilizes with reads of 250 bases and above).

In contrast, the three tree-based approaches are much more sensitive to the region of sequence examined. Extra caution is necessary when these methods were used for taxonomy assignment. For example, with one of the worst regions, reading forward 250 bases from F1099, the reported frequency of euryarcheota can be up to 39% (considerably more than is reasonable from current estimates of the representation of this group). However, certain regions, such as F343 and R357 with 250-base reads, recapture the results of the more stable methods almost exactly.

One common procedure, often manually applied, is to import the alignments into ARB (20), put the sequences into a standard tree using ARB parsimony insertion and then use either the group names in the standard tree (based on the Hugenholtz taxonomy) or the Fitch parsimony status from known tip taxa to infer the taxonomy. Figure 6 shows the results of performing this procedure using 100-base fragments (because the procedure is extremely time-consuming on datasets of this size, we did not repeat the procedure for the other fragment lengths). Coverage and accuracy are within the range of the methods shown in Figure 4, ~100% at the phylum level on the combined dataset and somewhat lower at the family/class level. Again, the regions close to V6 (e.g. R1114, 100 bases) perform exceptionally poorly. Another important conclusion is that the 'group name' algorithm (see Methods section) performs no better than the Fitch parsimony algorithm, so the latter can be used to perform equivalent analyses in a fully automated way.

Finally, we compared the speed of the different methods: all methods were run locally on the same hardware except RDP, NAST alignment and

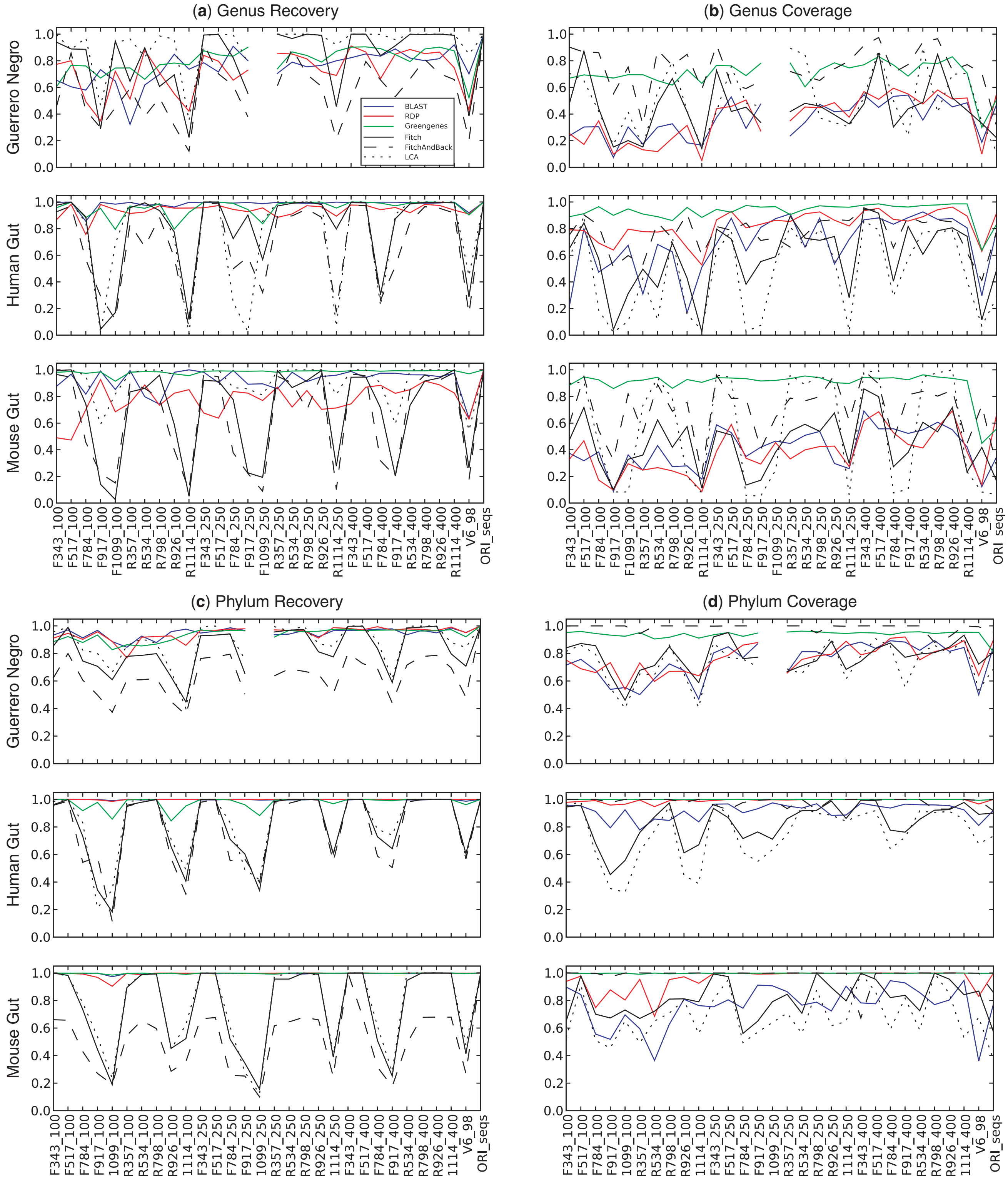

Figure 4. Recoveries and coverage at the genus level (**a** and **b**) and phylum level (**c** and **d**) for each of the three datasets: the Guerrero Negro microbial mat, the mouse gut and the human gut. The legend for the series in the first panel applies to all panels. Each line represents the performance (recovery or coverage) of one method on one dataset. The *x*-axis represents primer name and sequence lengths. Apart from the coverage of 'ORI_seqs', which is the fraction of full-length sequences with an assignment at a certain rank, recovery and coverage are measured relative to the results of the full-length sequence. Missing data points are for reads that extend past the length of the near full-length amplicons used for this study. Recovery and coverage are defined as in Figure 2.

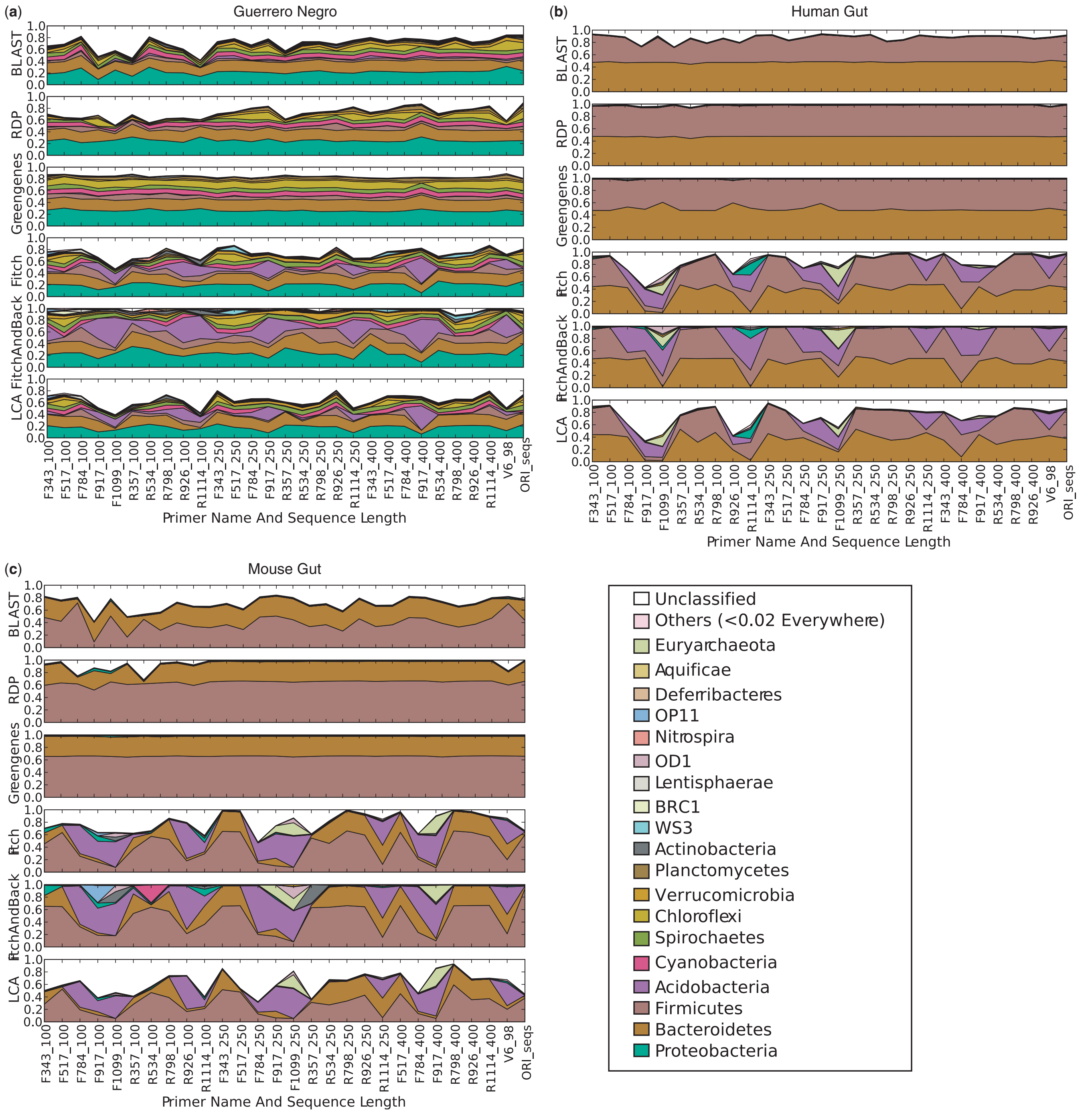

Figure 5. Compositions at the phylum level for each of the three datasets: (**a**) Guerrero Negro mat, (**b**) Human gut and (**c**) Mouse gut, using a range of different methods (separate subpanels within each group). The *x*-axis of each graph shows region sequenced. The *y*-axis shows abundance as a fraction of the total number of sequences in the community. The legend shows colors for phyla (consistent across graphs).

Greengenes, which were run at the respective websites. Figure 7 shows the results. The RDP method is by far the fastest (~80 s for 1000 full-length sequences), and Greengenes the slowest (~3000 s for 1000 sequences): the other methods performed at intermediate speeds. The BLAST methods can be accelerated by running multiple tasks in parallel. In general, parsing the files and/or post-processing the trees took negligible time compared to building trees and/or performing the classification task. In general, performance was affected substantially by the lengths of the sequences, but as expected, scaled approximately linearly with the number of sequences.

DISCUSSION

Overall, our analyses show that taxonomic assignment is highly sensitive to the region of the 16S rRNA

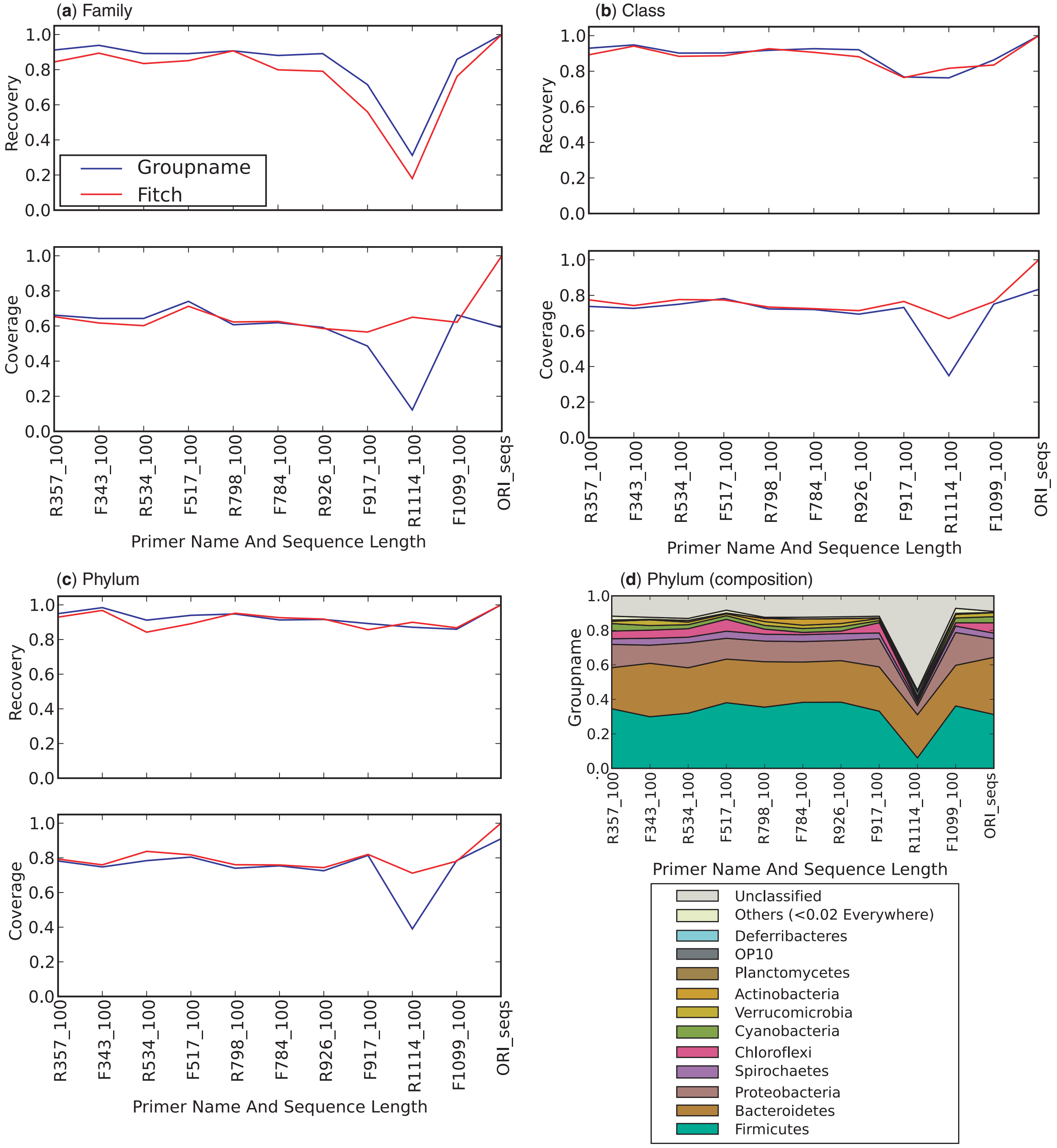

Figure 6. Comparison of recoveries and coverage using ARB and either the group name or Fitch parsimony criteria for grouping sequences. The *x*-axis of each graph shows the region of the gene encompassed by the sequence (all 100-base clipped sequences). The *y*-axis plots either coverage or recovery, defined as in Figure 2. Results are shown for (**a**) family, (**b**) class and (**c**) phylum. (**d**) Compositions at the phylum level obtained using the Group Name method for the combined dataset (i.e. Guerrero Negro mat, mouse gut and human gut).

gene sequenced, to the assignment method used, and in some cases, to the length of the region sequenced. However, the results are encouraging: most regions, no matter how short, provide stable estimates of the abundance of each phylum in the dataset, which is all that is reported in many 16S rRNA sequence-based analyses. Thus, even results collected using earlier generation pyrosequencers (GS20 with average read lengths of 80–100 bases) should be stable to future re-analyses provided that the Greengenes or RDP classifiers are used. Similarly, many regions of the 16S rRNA gene, e.g. 100-base reads forward from positions 343, 517, 784, or backwards from positions 357, 534, 798 or 926 relative to the *Escherichia coli* sequence, allow excellent coverage

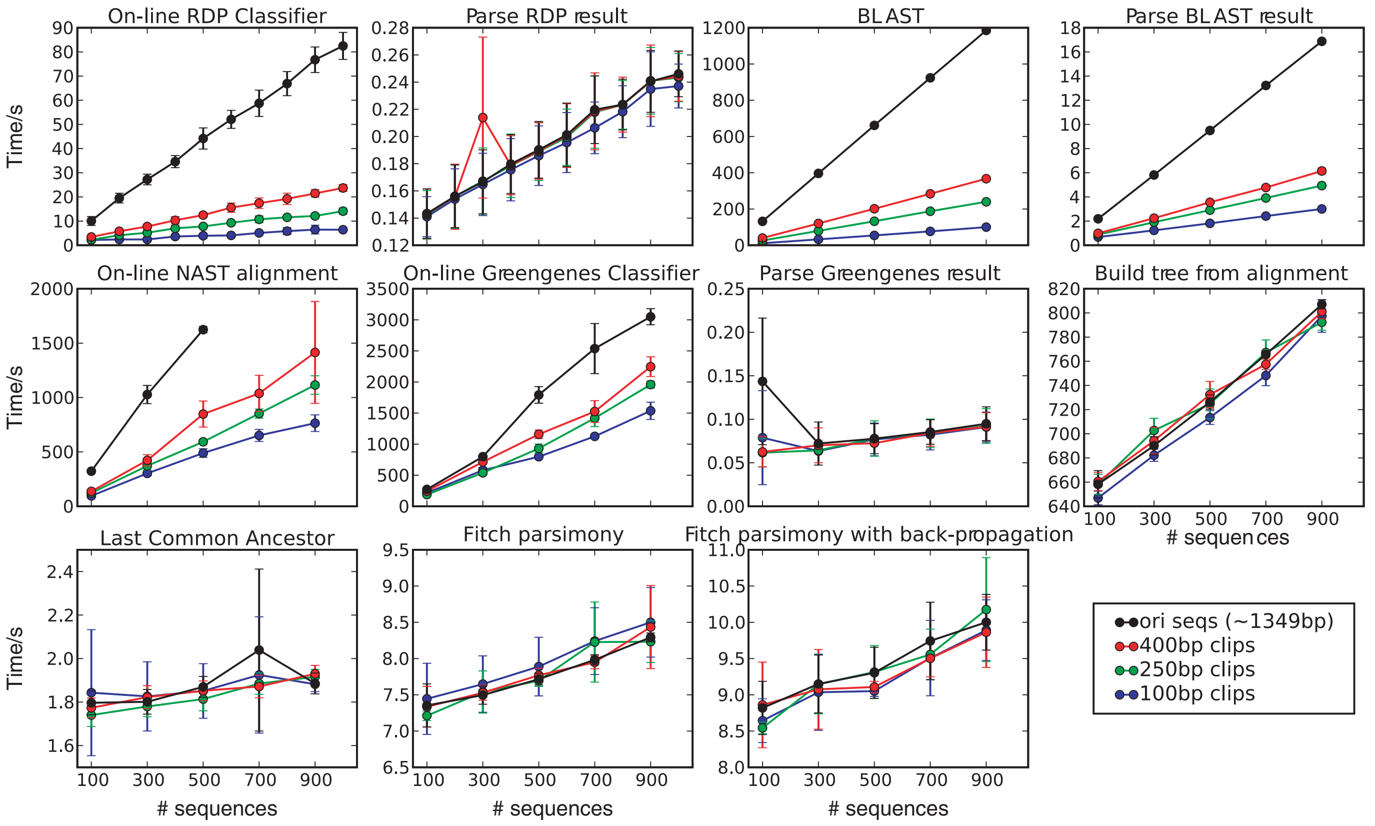

Figure 7. Run time performance of the different methods as a function of the number and length of sequences. The *x*-axis plots the number of sequences, the *y*-axis time (in seconds). The legend shows colors for length of sequence. The error bars represent SDs from 10 replicates.

and recovery even at the genus level with relatively short reads. Moreover, reads produced by the current GS FLX apparatus (average of 250 bases) will for the most part hold up to re-sequencing using the newer XLR instrument, which produces longer reads. In general, the Greengenes and RDP classifiers produce highly stable and accurate results even on very disparate datasets, so we can consider taxonomy assignment tools mature in that sense.

One important cautionary note about the leave-one-out results using Bergey's manual is that different groups of bacteria have been studied at different levels of effort (e.g. human-associated commensals and pathogens, and their close relatives, have been far more extensively studied than other taxa), and only organisms that can be cultured are included. Both of these factors substantially bias the phylogenetic representation, and thus Bergey's provides an extremely biased test set. As Figure 2 shows, exclusion of single-sequence genera can have large effects both on coverage and recovery, e.g. increasing the apparent error rate of family-level assignments by 50–100% depending on the method. All analyses we performed, except those indicated in Figure 2, included the single-sequence genera.

Our analysis suggests that the V6 region is not optimal for pyrosequencing analyses that are directed at taxonomic assignment, as opposed to measuring levels of diversity. These results are consistent with our previous analysis of pyrosequencing and community clustering (7), which showed that regions overlapping V6 (R1114, 100 bases) were much less suitable for community clustering than other regions, such as R357. Therefore, we recommend using primers R357 and F8 to generate 250-base reads starting at R357: this region which spans V2 and V3 performs well for both community clustering and taxonomic assignments in a wide range of datasets (e.g. the mouse and human gut, and the Guerrero Negro microbial mat).

Automated methods that can be incorporated into high-throughput workflows perform at least as well as manual analyses with ARB, suggesting that these manual analyses are no longer required except perhaps for independent confirmation of suspect taxa. We therefore suggest that the Greengenes and/or RDP classifications be treated as sufficient evidence to support taxonomic classifications: although they are not perfect, they achieve high accuracy under a wide range of conditions. However, we recommend that future development should focus on improving the run-time performance, especially of the Greengenes classifier.

Finally, one important consideration with this and all other taxonomic analyses is that the taxonomy assignment for each read is only as good as the underlying taxonomy and the phylogeny on which it is based. New taxa by definition cannot be identified using these types of techniques, although sequences that remain unclassified are often fertile grounds for new lineage discovery. *De novo* tree-building is still required for identifying new lineages, and thus support for improvement of taxonomies and iterative tree-building approaches are essential for maximizing the utility of rapid classification methods for pyrosequencer datasets—datasets that will undoubtedly grow in number and size given the impending explosion in comparative metagenomic surveys of human body as well as terrestrial and oceanic habitats.

ACKNOWLEDGEMENTS

We thank Catherine Lozupone, Phil Hugenholtz, Frederic Bushman, Peter Turnbaugh and Jeffrey I. Gordon for critical discussion of the article. We thank J. K. Harris, J. J. Walker and N. R. Pace for sharing unpublished data. This work was supported by NIH grant P01DK078669 and the Crohn's and Colitis Foundation of America. Analyses were performed using the W.M. Keck Bioinformatics Facility at CU Boulder. Additional support was provided by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics: GTL Program, through contract DEAC02-05CH11231 (T.Z.D., G.L.A.).

FUNDING

Funding to pay the Open Access publication charges for this article was provided by NIH P01DK078669.

Conflict of interest statement. None declared.

REFERENCES

1. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
2. Rappe,M.S. and Giovannoni,S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.*, **57**, 369–394.
3. Pace,N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science*, **276**, 734–740.
4. Binladen,J., Gilbert,M.T., Bollback,J.P., Panitz,F., Bendixen,C., Nielsen,R. and Willerslev,E. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.
5. Hamady,M., Walker,J.J., Harris,J.K., Gold,N.J. and Knight,R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods.*, **5**, 235–237.
6. McKenna,P., Hoffmann,C., Minkah,N., Aye,P.P., Lackner,A., Liu,Z., Lozupone,C.A., Hamady,M., Knight,R. and Bushman,F.D. (2008) The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog.*, **4**, e20.
7. Liu,Z., Lozupone,C., Hamady,M., Bushman,F.D. and Knight,R. (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res.*, **35**, e120.
8. Lozupone,C., Hamady,M. and Knight,R. (2006) UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, **7**, 371.
9. Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.
10. Lozupone,C.A., Hamady,M., Kelley,S.T. and Knight,R. (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **73**, 1576–1585.
11. Ley,R.E., Backhed,F., Turnbaugh,P., Lozupone,C.A., Knight,R.D. and Gordon,J.I. (2005) Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA*, **102**, 11070–11075.
12. Ley,R.E., Turnbaugh,P.J., Klein,S. and Gordon,J.I. (2006) Microbial ecology: human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.
13. Turnbaugh,P.J., Ley,R.E., Mahowald,M.A., Magrini,V., Mardis,E.R. and Gordon,J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.
14. Barker,D. and Pagel,M. (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput. Biol.*, **1**, e3.
15. Krause,L., Diaz,N.N., Goesmann,A., Kelley,S., Nattkemper,T.W., Rohwer,F., Edwards,R.A. and Stoye,J. (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.*, **36**, 2230–2239.
16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
17. DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
18. Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Bandela,A.M., Cardenas,E., Garrity,G.M. and Tiedje,J.M. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.*, **35**, D169–D172.
19. Turnbaugh,P.J., Ley,R.E., Hamady,M., Fraser-Liggett,C.M., Knight,R. and Gordon,J.I. (2007) The human microbiome project. *Nature*, **449**, 804–810.
20. Eckburg,P.B., Bik,E.M., Bernstein,C.N., Purdom,E., Dethlefsen,L., Sargent,M., Gill,S.R., Nelson,K.E. and Relman,D.A. (2005) Diversity of the human intestinal microbial flora. *Science*, **308**, 1635–1638.
21. Ley,R.E., Harris,J.K., Wilcox,J., Spear,J.R., Miller,S.R., Bebout,B.M., Maresca,J.A., Bryant,D.A., Sogin,M.L. and Pace,N.R. (2006) Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Appl. Environ. Microbiol.*, **72**, 3685–3695.
22. DeSantis,T.Z. Jr, Hugenholtz,P., Keller,K., Brodie,E.L., Larsen,N., Piceno,Y.M., Phan,R. and Andersen,G.L. (2006) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.*, **34**, W394–W399.
23. DeSantis,T.Z., Brodie,E.L., Moberg,J.P., Zubieta,I.X., Piceno,Y.M. and Andersen,G.L. (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb. Ecol.*, **53**, 371–383.
24. Hugenholtz,P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, **3**, REVIEWS0003.
25. Sheneman,L., Evans,J. and Foster,J.A. (2006) Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics*, **22**, 2823–2824.
26. Ludwig,W., Strunk,O., Westram,R., Richter,L., Meier,H., Yadhukumar, Buchner,A., Lai,T., Steppi,S., Jobb,G. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363–1371.
27. Hao,B. and Qi,J. (2004) Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. *J. Bioinform. Comput. Biol.*, **2**, 1–19.
28. McHardy,A.C., Martin,H.G., Tsirigos,A., Hugenholtz,P. and Rigoutsos,I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, **4**, 63–72.
29. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 83–92.
30. Knight,R., Maxwell,P., Birmingham,A., Carnes,J., Caporaso,J.G., Easton,B.C., Eaton,M., Hamady,M., Lindsay,H., Liu,Z. *et al.* (2007) PyCogent: a toolkit for making sense from sequence. *Genome Biol.*, **8**, R171.
31. Huber,J.A., Mark Welch,D.B., Morrison,H.G., Huse,S.M., Neal,P.R., Butterfield,D.A. and Sogin,M.L. (2007) Microbial population structures in the deep marine biosphere. *Science*, **318**, 97–100.