Published online 24 April 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 14* ***e111***
doi:10.1093/nar/gks335

Rapid identification of high-confidence taxonomic assignments for metagenomic data

Norman J. MacDonald, Donovan H. Parks and Robert G. Beiko*

Faculty of Computer Science, Dalhousie University, 6050 University Avenue, PO BOX 15000, Halifax, NS B3H 4R2, Canada

Received December 3, 2011; Revised March 30, 2012; Accepted April 4, 2012

ABSTRACT

Determining the taxonomic lineage of DNA sequences is an important step in metagenomic analysis. Short DNA fragments from next-generation sequencing projects and microbes that lack close relatives in reference sequenced genome databases pose significant problems to taxonomic attribution methods. Our new classification algorithm, RITA (Rapid Identification of Taxonomic Assignments), uses the agreement between composition and homology to accurately classify sequences as short as 50 nt in length by assigning them to different classification groups with varying degrees of confidence. RITA is much faster than the hybrid PhymmBL approach when comparable homology search algorithms are used, and achieves slightly better accuracy than PhymmBL on an artificial metagenome. RITA can also incorporate prior knowledge about taxonomic distributions to increase the accuracy of assignments in data sets with varying degrees of taxonomic novelty, and classified sequences with higher precision than the current best rank-flexible classifier. The accuracy on short reads can be increased by exploiting paired-end information, if available, which we demonstrate on a recently published bovine rumen data set. Finally, we develop a variant of RITA that incorporates accelerated homology search techniques, and generate predictions on a set of human gut metagenomes that were previously assigned to different 'enterotypes'. RITA is freely available in Web server and standalone versions.

INTRODUCTION

Culture-independent sequencing is rapidly filling the gap in our physiological and ecological understanding of the world. For instance, 148 new bacterial phylotypes and over 1.2 million novel genes were found by direct sampling of environmental sequences in the Sargasso Sea (1). Studies of the human gut microbiome have revealed large variation in taxonomic community profiles, but stronger apparent conservation of function at the molecular level (2). Assembly of complete genomes from metagenome samples is typically only feasible for abundant members of a community (3–5), and many organisms in complex communities are represented only by short contiguous regions or singletons. Communities as a whole can be characterized using small-subunit rDNA (SSU or 16S) or an expanded range of taxonomic markers (6–7), but doing so may lead to a skewed view of a microbial community's composition (8–10) and does not reveal the metabolic potential of different members of a population. An alternative is to characterize the functional capabilities of communities by using random shotgun sequencing of DNA or RNA (11,12). An important step in this approach is the assignment of short reads to taxonomic lineages using composition- or homology-based approaches. While genome composition is an informative trait for the classification of DNA sequences, the presence of confounding factors such as G+C variability, novel genomes and inter-group genetic similarities complicates fragment annotation (13).

Homology-search approaches such as BLAST (14), and classifiers such as CARMA (15), TreePhyler (16) and MetaDomain (17) can be used to identify evolutionary relationships through statistical comparison of a query fragment to a set of annotated reference sequences. These methods are highly accurate as long as there is a similar sequence within the reference set, but sequences

*To whom correspondence should be addressed. Tel: +1 902 494 8043; Fax: +1 902 492 1517; Email: beiko@cs.dal.ca

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

that are evolutionarily distant from examples in the reference set can be difficult or impossible to match with homology-based methods. Composition-based classifiers based on the k-nearest neighbors (18), support vector machine (SVM) (19), Naïve Bayes (NB) (20–22), or Markov model (23,24) paradigms compare signatures of short motifs from a query fragment to models capturing the motif signature of a complete genome. These approaches, though inferior to homology when a similar genome is part of the training set (23), have been used to find distant taxonomic signal to classify novel sequences correctly to higher ranks. NB also underpins the classifier used by the Ribosomal Database Project to assign SSU rDNA sequences to taxonomic groups (25).

When novel taxonomic groups (i.e. organisms from genera, families, or even broader groups from which no sequenced representative is available) are present in a metagenome, an additional type of error is possible, where a fragment is classified to a taxonomic group that is too specific (e.g. a member of a novel family being classified to a species of the same order). Although part of the lineage is correct, the overly specific classification could produce misleading estimates of microbial diversity or assign critical metabolic reactions to organisms incorrectly. Parks *et al.* (20) distinguish between *rank-specific* approaches that target one rank (e.g. genus or phylum) for all classifications, and *rank-flexible* methods that can make precise predictions when strong evidence is available, but can fall back to higher ranks when such evidence is lacking. Automatically choosing the appropriate rank of a classification is a more difficult problem, but is essential when dealing with varying degrees of novelty; otherwise all classifications may need to be made at a conservative rank such as phylum, even though identification of individual genera may be possible for some fragments. MEGAN (26) uses a lowest common ancestor (LCA) approach to assign rank-flexible taxonomic classifications to DNA sequences analyzed by a rank-specific classification algorithm, in this case the assignment of taxonomy based on best BLASTX matches. TACOA (18) classifies sequences to an appropriate rank based on evidence from a modified nearest-neighbor classifier. PhyloPythiaS (19) also provides rank-flexible attributions by assigning fragments to the LCA in agreement between linear SVMs trained on genomic fragments of different lengths.

Shorter DNA sequence lengths present a greater challenge to classifiers and require more-careful modeling of composition and homology to yield accurate classifications (15,27). Sequencing technologies that generate many short reads are now very common: for example, GS-FLX sequencing was used to study the diversity of a glacier metagenome, resulting in a set of 1 076 539 reads with a median length of 243 nt (26). The bovine rumen data set of (5), generated with Illumina sequencing, produced paired-end reads of length 36 to 125 nt. Phymm (23) computes Markov models of sequence composition on overlapping *n*-mer windows to learn taxonomic signatures, and was validated using the metagenome of an acid mine drainage environment (3). With 1000-nt fragments and removal of sequences from other genomes of the same species, Phymm yielded accuracies comparable to a best BLAST-based approach, but with shorter sequences the classification accuracy is considerably worse than that obtained using best BLAST. PhymmBL takes a linear combination of log-scores from BLASTN and Phymm in order to achieve higher accuracy than can be obtained by using either approach in isolation.

In this article, we propose a sequential classification pipeline, RITA (Rapid Identification of Taxonomic Assignments), which uses homology- and composition-based approaches to obtain sets of highly confident taxonomic assignments determined by progressively more time-consuming algorithms. Although it is a hybrid approach like PhymmBL, RITA uses our previously developed NB classifier (20) to generate compositional predictions, and uses an optimized choice of homology detection algorithms to accelerate the essential homology assignment step. Furthermore, we present an approach for using 16S rDNA profiles of a community to help estimate the novelty in a microbiome which provides a method for rank-flexible classification of sequences that does not require multiple taxonomically distinct BLAST hits for a fragment. RITA is able to achieve very high specificity on short reads, is robust to sequencing errors, and can deal with novel sequences by classifying them to an appropriate taxonomic rank. We also show that exploiting paired-end information from short reads can substantially increase the accuracy of classifications, reducing the number of sequences that are assigned to implausible groups. Finally, we test a version of RITA that uses USEARCH (27) as an initial filter to reduce the number of BLAST comparisons that need to be carried out, and show similar classifications to the set of gut metagenomes used to propose the enterotype hypothesis (28).

MATERIALS AND METHODS

The RITA pipeline

The RITA pipeline (Figure 1) is a hybrid classifier that, like PhymmBL, uses both homology and compositional information. However, RITA differs from PhymmBL in three critical ways. First, RITA uses the NB classifier developed by (20) in place of the interpolated Markov models of Phymm, which increases the speed over 10-fold with no loss of accuracy. Second, RITA uses an approach that places greater weight on homology-based predictions, first testing whether homology and compositional predictions agree, and then checking whether homology results alone strongly favor one taxonomic label over all others. Third, a rank-flexible version of RITA allows the user to provide a list of expected taxonomic groups (e.g. established from a marker-gene study) to restrict the set of predictions that can be made by RITA.

Homology-based predictions made with BLAST can be very time-consuming, and RITA uses three BLAST algorithms (Figure 1) in order to save computational running time. Discontiguous MEGABLAST (D-BLASTN) is executed first in our pipeline as it is the fastest of the BLAST algorithms. The best hits of D-BLASTN are

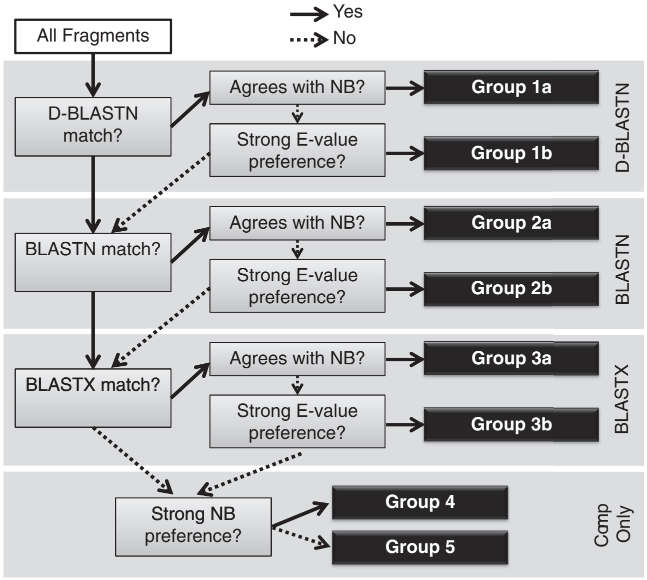

Figure 1. RITA pipeline schematic. Input DNA fragments are sorted into different classification groups based on the outcome of homology searches and compositional classifications.

combined with the best hits of the NB classifier to generate our most confident set, Group 1a. Group 1b is composed of highly confident D-BLASTN matches that do not agree with the best-scoring NB model, since homology-based classification is superior to composition-only classification (Supplementary Figure S1). The threshold for this group is based on the orders of magnitude separating the best overall *E*-value from the *E*-value corresponding to the best model genome from a different taxonomic group. For example, if the top two matches are to genus *Polaromonas* but the third is to genus *Streptococcus*, then the magnitude difference is calculated between the best match and the third match. If this difference exceeds the pre-specified threshold, then the label of the best match is assigned to the sequence, otherwise the sequence is passed to the next step in the pipeline. Group 2a and b are assembled analogously to Groups 1a and 1b, using the BLASTN algorithm. Sequences still unassigned are considered for Group 3a and b which are based on the more-sensitive but slower BLASTX algorithm. Finally, Groups 4 and 5 are composed of the fragments that share no unambiguous homology with our training set. These fragments can be classified with NB and thresholded based on the likelihood ratio of the best prediction to the next-best prediction of a different target rank. The modular nature of the RITA pipeline allows the substitution of different algorithms at different points in the pipeline: in the analysis of the 'enterotype' data set of (29) below, we replaced the three BLAST homology search steps with an approach that first identifies potential homologous matches using UBLAST (30), and then runs BLASTX on the reduced set of candidate matches (see Supplementary Methods section for details).

Genomes, metagenomes and reference models

A total of 2708 completely sequenced and draft bacterial and archaeal genomes were downloaded from the NCBI RefSeq database (31) in February 2011. Subsets of these genomes were selected to generate various validation sets throughout this work. Genomes were merged at the level of species, and used to construct reference homology and compositional models. The homology model consisted of nucleotide sequence databases containing all contiguous sequences from all genomes for D-BLASTN and BLASTN, and a protein sequence database containing all conceptually translated, predicted open reading frames from the full set of reference genomes. Individual NB models were constructed for each species. Each model consisted of conditional probabilities of *n*-mers of length 10 estimated by counting observed *n*-mers over all contigs within a species on both the forward and reverse DNA strands (20). *N*-mers containing characters other than A, C, G or T were discarded. Phymm ICMs were built for each species file using the software provided by (23) that was downloaded on 10 May 2010. PhyloPythiaS models were built using version 1.1 of the software provided by (19). Models were trained using two sets of fragments with varying lengths: (100, 200, 300, 400, 500, 1000, 2000, 3000) and the default (1000, 3000, 5000). Results are reported for the set of models obtained using shorter training fragments as these models produce slightly better results. RAIphy models were built using version 1.0 of the software provided by (24).

We used the following data sets for the analyses described in the Results section (see Supplementary Methods section for further details). The artificial 'leave-one-out' data set consisted of 534 reference sequenced genomes from NCBI, covering a total of 334 named species. The glacier metagenome and associated 16S profiles of (28) were used as the basis for a simulated 'pseudometagenome' analysis and the validation of the rank-flexible classifier. Paired-end Illumina data from the bovine rumen data set of (5) were used to test the agreement between paired ends. We also tested the 'enterotype' hypothesis of (29) using both predictions from SmashCommunity (32) and RITA.

Classification strategy

Classification by rank-specific RITA was performed by finding the best-matching compositional (i.e. maximum likelihood) and homology (i.e. best BLAST match) models from among the set of reference models. Rank-specific classification was carried out with a uniform target taxonomic rank; a true positive prediction occurs when the model chosen is a member of the correct taxonomic group at the appropriate rank. Agreement among compositional and homology classifiers is assessed at the target rank We used a maximum *E*-value threshold of 10^{-5} to assign BLAST matches, and set the interval for assigning sequences to Groups 1b, 2b and 3b to 20 orders of magnitude.

Rank-flexible classification with RITA follows the same procedure for finding a best-matching model, but uses additional information from a 16S phylogeny to assign models to appropriate taxonomic groups and ranks (see Supplementary Methods section for details). One 16S rDNA sequence was selected from each genome in our 2708-genome set by best BLAST match with sequences from the Ribosomal Database Project (25) with the following parameters: strain = type; source = isolates; quality = good; size $\geq$ 1200. These sequences, along with the 16S reads associated with a metagenome sample, were aligned and hypervariable regions were masked using mothur 1.15.0 (33). A phylogenetic tree was constructed from the alignments using FastTree 2.1 (34) and rooted between Archaea and Bacteria. The best-matching homology model was used to set the appropriate target taxonomic rank and group for classification; agreement was concluded if the best-matching compositional model belonged to the same taxonomic group.

Sensitivity, specificity, false negative rate and unclassified rate are defined appropriately for multi-class problems (35). We use P_i,TP_i, FP_i, TN_{i}, FN_i and U_i to represent the total positives, true positives, false positives, true negatives, false negative, and unclassified samples of class label i respectively. Measures of accuracy are then defined as follows: sensitivity (Sn_i) = TP_i/P_i, specificity (Sp_i) = $TP_i/(TP_i+FP_i)$, false negative rate (FNr_i) = FN_i/P_i, and unclassified rate ($Urate_i$) = U_i/P_i, where $P_i = TP_i+FN_i+U_i$ is the set of examples for class i. Summary statistics over all class labels of a taxonomic rank are the mean of the rate over each class label.

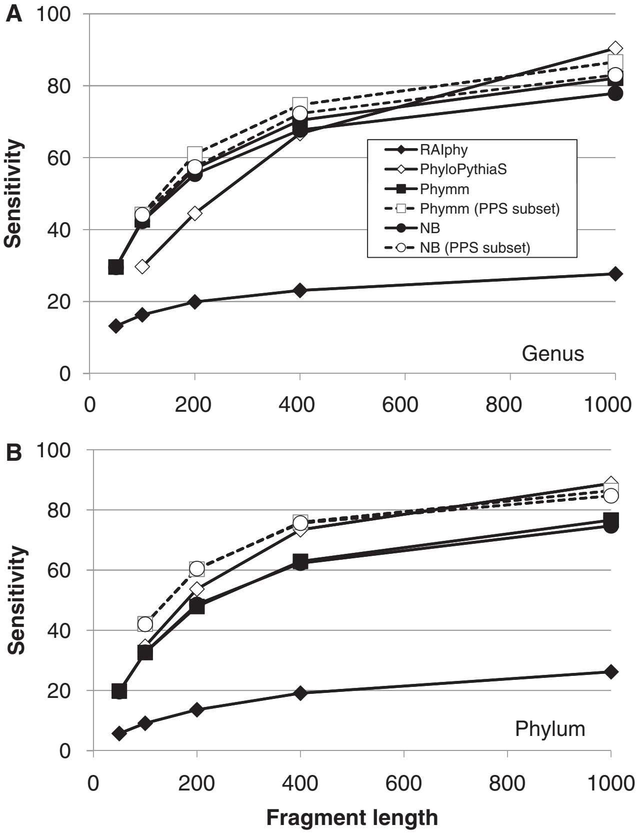

Figure 2. Sensitivity of composition-based classifiers on a synthetic 'leave-one-out' test set. Results are reported for both phylum (**A**) and genus (**B**) predictions. PhyloPythiaS (PPS) is a rank-flexible classifier so may not provide a prediction for all fragments at a given taxonomic rank. To compare against this classifier, the sensitivity of NB and Phymm over the subset of fragments classified by PPS are given.

RESULTS

Leave-one-out synthetic data set

We performed a detailed evaluation of three variants of the BLAST algorithm and several leading composition-based classifiers on a synthetic 'leave-one-out' data set built from 534 completely sequenced genomes. While the genomes in this set do not constitute a real community, using reference genomes allows us to vary the simulated read length, introduce different amounts of error into sequences, and control the degree of taxonomic novelty (and hence the difficulty) in the data set. Our NB classifier and Phymm showed similar sensitivity at different fragment lengths at both the genus and phylum prediction levels, and both were more accurate than RAIphy and outperformed PhyloPythiaS on the subset of fragments that were classified by this algorithm for all fragment lengths <1000 (Figure 2). Furthermore, PhyloPythiaS had a significant advantage in that the same genomes used for testing were present in the training set: this was not the case for either NB or Phymm. Our NB implementation is considerably faster than Phymm, requiring only 1 min for NB versus 26 min for Phymm to classify 33 400 fragments of length 200 nt [see also (20)]. Since the homology and composition-based approaches are complementary and can lead to higher-confidence predictions, we also benchmarked approaches that combine these two types of classifier (see Supplementary Methods section and Supplementary Figures S1–S3). The combined scoring scheme of PhymmBL has lower specificity than the agreement-based classifier defined here largely due to the Phymm classifications being used in isolation whenever there are no BLAST matches. This also leads to a higher sensitivity at the expense of a greatly increased number of false negative classifications.

The accuracy of RITA on the leave-one-out data set is shown at the genus and phylum levels, with self-matching species excluded, in Figure 3a and b and for all levels in Supplementary Tables S1–S4. The only parameters that need to be set in RITA are the E-value interval that defines Groups 1b, 2b and 3b, and the likelihood ratio used to distinguish Groups 4 and 5. We chose 20 as the E-value interval to enforce a strong preference of one group over another in homology terms, and for the purposes of reporting accuracy here did not distinguish Groups 4 and 5 (see Supplementary Methods section and Supplementary Figure S4 for an exploration of ratio settings). A user of RITA might wish to use only the highest-confidence set consisting of Group 1a and b, an 'all-homology' set which includes Group 1–3, or a set that also includes some (Groups 1–4) or all (Groups 1–5) of the predictions that are based on composition only. Similarly, one can take the full set of PhymmBL predictions, or only the subset of PhymmBL predictions that are based on homology as well as

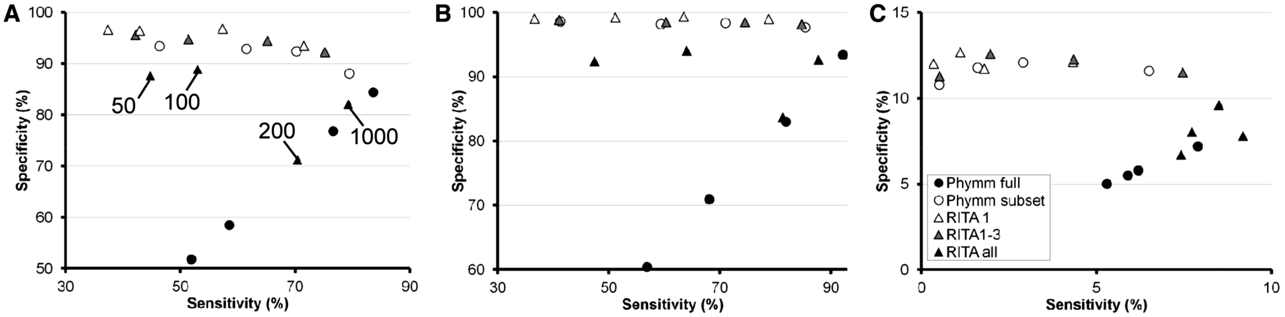

Figure 3. Performance comparison of RITA and PhymmBL on leave-one-out data set. Sensitivity and specificity of different subsets of RITA predictions, full PhymmBL predictions, and the subset of PhymmBL predictions that are based on both homology and composition are shown. Simulated DNA fragments of lengths 50, 100, 200 and 1000 nt are reported; sensitivity increases with increasing fragment length such that the leftmost symbol of a particular group always corresponds to the 50-nt fragment. (**A**) Genus-level predictions with same-species matches excluded; (**B**) phylum-level predictions with same-species matches excluded; (**C**) phylum-level predictions with same-class matches excluded.

composition (which we refer to as the 'PhymmBL subset'). A common feature of homology-inclusive sets of predictions is their relatively high specificity: RITA Group 1, Groups 1–3 and the PhymmBL subset all have specificity in excess of 90% in Figure 3a–b, with the lone exception of the PhymmBL subset of genus-level predictions on fragments of length 1000. The PhymmBL subset and RITA Groups 1–3 have similar accuracy, although at the genus level the PhymmBL subset tends to have higher sensitivity but lower specificity. The full prediction sets (RITA Groups 1–5 and PhymmBL) tend to have similar sensitivity and specificity that increases from ~50% on fragments of length 50 nt, to ~80% on fragments of length 1000 at the genus level, and from ~60% to 90% at the phylum level. PhymmBL showed slightly higher sensitivity and specificity on the longest fragments. However, RITA tends to have much higher specificity than PhymmBL on fragments of length 50 and 100 nt: accuracy across different genera is highly variable (20), and the main reason for this appears to be the tendency of RITA to assign many false positive predictions to five genera in particular, namely *Rhizobium*, *Bacillus*, *Staphylococcus*, *Clostridium* and *Streptomyces*. Since the majority of genera have a very small number of false positives assigned to them, the overall average specificity is very high. In (20) the composition-only approaches Phymm and NB were shown to have very similar accuracy properties, albeit with somewhat different performance on some genera: the difference in specificity between the hybrid approaches seen here is likely due to RITA's use of agreement as a criterion, which will discard some sequences that would be classified by PhymmBL's combined scoring function. Since RITA Groups 4–5 have very low accuracy, and are enriched in non-coding and unclassified sequences (Supplementary Figure S5), for many applications we recommend the use of Groups 1–3 for classification, especially if reads are short. However, NB often ranks the correct taxonomic group near the top in cases where an incorrect prediction is given (Supplementary Figure S6), so Groups 4–5 may still be useful if combined with other information such as taxonomic priors or read depth.

Classification at the phylum level with exclusion of species-level lineages (Figure 3b) can be successful if informative homology and compositional information is present in reference genomes from the same genus, family, order, class or phylum as the test genome. A much more challenging problem is classification at the phylum level with removal of all members of the same class as the test genome, since members of different classes (e.g. Alphaproteobacteria and Gammaproteobacteria, or Clostridia and Firmicutes) must provide enough compositional and homology information to classify the fragment. The results in Figure 3c confirm that this is an extremely difficult problem: regardless of the choice of classification algorithm, subset and fragment length, the sensitivity and specificity scores are appallingly low. Specificity never exceeds 13% and sensitivity is always <10%. From these results it is evident that the compositional profiles from one taxonomic class are not likely to be characteristic of the entire phylum, which is consistent with the observations of (13). Homology information is clearly unhelpful too, as lateral gene transfer, lineage-specific proteins and highly divergent sequences are very likely to occur at these deep taxonomic levels.

Glacier ice metagenome: rank-flexible classification and the effects of sequencing errors

The results in the previous section suggest that RITA predictions are comparable in accuracy to a carefully chosen subset of PhymmBL predictions, but highlight the confounding role that is played by taxonomic novelty. To evaluate a data set with varying and realistic degrees of novelty but still with a known true origin for each fragment, we created a *pseudometagenome* with simulated reads sampled from sequenced genomes in a way that mirrors the read length and taxonomic distribution of the glacier ice metagenome of (28). This data set consists of short reads and is sampled from a diverse community with varying degrees of taxonomic novelty: 16S rDNA sequence analysis suggested the presence of 11 bacterial phyla, with ~60% of the reads from phylum Proteobacteria, and a large proportion of those from the genera *Sphingomonas, Stenotrophomonas, Cryobacterium and Polarmonas*. A similar profile was found with a random read analysis by the developers of CARMA

(15) and Treephyler (16). We randomly sampled reads from 56 reference genomes to mimic the composition of this metagenome (Supplementary Table S5), and removed these genomes from the set of reference models in our database. The resulting data set has varying degrees of novelty with respect to our reference database at all taxonomic ranks from species to phylum: the extensive novelty in this data set means that many community members cannot be correctly classified at more-precise ranks such as genus or family. We compared two rank-specific classifiers (RITA and PhymmBL) at the level of phylum, and two rank-flexible classifiers (RITA and PhyloPythiaS), which can classify sequences to any taxonomic level, and tested the effects of introducing sequence changes that mimic different degrees of DNA sequencing error. A test of sequencing error rates of up to 5% (Supplementary Figure S7) on the leave-one-out data set used above showed little difference between either the base classification accuracy or the sensitivity to error of NB versus Phymm [see also (20)], but differences may arise due to the different ways in which RITA and PhymmBL combine homology and compositional information.

Rank-specific classifications at the phylum level for both PhymmBL and RITA are shown in Figure 4a. RITA Group 1 is a very conservative subset, with specificity = 69.8% and sensitivity only 37%. Expanding the RITA set to include all homology-based Groups 1–3 increased sensitivity by 11.1%, while decreasing the specificity by only 3.5%. Including all RITA predictions (Groups 1–5) increased sensitivity and decreased specificity by a further 2–3%. Applying RITA to pseudometagenome fragments with introduced errors showed some sensitivity to sequencing 'noise', but the overall sensitivity decreased by <6% as error rates increased from 0% to 5%, and specificity remained stable. The sensitivity of the PhymmBL subset on error-free sequences was less than the RITA Group 1–3 subset, even when RITA was classifying sequences with error rates of 5%. The full set of PhymmBL predictions had the highest overall sensitivity (11.9% greater than the PhymmBL subset, and 5.4% greater than RITA Groups 1–3), but with a specificity ~20% lower than that of the other classifiers.

Given the composition of our reference database and the set of reference genomes, it is in many cases impossible to classify a sequence correctly to a taxonomic level more precise than order or family. The choice to classify at the phylum level above reflects this limitation, but ignores cases where precise matches *can* be made at more-precise taxonomic ranks. We therefore modified RITA to perform rank-flexible classification that uses a user-provided file containing representative 16S rDNA sequences to map reference genome models to appropriate taxonomic ranks (see Supplementary Methods section and Supplementary Figure S8). For comparative purposes, we also applied PhyloPythiaS in a way that limits its taxonomic predictions to categories that are expected based on the 16S profile. The performance of these two rank-flexible classifiers, broken out by phylum, is compared in Figure 4b. PhyloPythiaS is very conservative on this data set, with most fragments classified on average to either the phylum or domain level. Greater distances from the diagonal line indicate a greater degree of over-classification. RITA assigned fragments from some phyla such as Gemmatimonadetes (one reference genome) and Verrucomicrobia (seven reference genomes) to an overly specific rank relative to PhyloPythiaS which made less-precise calls on these same phyla. However, RITA dramatically outperformed PhyloPythiaS on two critical phyla, Proteobacteria and Actinobacteria, increasing the precision of assignment by nearly two full ranks (e.g. from phylum to order) with negligible loss of accuracy. An additional two phyla, Deinococcus-Thermus and Chloroflexi, were classified with extremely high precision and accuracy by RITA, possibly owing to the relatively sparse

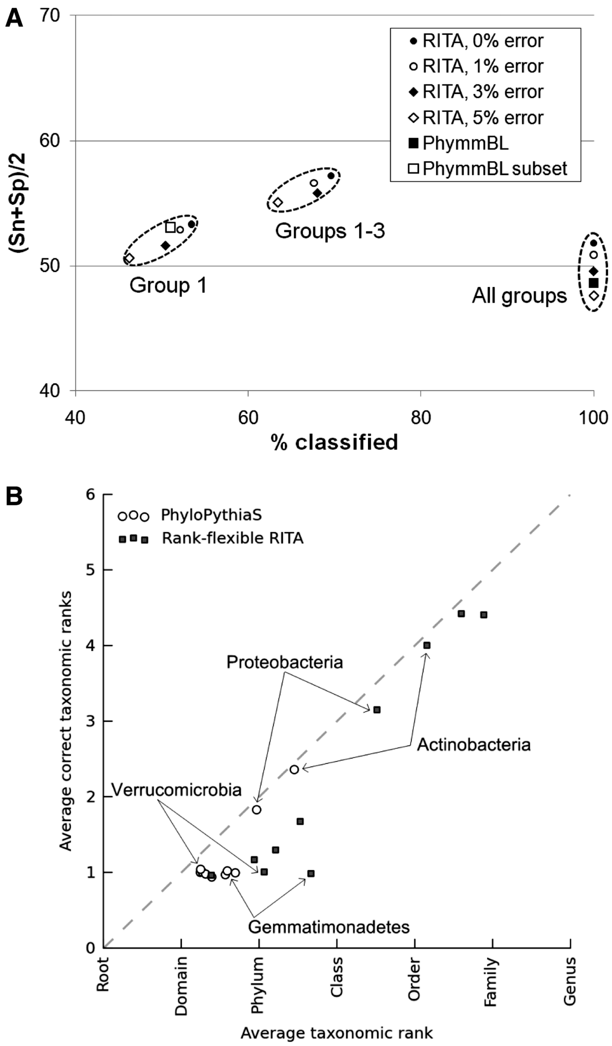

Figure 4. Performance of RITA in comparison with other leading methods on a glacier pseudometagenome. (**A**) Comparison of sensitivity and specificity of different RITA subsets and PhymmBL (rank-specific, phylum-level classifications). Dashed ovals surround different RITA subsets with varying amounts of random sequence error introduced. (**B**) Precision and accuracy of rank-flexible RITA and PhyloPythiaS on the glacier ice pseudometagenome. Values on the x-axis indicate the average taxonomic precision for each phylum on a scale from root = 0 to genus = 6, while the y-axis indicates the average number of ranks that were correct. The dashed line is equivalent to $y = x$ and shows the theoretical maximum accuracy.

taxonomic representation of these phyla in the set of sequenced genomes.

Finally, we ran the RITA pipeline on the actual metagenome data collected by (28), in rank-flexible mode and in rank-specific mode at the taxonomic levels of phylum and genus. Of the 1 076 539 reads, a total of 599 048 (55.6%) were classified in Groups 1–3, with 219 366 (20.4%) classified at a rank more precise than domain by the rank-flexible classifier. Although this set is a small fraction of the total reads, these are high-confidence predictions and are summarized in Figure 5a and Supplementary Table S6. The distribution of assignments to different ranks between class and genus was roughly equal, with phylum-level predictions slightly more numerous and dominated by Bacteroidetes, and every other rank dominated by Proteobacteria. Highly precise predictions were made for the genera *Polaromonas*, *Flavobacterium*, *Pedobacter* and *Deinococcus*, and the species *Stenotrophomonas maltophila*: in addition, many predictions at ranks higher than genus corresponded to the taxonomic lineage of *Polaromonas*. For the rank-specific classifier, the phylum-level breakdown of predictions was similar whether predictions were made at the level of phylum or genus (Figure 5b and Supplementary Table S7). A total of 295 880 reads were assigned to Groups 1–3 at the genus level, and 574 029 at the phylum level. Frequently observed genera include *Polaromonas* (41 187 reads, 13.9% of all assignments made at the genus level), *Methylibium*, *Variovorax*, *Burkholderia* and *Leptothrix* (9000–12 000 reads each) from the Proteobacteria; *Chitinophaga*, *Flavobacter* and *Pedobacter* from the Bacteroidetes (6000–11 000 reads each); *Deinococcus* (4264 reads); *Gemmata* (3415 reads); and *Gemmatimonas* (4388 reads). Of the other major genera identified by (28), *Sphingomonas* had 1594 associated reads. While *Cryobacterium* was absent from our reference set of genomes, *Clavibacter*, a member of the same family (Microbacteriaceae), had 1889 associated reads, second in the Actinobacteria after *Mycobacterium*.

Bovine rumen metagenome: using paired-end, short-read data

The metagenome of a cellulose-degrading cow rumen community was studied using paired-end Illumina sequencing technology (5), which generated reads of length 75 and 125 nt, with spacers of length 200, 300, 3000 and 5000 nt. Although the read lengths are extremely short for the purpose of taxonomic identification, the presence of paired ends in the data suggests two alternative approaches to improve classification accuracy. This arrangement allows the combination of predictions associated with both ends in a more-stringent way, for instance by requiring both ends to match the same reference genome in order to mark the fragment as classified. We explored the impact on classification of using such an agreement-based approach by constructing a pseudometagenome from reference genomes with varying degrees of taxonomic novelty which mimics the sequence and spacer length distribution of the rumen metagenome. Our pseudometagenome incorporated fragments from fifteen genomes, five of which were novel at the levels of strain, species and genus. Each genome was sampled in the same proportion (see ‘Materials and Methods’ section and Supplementary Table S8). We also examined the degree of agreement and frequency distribution of different taxonomic groups for a subsample of the rumen data set of (5), consisting of 200 000 reads lacking ambiguous characters.

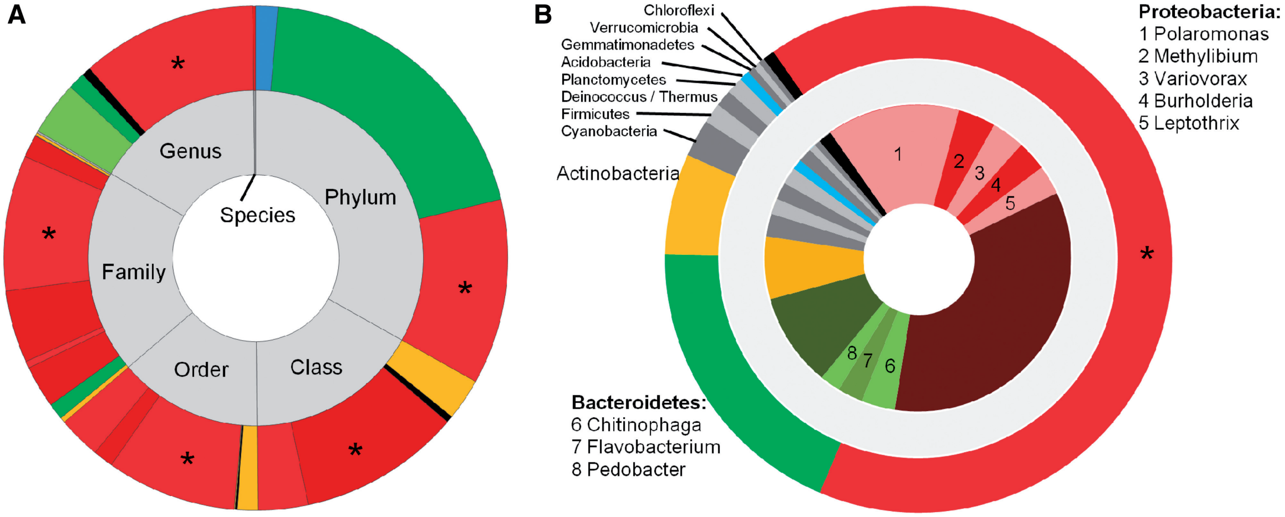

Figure 5. RITA classifications of the glacier metagenome of (26). (**A**) Rank-flexible classifications in Groups 1–3 to ranks between species and phylum. The inner ring identifies the rank at which different fragments were classified, while the outer ring shows the distribution across different labels at that rank, colored by the phylum to which the taxon belongs. Phylum colors: blue = Acidobacteria, green = Bacteroidetes, red = Proteobacteria, orange = Actinobacteria, black = other. Alternating shades of the same color are used to distinguish different taxa at the same rank from the same phylum. The taxonomic lineage of *Polaromonas* is identified with asterisks. (**B**) Rank-specific classifications at the phylum (outer ring) and genus (inner ring) levels, with color scheme as in panel A. Deepest red and green represent aggregated ’other’ genera of Proteobacteria and Bacteroidetes.

A comparison of classification accuracy on pseudometagenome sequences, separated by degree of taxonomic novelty and the inclusion or exclusion of a paired-end agreement criterion, is shown in Figure 6a (genus-level classification) and 6b (phylum-level classification). Given the accuracy of the Group 1–3 set in previous analyses, we required at least one of the two ends in an agreeing pair be assigned to one of these groups to successfully classify both ends. Using the paired-end constraint in classification invariably has the effect of decreasing the number of sequences that are both correctly and incorrectly classified. The benefits of this approach are dependent on the degree of taxonomic novelty: since the classification accuracy of genome fragments from ‘known’ species (i.e. novel strains) is extremely high, little benefit is achieved by adding the paired-end constraint. When classifying novel species, the total number of classified sequences drops by >10%: when classifying fragments at the genus level, this reduced the false negative rate by a factor of >10, from 5.8% to 0.5%. However, when classifying at the phylum level, the false negative rate was only 1% in the absence of the paired-end restriction, so many more true positive than false negative classifications were discarded. Classifying novel genera at the genus level is impossible, so all predictions in this set are incorrect: the paired-end constraint reduced the false negative rate from 31.1% to 19.9%. When classifying to the phylum level, the false negative rate dropped from 5.5% to 0.6%.

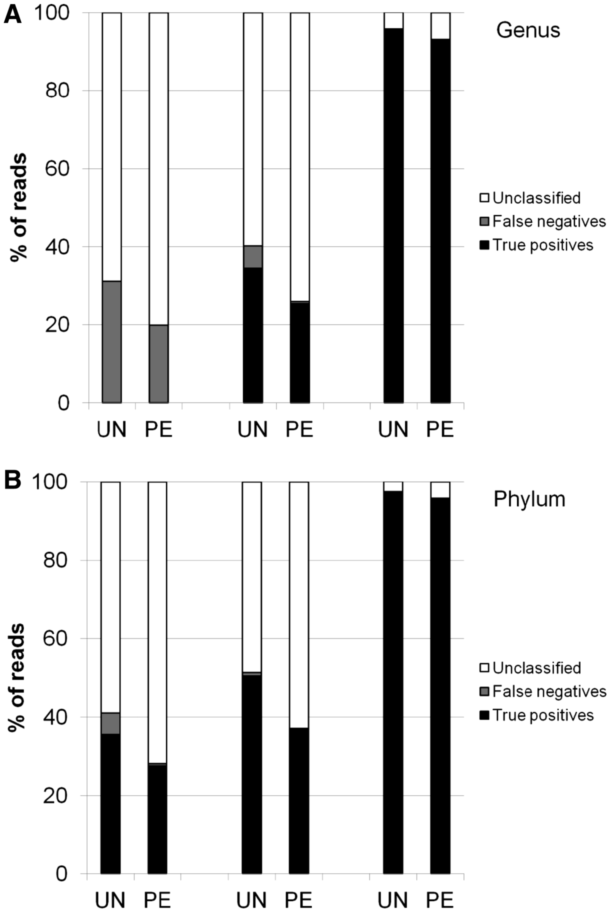

Figure 6. Classification accuracy of rank-specific RITA on constituents of a rumen pseudometagenome at the genus (**A**) and phylum (**B**) levels. Accuracy is shown for three levels of genomic novelty: genomes that are novel at the level of genus (leftmost pair of bars), novel at the level of species (middle pair), and novel at the level of strain (rightmost pair). Paired bars for each combination of classification rank and degree of novelty represent the scoring of reads in an unconstrained way (UN) and using a paired-end constraint (PE).

We also applied RITA to a sample of 200 000 reads from the bovine rumen metagenome data of (5). Although accuracy cannot be assessed given the lack of ground truth, we can evaluate the effect of applying or not applying the paired-end requirement for classification, in terms of the effect on the total number of classified sequences, the agreement with the broad taxonomic results (considered only at the taxonomic level of order) of (5), and the reported taxonomy in light of known rumen organisms. Rank-specific RITA classified a relatively small number (39 149/200 000, with a total of 608 genera predicted) of reads from the sample, and this number dropped even further (14 016/200 000, with a total of 178 genera predicted) with the use of the paired end requirement. The tighter distribution of genera was reflected in the percentage of reads that were mapped to the top 15 most frequent genera: in the absence of the paired-end constraint, only 67.1% of classified reads were assigned to this set, while 91.3% were assigned to the top 15 genera when the paired-end constraint was used. Genera are shown, sorted by decreasing number of assignments, in Supplementary Table S9. Not surprisingly, there is strong agreement between the two lists, with 13 genera appearing in the top 15 of each list. Nine of these fifteen genera fall into the two orders highlighted in the original paper (Clostridiales, including *Butyrivibrio* which was identified by the original authors, and Bacteroidetes), while five of the remaining genera (*Fibrobacter*, *Eubacterium*, *Methanobrevibacter*, *Victivallis* and *Slackia*) are frequently found in rumen or gut samples. Genus *Acinetobacter* is not normally associated with the gut and is aerobic, in contrast with the other identified groups, and its drop in rank from tenth (587 total reads) in the unconstrained set to 99th (one paired set of reads) in the constrained set may indicate that it is a frequent false positive that is effectively filtered using the paired-end constraint. Two orders identified in the original study, Spirochaetales and Myxococcales, do not appear in the top 15 genera from the unconstrained classifications, but the Spirochaete *Treponema* was ranked 12th in the constrained list; furthermore, the composition-based approach used by (5) may have misassigned some novel lineages to the incorrect order. We note that the Myxococcales contain several large genomes (>5 Mb) including members of genus *Anaeromyxobacter*, *Myxococcus* and the titanic *Sorangium cellulosum*, which have been impacted by extensive LGT (36,37); consequently, rumen-associated organisms of many other lineages may share a great deal of compositional and genetic similarity to members of this order.

Increasing the speed of predictions and hypothesis testing on a set of gut microbiomes

Although our NB classifier is considerably faster than Phymm, the homology search component (particularly BLASTX, which needs to search six-way conceptual translations of nucleotide sequences) is the principal limiting step. USEARCH (30) attempts to optimize homology search using a range of techniques including k-mer based prioritization of search candidates, and search termination when a target number of successes or failures have been observed. Since RITA is only concerned with best and near-best matches, and since the majority of BLASTX runtime is spent comparing query sequences with non-homologous targets in the large reference database, the USEARCH strategy may be ideal for the RITA pipeline. Comparison of the UBLASTX and BLASTX algorithms showed that USEARCH was approximately 14 times faster than BLASTX, with similar accuracy (Supplementary Figure S9). Since BLASTX reports *E*-values with higher precision than UBLASTX, our variant of the pipeline identifies all the best hits for each fragment, rebuilds a BLASTX database with just these genomes, and searches that database with the fragments. We applied our modified RITA pipeline (RITA–UB), which replaces the three BLAST steps of RITA (RITA–BL) with a single step that screens sequences with UBLASTX before running a reduced BLASTX search, to the data set used by (29). The authors of this article used a series of statistical techniques to assess the similarity of gut microbiome samples from individuals in several countries, and claimed the discovery of three distinct ‘enterotypes’ that differed substantially in their taxonomic composition. Here we assign taxonomic information to these sequences using RITA–UB to assess the feasibility of applying RITA to this large data set, and use clustering and ordination techniques to assess whether the three enterotypes can indeed be recovered.

The total running time to assign all 6 734 462 reads was approximately 2.1 days on 100 CPU cores, as compared with an estimated 13.7 days (based on a trial run with enterotype sample A) that would have been required by RITA–BL. A total of 5 134 129 reads received taxonomic assignments, and we examined the frequency of predicted genera in comparison with the results of (29), which were generated using SmashCommunity which uses thresholded homology searches to assign taxonomic labels to sequence reads (32). Although some amount of variation in the frequency of *Bacteroides*, *Prevotella*, *Ruminococcus* (Figure 7a and Supplementary Figure S10) and other genera (Supplementary Table S10) was observed between the RITA and SmashCommunity results, the relative frequencies of particular genera across the set of samples was largely consistent. Similarly, principal components analysis (PCA) applied to both sets of taxonomic predictions (Figure 7b) showed that the relative positioning of different samples was similar in spite of differences in the predicted frequencies of genera. While hierarchical clustering of RITA profiles using UPGMA (Figure 7c) suggested some amount of sample clustering by reported enterotype, none of the three sets of samples assigned to a single enterotype constituted a single cluster within the tree. Hierarchical clusters based on SmashCommunity profiles (Supplementary Figure S11) were slightly more cohesive, but still failed to resolve the three enterotypes. Differences in frequencies are likely due to both the use of a different annotation pipeline (i.e. RITA versus SmashCommunity) and the normalizations performed by (29), but PCA and UPGMA clustering of either set of predictions fails to recover the proposed enterotypes. Instead of the standard PCA technique which is unsupervised, the original authors provided the enterotype assignments as an ‘instrumental variable’ which serves as an optimality criterion for construction of the ordination plot. This choice of technique therefore yielded greatly enhanced visual separation of the clusters, whereas our completely unsupervised approach does not give clear separation between the three reported enterotypes.

Runtime comparison of PhyloPythiaS, PhymmBL and RITA variants

Using the same enterotype sample (sample ‘A’, which contained 116 244 sequences) that was used above for benchmarking, we compared the performance of PhyloPythiaS, PhymmBL and several RITA variants developed in this article (Supplementary Figure S12). While the RITA variant developed in the section above coupled an initial pass with UBLASTX with a subsequent reanalysis using BLASTX in order to achieve more-precise *E*-value estimates, we also benchmarked a ‘Fast’ series of variants which omitted the followup BLASTX step in order to reduce computation time. Since the ‘maxaccepts’ and ‘maxrejects’ parameters of UBLASTX can influence the thoroughness of the search and the corresponding runtime, we tried setting these values in tandem to 5, 10 and 100. We used the same query data set as above, sample ‘A’ from the enterotype data set, and in this case used a reduced reference database of 428 draft and completely sequenced genomes of known gut-associated organisms. Our composition-only approach (NB) completed in 5 min 53 s, 15.2 times faster than PhyloPythiaS. Among hybrid classifiers, five RITA variants were faster than PhymmBL (16.4 h), including the three ‘Fast’ variants and a further ‘Fast’ approach coupled with D-BLASTN, and a variant in which D-BLASTN was the only homology search algorithm used. Notably, the ‘Fast’ variants with ‘maxaccepts’ and ‘maxrejects’ set to 5 or 10 required 3–4 h to run, and adding a follow-up D-BLASTN step increased the runtime to 7.36 h, still less than half of that of PhymmBL. The three RITA variants that included a full BLASTX step ran in considerably longer time, demonstrating the time-consuming nature of the translated homology search: all required in excess of 300 h to complete, and the UBLASTX variant required over 500 h. The advantage of using UBLASTX followed by BLASTX in this instance is negligible, which we attribute to having a much more ecologically focused database: since an increased proportion of reference database genomes are likely to have hits to query sequences, UBLASTX will filter out far fewer genomes

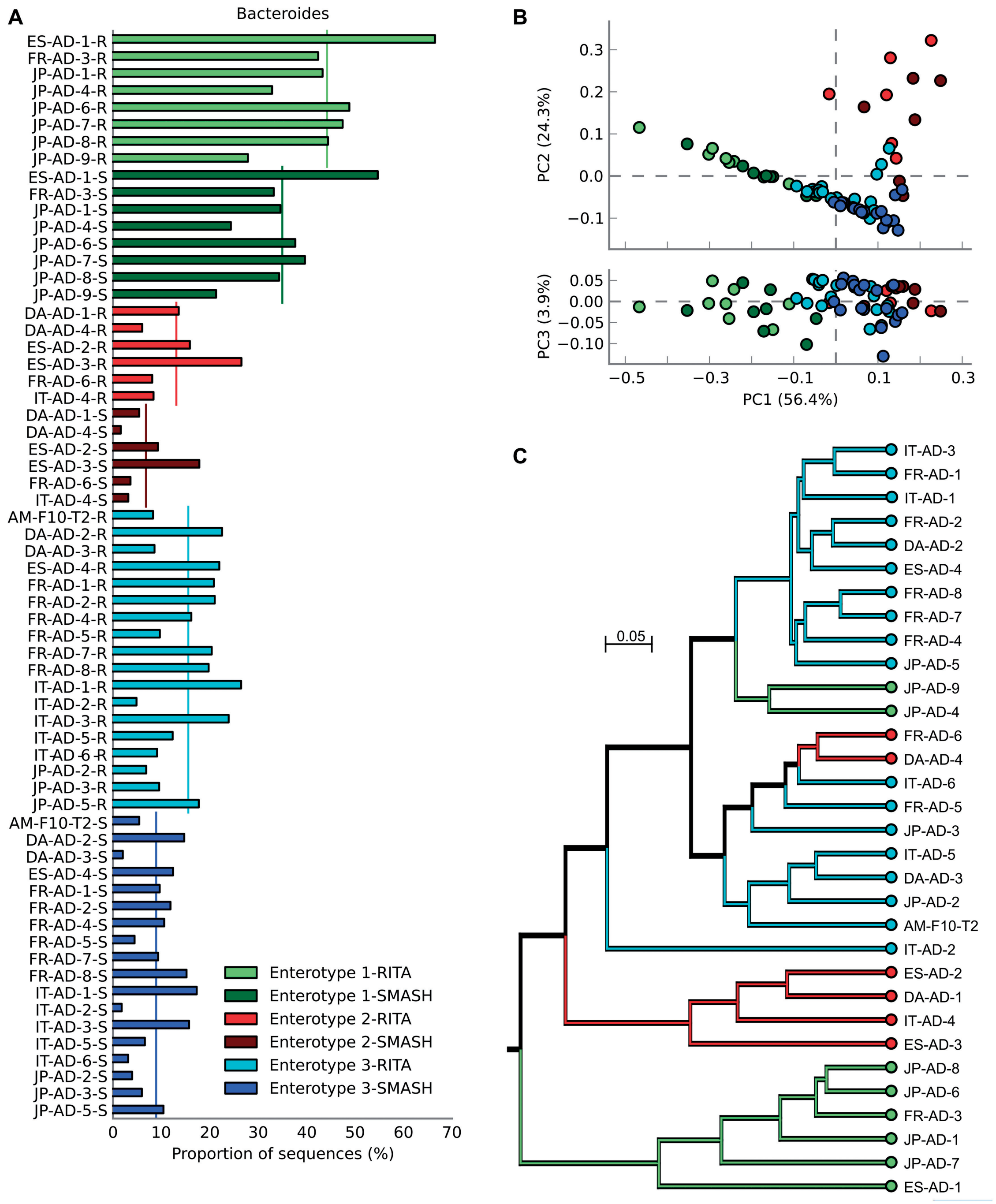

Figure 7. Taxonomic attributions obtained with RITA and SmashCommunity for the gut microbiota of 32 individuals from Japan, France, Italy, Spain, Denmark and America. (**A**) Comparison of the proportion of fragments within each sample assigned to the genus *Bacteriodes* by RITA (R; light colors) and SmashCommunity (S; dark colors). (**B**) PCA plot of genus-level profiles obtained with RITA and Smash Community. (**C**) UPGMA clustering of genus-level profiles obtained with RITA. Coloring is consistent across all three panels and corresponds to the enterotypes proposed by Arumugam *et al.* (2011).

and the majority of searches will still need to be performed by BLASTX.

The choice of homology algorithm(s) will influence the recovery of homologous matches. Since homology-based predictions are in general better than those based on composition alone, we examined the proportion of the 116 244 sequences in sample ‘A’ that were assigned to homology categories by different RITA variants (Supplementary Figure S13). There is a clear distinction between approaches that use only one type of homology search algorithm (i.e. nucleotide–nucleotide, or translated nucleotide–protein), and approaches that combine both. Fewer than 80 000 sequences were assigned homology-based classifications by RITA variants based only on

D-BLASTN, BLASTX, or Fast UBLASTX. PhymmBL, which uses BLASTN only, classified over 84 000 sequences because it does not require agreement between homology and composition, or a strong homology preference for one group over another. By contrast, hybrid approaches that use both types of homology-based prediction classified 89 744 (Fast UBLASTX + D-BLASTN), 101 456 (D-BLASTN + UBLASTX followed by BLASTX) and 110 594 (Full RITA pipeline) sequences using homology information. Although using two homology-based approaches rather than one imposes an additional runtime cost, the results of the leave-one-out analysis above and the comparative benchmarking here demonstrate the value of using both approaches.

DISCUSSION

The RITA pipeline uses agreement among homology and composition-based classifiers as a measure of confidence to produce a subset of read classifications that are more accurate than those obtained from either type of classifier alone. NB, due to its fast run-time, simple model and similar performance to the state-of-the-art compositional classifier Phymm (23), makes it the perfect candidate for compositional analysis. Discontiguous MEGABLAST, BLASTN and BLASTX give us different sets of assignments with different levels of sensitivity, specificity and run-time, while UBLASTX (30) provides an effective filter to further accelerate classification. Through several steps of validation on sequenced genomes and pseudometagenomes, we found that predictions based only on composition (i.e. Groups 4 and 5, and similarly the PhymmBL predictions that were based on Phymm alone) tended to be highly inaccurate, and we recommend that only predictions that include both a strong homology component and potential agreement with composition (i.e. RITA Groups 1–3) be used for any classification task. However, the accuracy of compositional classifiers such as Phymm, NB and PhyloPythiaS increase with increasing sequence length, Group 4 and 5 assignments of fragments several thousand nucleotides in length might be treated with greater confidence. Also, since the correct model is still often highly ranked by NB in cases where the best-scoring model is incorrect, additional information might be used to select the correct classification from a list of the best-scoring NB models for a given fragment. Thus the compositional predictions can be considered valuable in their own right. One promising type of approach is that of (38), who use taxonomic information from all reads to improve the taxonomic precision of sequence assignments. In the context of RITA, predictions from the higher-confidence groups might be used to select appropriate assignments from the list of possibilities identified for lower-confidence groups.

The simulated and real data sets we examined were challenging for two important reasons. The short reads generated by sequencing machines such as the Illumina platform provide very little information with which to assign sequence reads to the correct genome. Since these sequencers typically yield very deep coverage (>10× or even 100×) of dominant community members, one solution is to perform assembly prior to taxonomic assignment, as was done by (5), which will yield longer contiguous regions which can be assigned with higher confidence. We have shown that RITA retains high specificity on short reads, so although many such reads may remain unclassified, those assigned to RITA Groups 1–3 have a high probability of being correct. Our paired-end classification strategy, tested on the rumen data set, provides an additional mechanism to increase the confidence of predictions: we showed that this strategy removes many implausible genera from the set of classifications. The other significant challenge posed by metagenomes is that of taxonomic novelty: although members of the same species and genus are very likely to have similar gene content and compositional biases, Figure 3c in particular shows that the essential signals for classification are no longer present due to compositional change, gene gains and losses and lateral gene transfer when the closest relatives of the organism being classified are from a different class in the same phylum. To deal with such cases, we recommend the use of a rank-flexible classifier such as our 16S-restricted version of RITA, which can at least limit the set of predictions to lineages that are known to be present based on well-sampled marker gene data. An alternative is to use an unsupervised technique in order to identify sets of reads with strong affiliations to one another, without the need for closely related reference genomes and without assigning explicit taxonomic labels to inferred clusters. If a community consists of a mix of taxonomically well-sampled and poorly sampled organisms, it may be worthwhile to use a supervised classifier to associate reads with strongly matching reference genomes, and then used an unsupervised approach to cluster the remaining reads.

Analysis of the ‘enterotype’ data set of (29) showed how taxonomic assignments can be used to identify patterns of similarity among many sampled microbial communities. Although there was reasonable agreement between RITA predictions and those of SmashCommunity, our use of a completely unsupervised ordination technique did not support the recovery of the three enterotypes as distinct clusters. This raises the question of whether enterotypes are truly distinct, or whether they represent a continuum of diversity that can shift gradually in response to a number of factors. More-intensive microbiome sampling and the use of additional statistical techniques will be necessary to confirm or refute the enterotype hypothesis.

Metagenome projects now cover a wide range of communities with varying degrees of taxonomic novelty, use sequencing technologies that generate massive numbers of reads of varying length, and produce data that can be analyzed in various ways, e.g. with or without the use of sequence assembly. Our RITA pipeline has several properties that make it useful for a wide range of projects. First, it is fast, making use of simplified statistical models and fast homology searching such as UBLASTX to expedite the classification process. Second, the requirement for either agreement among different types of classifier or a strong homology result leads to very high specificity in relation to other published classifiers. Third, the rank-flexible approach of RITA allows the classification of fragments from metagenomes with varying degrees of taxonomic novelty relative to the reference database. Finally, the modular nature of RITA makes it adaptable to new tools and techniques. As we demonstrated with the comparison of homology search techniques, the use of multiple search algorithms yields more high-confidence classifications than any single method alone, and heuristics such as those in UBLASTX can be used to offset the increased computational cost. As new algorithms emerge, they can be incorporated into the RITA pipeline to complement existing approaches.

AVAILABILITY

RITA can be downloaded as a standalone Python program at (http://kiwi.cs.dal.ca/Software/RITA). The standalone version allows the user to implement the components of the RITA pipeline (i.e. the BLAST algorithms, Fast UBLASTX and NB) in any order, with custom homology and NB thresholds. Users can also add new elements to the RITA pipeline by writing small Python classes to execute and parse the results of other programs. RITA has also been made available as a web application available at (http://ratite.cs.dal.ca/rita). Users can submit FASTA formatted metagenomic read files to the server, along with an optional set of 16S rDNA sequences which are aligned with mothur (33) and subsequently placed in a tree with the reference genomes using FastTree (34). The application emails the user when the processing has completed at which point they can download a list of classifications and pipeline group assignments for each fragment.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–10, Supplementary Figures 1–13, Supplementary Methods and Supplementary References [39–41].

ACKNOWLEDGEMENTS

The authors thank Laura Hug for helpful discussions regarding the pseudometagenome analysis, and members of the Beiko lab for helpful discussions and testing.

FUNDING

Natural Sciences and Engineering Research Council of Canada (to N.J.M. and D.H.P.); the Killam Trusts (to D.H.P.); Genome Atlantic; the Canada Foundation for Innovation; the Canada Research Chairs program (to R.G.B.); and the Government of Canada through Genome Canada and the Ontario Genomics Institute [2009-OGI-ABC-1405]. Funding for open access charge: The Canadian Natural Sciences and Engineering Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–73.
2. Turnbaugh,P.J., Hamady,M., Yatsunenko,T., Cantarel,B.L., Duncan,A., Ley,R.E., Sogin,M.L., Jones,W.J., Roe,B.A., Affourtit,J.P. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **22**, 480–484.
3. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
4. García Martín,H., Ivanova,N., Kunin,V., Warnecke,F., Barry,K.W., McHardy,A.C., Yeates,C., He,S., Salamov,A.A., Szeto,E. *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.*, **24**, 1263–1269.
5. Hess,M., Sczyrba,A., Egan,R., Kim,T.W., Chokhawala,H., Schroth,G., Luo,S., Clark,D.S., Chen,F. and Zhang,T. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, **331**, 463–467.
6. Case,R.J., Boucher,Y., Dahllö,I., Holmström,C., Doolittle,W.F. and Kjelleberg,S. (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.*, **73**, 278–288.
7. Liu,B., Gibbons,T., Ghodsi,M. and Pop,M. (2010) MetaPhyler: taxonomic profiling for metagenomic sequences. 2010 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 18–21 December 2010*, pp. 23–28.
8. Crosby,L.D. and Criddle,C.S. (2003) Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. *Biotechniques*, **34**, 790–794.
9. Forney,L.J., Zhou,X. and Brown,C.J. (2004) Molecular microbial ecology: land of the one-eyed king. *Curr. Opin. Microbiol.*, **7**, 210–220.
10. Manichanh,C., Chapple,C.E., Frangeul,L., Gloux,K., Guigo,R. and Dore,J. (2008) A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res.*, **36**, 5180–5188.
11. Green Tringe,S., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J., Detter,J.C. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.
12. Gilbert,J.A., Field,D., Huang,Y., Edwards,R., Li,W., Gilna,P. and Joint,I. (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One*, **3**, e3042.
13. Perry,S.C. and Beiko,R.G. (2010) Distinguishing microbial genome fragments based on their composition: evolutionary and comparative genomic perspectives. *Genome Biol. Evol.*, **2**, 117–131.
14. 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.
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. Schreiber,F., Gumrich,P., Daniel,R. and Meincke,P. (2010) Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, **26**, 960–961.
17. Zhang,Y. and Sun,Y. (2012) MetaDomain: a profile HMM-based protein domain classification tool for short sequences. *Pac. Sym. Biocomput.*, **17**, 271–282.
18. Diaz,N.N., Krause,L., Goesmann,A., Niehaus,K. and Nattkemper,T.W. (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbour approach. *BMC Bioinformatics*, **10**, 56.

19. Patil,K.R., Haider,P., Pope,P.B., Turnbaugh,P.J., Morrison,M., Scheffer,T. and McHardy,A.C. (2011) Taxonomic metagenome sequence assignment with structured output models. *Nat. Methods*, **8**, 191–192.
20. Parks,D.H., MacDonald,N.J. and Beiko,R.G. (2011) Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics*, **12**, 328.
21. Rosen,G., Garbarine,E., Diamantino,C., Polikar,R. and Sokhansanj,B. (2008) Metagenome fragment classification using n-mer frequency profiles. *Adv. Bioinformatics*, **2008**, 1–12.
22. Sandberg,R., Winberg,G., Bränden,C.I., Kaske,A., Ernberg,I. and Cöster,J. (2001) Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res.*, **11**, 1401–1409.
23. Brady,A. and Salzberg,S.L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods*, **6**, 673–676.
24. Nalbantoglu,O.U., Way,S.F., Hinrichs,S.H. and Sayood,K. (2011) RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics*, **12**, 41.
25. Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M. *et al.* (2009) The ribosomal database project: improved alignments and new tools for rDNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.
26. Huson,D.H., Auch,A.F., Qi,J. and Schuster,S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.
27. McHardy,A.C. and Rigoutsos,I. (2007) What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr. Opin. Microbiol.*, **10**, 499–503.
28. Simon,C., Wiezer,A., Strittmatter,A.W. and Daniel,R. (2009) Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Appl. Environ. Microbiol.*, **75**, 7519–7526.
29. Arumugam,M., Raes,J., Pelletier,E., Le Paslier,D., Yamada,T., Mende,D.R., Fernandes,G.R., Tap,J., Bruls,T., Batto,J.M. *et al.* (2010) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.
30. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
31. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
32. Arumugam,M., Harrington,E.D., Foerstner,K.U., Raes,J. and Bork,P. (2010) SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics*, **26**, 2977–2978.
33. Schloss,P.D., Westcott,S.L., Ryabin,T., Hall,J.R., Hartmann,M., Hollister,E.B., Lesniewski,R.A., Oakley,B.B., Parks,D.H., Robinson,C.J. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.
34. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
35. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classifications: an overview. *Bioinformatics*, **16**, 412–424.
36. Goldman,B.S., Nierman,W.C., Kaiser,D., Slater,S.C., Durkin,A.S., Eisen,J.A., Ronning,C.M., Barbazuk,W.B., Blanchard,M., Field,C. *et al.* (2006) Evolution of sensory complexity recorded in a myxobacterial genome. *Proc. Natl Acad. Sci. USA*, **103**, 15200–15205.
37. Thomas,S.H., Wagner,R.D., Arakaki,A.K., Skolnick,J., Kirby,J.R., Shimkets,L.J., Sanford,R.A. and Löffler,F.E. (2008) The mosaic genome of Anaeromyxobacter dehalogenans strain 2CP-C suggests an aerobic common ancestor to the delta-proteobacteria. *PLoS One*, **3**, e2103.
38. Gori,F., Folino,G., Jetten,M.S.M. and Marchioni,E. (2011) MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. *Bioinformatics*, **27**, l196–203.
39. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. and Federhen,S. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
40. Legendre,P. and Legendre,L. (1998) *Numerical Ecology*, 2nd edn. Elsevier, Amsterdam.
41. Parks,D.H. and Beiko,R.G. (2010) Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, **26**, 715–721.