Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models

Arthur Brady & Steven L Salzberg

Metagenomics projects collect DNA from uncharacterized environments that may contain thousands of species per sample. One main challenge facing metagenomic analysis is phylogenetic classification of raw sequence reads into groups representing the same or similar taxa, a prerequisite for genome assembly and for analyzing the biological diversity of a sample. New sequencing technologies have made metagenomics easier, by making sequencing faster, and more difficult, by producing shorter reads than previous technologies. Classifying sequences from reads as short as 100 base pairs has until now been relatively inaccurate, requiring researchers to use older, long-read technologies. We present Phymm, a classifier for metagenomic data, that has been trained on 539 complete, curated genomes and can accurately classify reads as short as 100 base pairs, a substantial improvement over previous composition-based classification methods. We also describe how combining Phymm with sequence alignment algorithms improves accuracy.

Dramatic improvements in the speed and efficiency of DNA sequencing have encouraged the rapid growth of metagenomics, the study of DNA collected directly from environmental samples. This new field, which promises to uncover thousands of previously unknown species, has been compared to "a reinvention of the microscope in the expanse of research questions it opens to investigation"[1]. Only a small fraction of microbial organisms can be grown in a laboratory, a prerequisite for traditional genome sequencing and analysis. Single-organism genome sequencing projects have yielded a wealth of new scientific knowledge, which we are only beginning to exploit. Metagenomics promises to take these discoveries even further, by enabling scientists to study the full diversity of the microbial world. By sequencing collections of organisms from environments ranging from the human body to soil to the ocean floor, metagenomics projects are vastly increasing the range of organisms that can be analyzed, allowing for systems-level study of microbial environments and revealing a heretofore hidden world of biological complexity[2].

Although sequencing technology allows us to collect and sequence vast samples of DNA collected directly from organisms in the environment, many technical challenges must be overcome in order to make sense of these data. To identify the species and genes in a sample, DNA fragments ('reads') from common species need to be grouped together and assembled, if possible.

The newest sequencing technologies (increasingly preferred by metagenomics researchers owing to reduced cost) produce relatively short reads, 25–400 base pairs (bp), making the taxonomic classification problem considerably more difficult than with longer reads, which contain more identifying information. One algorithm, Carma[3], attempts to match short reads to known Pfam domains (structural components conserved across multiple proteins) and protein families. In a pilot study using Carma, however, only ~15% of random 100-bp shotgun reads could be matched to extant Pfam groups; even among this reduced set of reads (excluding 85% of the input data because of the absence of a match), the average sensitivity of this approach at the genus level was 40%, meaning only 6% of the input reads were correctly classified[3]. PhyloPythia[4], a classification method based on support vector machines, examines oligonucleotide frequencies to characterize taxonomic groups. This method is effective for DNA fragments of 3,000 bp and longer, but for 1,000-bp sequences, sensitivity drops drastically (to just 7.1% at the genus level). It has been observed[5] that 1,000 bp is a critical barrier that classification methods need to break.

Another approach is to use sequence homology, aligning reads to known sequences using the basic local alignment search tool (BLAST)[6] and assigning taxonomy based on the best match[7,8]. BLAST is highly accurate if the source organism's DNA has been sequenced; however, if the source species is missing from the database, accuracy drops dramatically, as we show below. The metagenome analysis (MEGAN) software system[9] classifies reads based on multiple high-scoring BLAST hits, assigning reads to a common ancestor of those BLAST matches that exceed a bit-score threshold. Notably, in a recent coral reef study[10], only 12% of reads had matches in a comprehensive microbial BLAST database.

Here we present Phymm, a new method for phylogenetically classifying short sequence fragments such as those generated by metagenomics sequencing projects. We use 'classification' to mean the assignment of a specific label (in our case, a phylogenetic group) to members of a dataset (in this case, DNA reads). We want to draw a clear distinction to 'binning', which though often

Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. Correspondence should be addressed to A.B. (abrady@umiacs.umd.edu).

RECEIVED 2 MARCH; ACCEPTED 25 JUNE; PUBLISHED ONLINE 2 AUGUST 2009; DOI:10.1038/NMETH.1358

used interchangeably with classifying, refers to the grouping of a dataset into subgroups which are distinct from one another; these subgroups may remain unlabeled.

Phymm uses interpolated Markov models (IMMs) to characterize variable-length oligonucleotides typical of a phylogenetic grouping. IMMs have been used with great success for bacterial gene finding in the Glimmer system[11], but this is to our knowledge the first use of IMMs for the general phylogenetic classification problem. Our results demonstrate that for short reads, Phymm is a dramatic improvement over previous methods such as PhyloPythia[4], accurately classifying unknown fragments as short as 100 bp. We also present a hybrid method that incorporates information from both Phymm and BLAST, and show that this hybrid method outperforms either of the two single methods.

In an earlier study describing the program Glimmer[12], IMMs had been shown to be highly accurate at distinguishing reads derived from a bacterial symbiont *Prochloron didemni* from those of its eukaryotic sea squirt host *Lissoclinum patella*. Glimmer discriminated with 99% accuracy between reads from two evolutionarily distant species. Our goal in the present study was to determine how well this discrimination generalizes to the problem of fully classifying much larger metagenomics samples that included many closely related organisms. We tested our method on both synthetic and real metagenome data.

RESULTS

Synthetic metagenome data

We conducted three groups of classification experiments, in which we assigned test sets of synthetic metagenomic reads taxonomic labels using Phymm, BLAST and Phymm plus BLAST (PhymmBL). Phymm contains a large suite of IMMs trained on chromosomes and plasmids from organisms collected from the US National Center for Biotechnology Information (NCBI) RefSeq database[13] (Phymm's 'reference library'). When used to score a DNA sequence, an IMM computes a score corresponding to the probability that the IMM generated that sequence, which can be used to estimate the probability that the sequence belongs to the class of sequences on which the IMM was trained. In the Phymm experiment, we scored each read with each IMM in the reference library, and the read was then classified using the clade labels belonging to the organism whose IMM generated the best score for that read. In the BLAST experiment, each read was submitted as a BLASTN query, searching against a background database built from the same genomes used to generate the IMMs, and clade labels were assigned using the known labels of the best BLAST hit. Finally, we used PhymmBL to score reads using both standalone methods in parallel, assigning a 'best hit' using a weighted combination of scores from both methods (Online Methods).

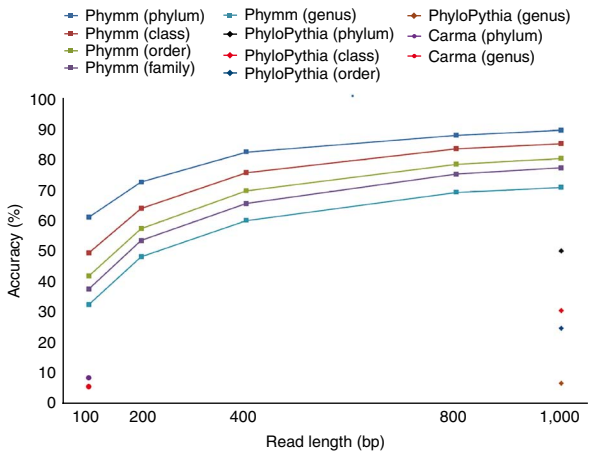

Figure 1 | Accuracy of Phymm, with species-level matches masked. Colored dots show classification accuracy reported for PhyloPythia at 1,000 bp for genus- through phylum-level predictions, and for Carma at 100 bp (as a percentage of the entire input dataset) for genus- and phylum-level predictions.

Clade-level exclusions

Our central goal was to model the problem of classifying a sequence from a species that has never before been observed; we expect that metagenomic data will lead to the discovery of thousands of new species and that this problem will be a common one. By definition, given a read from an organism whose genome has not been sequenced, no classification method can predict the correct species label (because that label does not exist). For higher-level phylogenetic classifications, in contrast, we may have previously seen the genus, family or other higher-level clade. We therefore repeated all three groups of experiments multiple times, with each iteration configured to explicitly exclude comparisons of each query read to related species at increasingly general clade levels.

Classification accuracy

We focused first on the experiment in which we used Phymm to classify reads for which species-level matches were masked. We repeated this experiment for reads of 100, 200, 400, 800 and 1,000 bp (Online Methods) and determined the accuracy of results (**Fig. 1** and **Supplementary Table 1**). As expected, classification at the genus level was the most difficult task, with 53 possible genera available as labels. Moving up the phylogenetic tree, the datasets contained 48 distinct bacterial and archaeal families, 34 orders, 21 classes and 14 phyla. Phymm's accuracy improved with greater read length, as expected, ranging from 32.8% for the problem of assigning the correct genus to 100-bp reads, up to 89.8% for assigning the correct phylum to 1,000-bp reads. At 400 bp,

Table 1 | Comparison of performance accuracy

Query length	Phymm	BLAST	PhymmBL	PhyloPythia
Genus	71.1%	73.8%	78.4%	7.1%
Family	77.5%	79.2%	84.8%	Not available
Order	80.6%	80.8%	86.9%	25.1%
Class	85.4%	84.1%	90.6%	30.8%
Phylum	89.8%	88.0%	93.8%	50.3%

Same-species matches were masked, for 1,000-bp reads. PhyloPythia accuracy was measured as the percentage of all reads for which each method produced the correct phylogenetic label. We performed this set of experiments once for each read length.

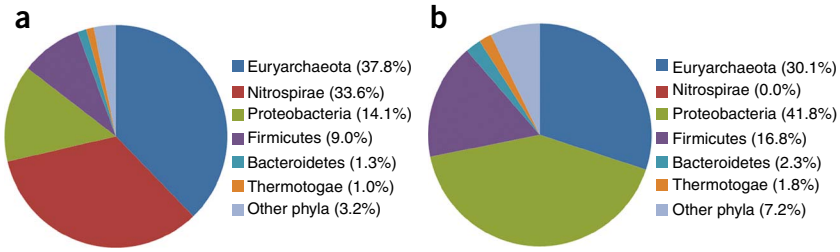

Figure 2 | PhymmBL's phylum-level population characterization of the AMD data. (**a**,**b**) Data were characterized using the RefSeq-generated IMMs plus IMMs generated from the draft genomes of the three dominant species in the AMD set (**a**) and the RefSeq-generated IMMs only (**b**). As no members of Nitrospirae were present in the RefSeq-only IMM set, no AMD reads could be classified with this label when PhymmBL was restricted to that set.

the read length provided by current-generation Roche 454 pyrosequencers, accuracy at the genus level was 60.3%. These results are substantial improvements over previous methods such as the support vector machine–based PhyloPythia, which reported 7.1% accuracy at the genus level for 1,000-bp sequences[4] and Carma, which labeled 6% of 100-bp reads correctly at the genus level[3]. In most cases, BLAST analysis by itself was superior to Phymm analysis by itself, but for the 800-bp and 1,000-bp reads, Phymm outperformed BLAST at the class and phylum levels (**Supplementary Table 2**).

The hybrid PhymmBL classifier produced additional improvements (**Table 1** and **Supplementary Table 3**). For all read lengths and clade levels, PhymmBL outperformed both Phymm and BLAST, showing approximately 6% improvement over BLAST by itself at all taxonomic levels for the 1,000-bp query set. These results indicate that the two approaches are somewhat complementary and that PhymmBL can use information from both. We compared of the results with all three of these methods, along with results for PhyloPythia, at the 1,000-bp read length (**Table 1**). Both Phymm and PhymmBL gave highly robust, reproducible results: in all cases, the observed standard deviation in accuracy was less than 1%. Mean accuracy results for experiments conducted on 100-bp reads at all levels of the phylogeny along with s.d. for each result are available in **Supplementary Tables 4–6**.

Finally, because more than 1,100 IMM scores were assigned to each read, we considered whether looking beyond the single top-scoring IMM might improve accuracy. We therefore examined the top five scores assigned by PhymmBL to each read in the 1,000-bp test set and counted how often the correct clade appeared in at least one of the top five predictions. This produced a 5–9% increase in accuracy at all levels, as follows (with the accuracy of the top-scoring prediction alone shown in parentheses): phylum, 97.9% (93.8%); class, 96.6% (90.6%); order, 94.5% (86.9%); family, 92.6% (84.8%); and genus, 89.3% (78.4%). This suggests that Phymm and PhymmBL might be improved if they can make use of additional signals that are sufficiently independent from those already detected by Phymm and BLAST to provide additional discriminatory information.

Metagenome data from an acid mine

Evaluating classification methods on real metagenomes can be problematic: ordinarily, the true taxonomic composition of the data cannot be established with certainty. The composition of an acid mine drainage (AMD) metagenome[14], however, has been substantially characterized. It contains three dominant populations, namely the archaeon *Ferroplasma acidarmanus* and two groups of bacteria, *Leptospirillum* sp. groups II and III.

We ran two experiments classifying the AMD data: one using the unaltered RefSeq library as a reference, and one in which we added the draft genomes for all three groups to the RefSeq data to create an augmented reference library. This allowed us to compare, as we did for the synthetic dataset, classification performance for Phymm, BLAST and PhymmBL in the presence of more- and less-specific reference data.

We performed three different characterizations of the AMD data in terms of population distribution. We characterized population breakdown at the phylum level as predicted by PhymmBL across two runs: one using the augmented reference library (including the three new draft genomes) and the other using only the RefSeq data (**Fig. 2**). We also characterized the population breakdown by species given by PhymmBL using the augmented reference library (**Fig. 3**).

In addition to these high-level characterizations, we examined how well, using only the RefSeq library without the added draft genomes, Phymm, BLAST and PhymmBL could characterize the three dominant species groups themselves. We aligned all the AMD reads to the three draft genomes to establish a 'correct' read set for each species group and then examined the classification performance for each of the three methods on each of the three positively identified groups of reads (**Supplementary Figs. 1–9**). PhymmBL correctly predicted the phylum (*Euryarchaeota*) for *F. acidarmanus* 61.0% of the time; it assigned approximately 80% of *Leptospirillum* sp. reads to *Proteobacteria* sp. (Note that members of the phylum (*Nitrospira*) assigned to the *Leptospirillum* genus were originally provisionally assigned to Deltaproteobacteria class[15], and indeed a majority of reads from both groups were labeled with this class: 59.9% for group II, and 51.7% for group III.)

When we analyzed the three positively identified sets of reads using the reference library augmented with the draft genomes for the three groups, PhymmBL's species-level prediction accuracy was 98% or higher for each group (data not shown). We believe these accuracy measurements to be more anecdotal than

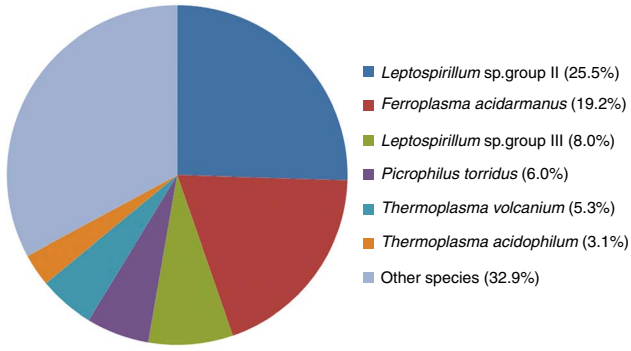

Figure 3 | PhymmBL's species-level population characterization of the AMD data. Data were characterized using the RefSeq-generated IMMs plus IMMs generated from the draft genomes of the three dominant species in the AMD set.

evidentiary, though, owing to the fact that the draft genomes, which we used to train IMMs, were presumably assembled from the same raw reads, which we used as test data for this particular run, although these results do give an indication of PhymmBL's robustness in the presence of sequencing errors and natural mutations, which were certainly present in the raw read data.

DISCUSSION

One advantage of using IMMs over other methods, particularly those that use oligonucleotide counts[16], is that IMMs use information from multiple oligonucleotides of different lengths and integrate the results. Thus instead of having to choose between 5-mers and 6-mers for classification at the class or phylum levels (as is done in PhyloPythia[4]), Phymm can use both. For our experiments we considered oligomers of 1–12 nucleotides, and Phymm automatically selected those oligomers that best characterized each species. Note also that these are comparable to 'spaced seeds' in that the positions from which information is extracted are not necessarily adjacent. The program Glimmer[12] showed the effectiveness of IMMs for a binary classification problem; here we classified reads from hundreds of species, all of which are more closely related to one another than the bacterium and sea squirt pair analyzed by Glimmer, and found that IMMs are also effective at this much more difficult task.

In contrast to other approaches, Phymm classified all reads, and its accuracy at the genus level was 32.8% for 100-bp reads (compared to 6% for CARMA); for 1,000-bp reads, genus-level accuracy was 71.1% (compared to 7.1% for PhyloPythia).

However, as our experiments demonstrate, the best stand-alone method for classifying reads from metagenomics projects, at least when other species from the same genus are known, is BLAST. Although BLAST has been the standard classification method for long reads[7,14], some previous studies excluded it from comparative performance reporting, hindering analysis of their results in the context of all major existing methods. BLAST has shortcomings when a sequence is truly different from anything in the database, as is often the case with actual environmental samples, but this is a universal problem. As has been previously observed[17], marker gene approaches based on 16S ribosomal DNA do not appear to improve classification accuracy over that of BLAST alone. In contrast, our IMM-based method provides a clear boost to BLAST: our hybrid method, PhymmBL, outperformed each of its two component methods alone, and we note that Phymm contributed substantial numbers of correct assignments that BLAST missed and vice versa.

One of the main advantages of our phylogenetic classification method is that preprocessing reads is unnecessary: no gene finding, protein-domain matching or conserved-sequence identification steps were needed, allowing predictions to be made for all reads in a query set without sacrificing accuracy as compared to existing methods. The 1,000 bp 'barrier'[5] did not seem to be a problem for Phymm, although as with all methods, accuracy improved with longer reads. As read lengths for new sequencing technologies increase, the ability to accurately classify short reads, directly as they emerge from the sequencers, will continue to improve. Current metagenomics analysis pipelines[5] postpone classification until after assembly has been attempted, owing to the unreliability of existing composition-based methods at accurately generating phylogenetic classifications for sequences less than 1,000 bp. As shown in the sea squirt and symbiont study[12], accurate binning can improve assembly: methods such as Phymm and PhymmBL (when used as binning tools) should thus improve all downstream analysis of metagenomes, including assembly and gene finding.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank A. Delcher for helpful discussions regarding IMM configuration. This work was supported in part by US National Institutes of Health grants R01-LM006845 and R01-GM083873.

AUTHOR CONTRIBUTIONS

A.B. performed the experiments and subsequent analysis. A.B. and S.L.S. designed the experiments and wrote the paper.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. National Research Council of the National Academies. The dawning of a new microbial age. in *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* p. 2 (The National Academies Press, Washington, DC, 2007).
2. Rondon, M.R. *et al.* Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**, 2541–2547 (2000).
3. Krause, L. *et al.* Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.* **36**, 2230–2239 (2008).
4. McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods.* **4**, 63–72 (2007).
5. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* **72**, 557–578 (2008).
6. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
7. Tringe, S.G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).
8. Tito, R.Y. *et al.* Phylotyping and functional analysis of two ancient human microbiomes. *PLoS One* **3**, e3703 (2008).
9. Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
10. Dinsdale, E.A. *et al.* Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS One* **3**, e1584 (2008).
11. Salzberg, S.L., Delcher, A.L., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**, 544–548 (1998).
12. Delcher, A.L., Bratke, K.A., Powers, E.C. & Salzberg, S.L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
13. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35** Database issue, D61–D65 (2007).
14. Tyson, G.W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
15. Bock, E. & Wagner, M. Oxidation of inorganic nitrogen compounds as an energy source. in *The Prokaryotes,* 3rd edn., vol. 3 (eds., Dworkin, M. and Falkow, S.) 457–495 (Springer, New York, 2006).
16. Chapus, C. *et al.* Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol. Biol.* **5**, 63 (2005).
17. Manichanh, C. *et al.* A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res.* **36**, 5180–5188 (2008).

ONLINE METHODS

Synthetic metagenome data. At present there are no benchmark metagenomics datasets based on real environmental sequences for which the correct taxonomy has been fully and confidently characterized. Indeed, all existing metagenomic samples from natural environments contain multiple unknown species. Following the work of others[18], therefore, we worked with simulated metagenomic samples drawn from existing sequences. For our main data source, we built a core library of all complete bacterial and archaeal genomes available in RefSeq as of October 2008 (ref. 13) comprising 539 distinct species containing 1,146 chromosomes and plasmids. (The full taxonomic composition of this dataset is available in **Supplementary Tables 7–11**, along with the amount of sequence data present for each clade.) All reads used in our experiments with synthetic data were drawn randomly from their source genomes, with no alteration. For the first set of experiments, in which no matches were masked, and comparisons were allowed between reads and their source species, query reads were extracted from each genome, then masked out before training IMMs or building our BLAST database, so that our training set (genomes with extracted reads masked) and our test set (extracted reads) did not overlap. For all other experiments, IMMs and BLAST databases were constructed using whole genomes, and separation of training and test data was automatic: the entire genome of the species from which each query read was drawn was masked, by design, during the classification process.

Synthetic test set construction and filtering. To control for under-representation of some clades in the available data, query sets were filtered so that all species under consideration had at least two sister species within the clade under consideration. For example, in the experiment that masked exact species matches but allowed intragenus comparisons, without this filtering step, if a given species was the only sequenced representative of its genus, then it would have been impossible to assign a correct genus label to reads from that species (because the species itself was excluded from the scoring process by design), which in turn would artificially depress the results. Analogous filters were used at higher levels; for example, phylum-level predictions were made only for reads which had at least two other species in their home phylum.

Each synthetic test set initially contained 5 randomly-selected "reads" from each of the 1,146 chromosomes and plasmids in the RefSeq reference data, totaling 5,730 reads representing 539 bacterial and archaeal species. After filtering out species that did not meet the criteria above, test set sizes for the species-masked, genus-masked, family-masked, order-masked and class-masked experiments contained 2,870, 3,255, 3,335, 4,575 and 4,390 reads, respectively.

The sparsest of these, the experiment in which exact species matches were masked and the lowest-level predictions were at the genus level, represented 573 chromosomes and plasmids from 254 species across 48 genera. This is still a very broad group of genera compared to previous studies; for example, another study used 31 genera[4].

Acid mine drainage test set. We downloaded the entire set of raw sequence reads for the AMD metagenome dataset presented in ref. 14 from the NCBI trace archive. Vector sequences were removed from reads using Figaro[19]. Each read was then truncated to include only the longest contiguous stretch of bases for which base-calling quality scores were at least 17. Finally, reads less than 100 bp were discarded. Of the original 180,713 raw reads in the dataset, 166,345 remained after all quality filtering was complete.

We established 'true positives' for reads belonging to *F. acidarmanus* and the two *Leptospirillum* sp. groups by using MUMmer[20] to align all reads in the dataset to the draft genomes for the three species groups. Three sets of positive matches, one for each species group, were identified by selecting reads which aligned to exactly one of the three draft genomes. When complete, the *F. acidarmanus* set contained 15,628 reads; the *Leptospirillum* sp. group II set contained 48,589 reads; and the *Leptospirillum* sp. group III set contained 10,104 reads. These sets were used to generate the per-group predictive accuracy results presented in **Supplementary Figures 1–9**.

Classification infrastructure. We built one IMM per molecule (chromosome or plasmid), with each IMM trained on the entire molecular input sequence, yielding a total of 1,146 IMMs. To compare our method to BLAST on the same data, we also constructed a BLAST database containing all 1,146 molecular sequences.

In designing PhymmBL, we attempted to boost accuracy by adding several common composition-based measures including G+C content and dinucleotide frequencies[21], but none was found to improve accuracy, likely because the statistics captured by these measures closely overlap those already captured by the Phymm IMMs.

Impact of training data size on classification accuracy. Taxonomic groups are represented in sequence databases to vastly different degrees; among the RefSeq genera we used for this study, for example, the amount of sequence data differed by up to two orders of magnitude. To explore potential biases in classification accuracy resulting from such a wide range in the volume of available training data for each clade, we plotted clade-specific classification accuracy as a function of training set size. Results at the phylum level for 100-bp queries, classified by Phymm (with exact species matches excluded) are given in **Supplementary Figure 10**. This analysis was conducted 10 times to establish variance in accuracy; a tabular version of **Supplementary Figure 10**, including s.d. in accuracy for each phylum, along with analogous results at the class, order, family and genus levels are available in **Supplementary Tables 7–11**.

Somewhat unexpectedly, little if any correlation was observed between the amount of sequence data available for IMM training and predictive power. Also unexpectedly, the s.d. for predictive accuracy, which one might expect to vary widely and exhibit correlations with amount of training data, exhibited neither of these trends, with deviations ranging from less than 1% to approximately 17%, with no apparent correspondence between variance and amount of training data. We hypothesize that accuracy is far more dependent on evolutionary diversity (that is, mutational diversity) within each clade: a phylogenetic group containing species that are more evolutionarily distant from one another will be inherently harder to classify, using methods based on sequence composition, than one containing more closely related species.

IMMs and phylogenetic classification. Interpolated Markov models are a form of Markov chain that uses a variable number of states to compute the probability of the next state. IMMs, which are also called variable-order Markov models, were described in detail in the original Glimmer papers[11,22]. For our purposes, the main idea is that IMMs can be used to classify sequences based on patterns of DNA distinct to a clade, whether the clade is a species, genus, or higher-level phylogenetic group. During training, the IMM algorithm constructs probability distributions representing observed patterns of nucleotides that characterize each species. The model used by Glimmer and Phymm captures nonadjacent patterns when necessary; for example, it can use 8 positions spaced across a window of 12 bases. During classification, we used each IMM as a scoring method: it examines the nucleotides in a given query sequence and outputs a score corresponding to the probability that the query was generated from the same distribution as that used to train the IMM.

Parameter settings for weighted voting in PhymmBL. For PhymmBL, we empirically we determined the combined score using the function Score = IMM + 1.2 (4 − log(E)), where IMM is the score from the best-matching IMM and *E* is the smallest (best) *E*-value returned by BLAST. Our IMMs return log-likelihood scores as integers, generally in the range between −500 and −100, with higher scores representing better matches; the log-transformation brought the BLAST scores into this scale. The constant 4 was determined experimentally to be optimal via binary search on small positive integers, and the weight of 1.2 was subsequently determined to be optimal via binary search on values between 1 and 3. The ranges for both searches (integers between 0 and 5 to find the additive constant 4, and values between 1 and 3, in increments of 0.1, to determine the multiplicative weight of 1.2) were established by identifying values at which the predictions of one method completely dominated those of the other. For example: multiplicative weights less than 1 generated combined scores essentially identical to those produced by the IMMs alone, while those greater than 3 generated combined scores which were the same as those produced by BLAST by itself. These settings may only represent a local optimum, but different values of the weights had only a marginal effect on overall accuracy.

Confidence scores. We conducted preliminary experiments with the goal of correlating raw Phymm and PhymmBL scores with predictive accuracy, to establish a function mapping these scores to the probability of generating a correct taxonomic prediction. A smooth, monotonic function mapping score to accuracy was indeed observed for reads 100 bp in length, but reads of different lengths exhibited different score ranges, and maps between score and predictive power across various lengths were not scaled, one to another, in any obvious (constant or linear) way. While we believe such a relationship can be established, more complex investigation will be required in order to properly determine a closed-form relationship between raw scores and predictive accuracy.

Data availability. A one-stop installer for the Phymm and PhymmBL system is available as **Supplementary Software** and at http://www.cbcb.umd.edu/software/phymm/ (where software updates will be released). The code is available along with a readme describing system requirements and configuration. Copies of all synthetic metagenomic data described in this paper are available for download and are linked off the Phymm and PhymmBL download page.

18. Mavromatis, K. *et al.* Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods.* **4**, 495–500 (2007).
19. White, J.R., Roberts, M., Yorke, J.A. & Pop, M. Figaro: a novel statistical method for vector sequence removal. *Bioinformatics.* **24**, 462–467 (2008).
20. Delcher, A.L., Salzberg, S.L. & Phillippy, A.M. Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinformatics* chapter 10, unit 13 (2003).
21. Karlin, S. & Burge, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**, 283–290 (1995).
22. Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).

doi:10.1038/nmeth.1358