# Accurate phylogenetic classification of variable-length DNA fragments

Alice Carolyn McHardy[1], Héctor García Martín[2], Aristotelis Tsirigos[1], Philip Hugenholtz[2] & Isidore Rigoutsos[1]

**Metagenome studies have retrieved vast amounts of sequence data from a variety of environments leading to new discoveries and insights into the uncultured microbial world. Except for very simple communities, the encountered diversity has made fragment assembly and the subsequent analysis a challenging problem. A taxonomic characterization of metagenomic fragments is required for a deeper understanding of shotgun-sequenced microbial communities, but success has mostly been limited to sequences containing phylogenetic marker genes. Here we present PhyloPythia, a composition-based classifier that combines higher-level generic clades from a set of 340 completed genomes with sample-derived population models. Extensive analyses on synthetic and real metagenome data sets showed that PhyloPythia allows the accurate classification of most sequence fragments across all considered taxonomic ranks, even for unknown organisms. The method requires no more than 100 kb of training sequence for the creation of accurate models of sample-specific populations and can assign fragments $\geq$1 kb with high specificity.**

The emerging field of metagenomics is dedicated to the study of sequences obtained by high-throughput sequencing of DNA samples from microbial communities. The approach has already provided exciting insights into the lifestyle, evolution and characteristics of microbial organisms[1–3]. Such insights could not have been obtained otherwise as the vast majority of microbes cannot be cultivated using standard techniques[4]. From a technical standpoint, the field has created new computational challenges. These include a need for assembly and gene-finding programs able to handle highly diverse sequence collections of organisms sampled with different coverage, and tools for characterizing the phylogenetic provenance of the vast amounts of generated short sequences.

One approach to classifying metagenomic sequence fragments is the use of 'marker genes', for example ribosomal RNAs (rRNAs), as phylogenetic anchors for identification of the source organism of a fragment. rRNAs are highly conserved and allow the most accurate placement within the tree of life of an organism or fragment containing the respective marker. See refs. 5,6 for the original observation and the resulting framework for the quantification of evolutionary relationships using rRNAs. Even though the marker-gene collection has been expanding through the inclusion of ubiquitous and slowly evolving or clade-specific proteins[7–9], the approach permits the characterization of only a limited number of fragments. Sensitive identification (to species level) using a marker-gene approach requires a very large database; presently the largest available database is for 16*S* rRNA[10]. In samples like the enhanced biological phosphorus-removing (EBPR) sludge[11], Sargasso Sea[1] and Minnesota soil[2] samples, however, a mere 0.17%, 0.06% and 0.017% of the contigs, respectively, carry 16*S* rRNA markers. Even if one includes other markers such as *recA* and *rpoB*, less than 1% of the contigs in a metagenomic assembly are identifiable with a marker-gene approach.

For a very low complexity community found in acid-mine drainage[3], whose dominant species exhibited considerable differences in their G+C content, the organism-specific 'binning' of fragments based on G+C content and read coverage retrieved near-complete genomes. The use of tetranucleotide signatures has also shown promise in the characterization of low-complexity communities, and for the dominant organisms of more complex populations[12]. However, these schemes are unable to characterize more diverse, and thus more challenging, metagenomes such as the Sargasso Sea sample, or samples of the highly complex soil communities that are estimated to contain millions of distinct taxa[13]. For an extremely complex sample of Minnesota soil, only a gene-centric characterization could be carried out, as less than 1% of reads could be assembled. Gene-centric analyses allow determination of genes important for the overall community function, but the taxonomic composition of the sample remains largely unresolved. It is thus imperative that fast and accurate tools be developed that allow the taxonomic characterization of short genomic sequence fragments and permit more comprehensive metagenome analyses.

Genomic sequence composition has been shown to reflect organism-specific characteristics and is dubbed the 'genome signature'[14–16]. The phenomenon is sufficiently pronounced to allow the simultaneous supervised or unsupervised discrimination among several different species[12,17–19]. Given the availability of near-complete genomes for training, 85% accuracy has been reported for fragments as short as 400 bp from a mixture of 28 organisms. Furthermore, genomic signatures also carry

[1]Bioinformatics and Pattern Discovery Group, IBM Thomas J Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, New York 10598, USA. [2]US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. Correspondence should be addressed to I.R. (rigoutso@us.ibm.com).

RECEIVED 4 APRIL; ACCEPTED 13 OCTOBER; PUBLISHED ONLINE 10 DECEMBER 2006; DOI:10.1038/NMETH976

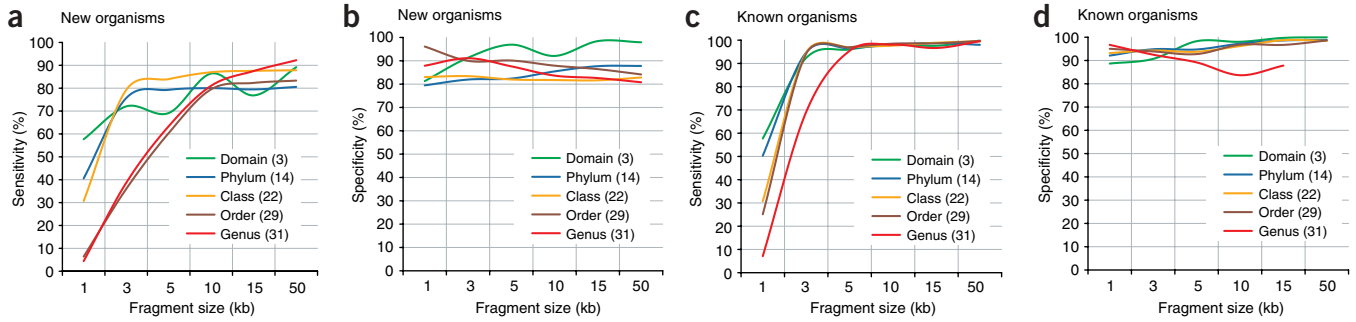


**Figure 1** | Accuracy of phylogenetic assignments for differently sized genomic fragments with PhyloPythia. (**a**,**b**) classification accuracy for fragments from unknown organisms (of which no fragments were included in the training data for the classifier). (**c**,**d**) Classification accuracy for organisms of which some fragments are known (other genomic fragments than the ones tested were included in the training data for the classifier). Class-normalized sensitivity (**a**,**c**) and specificity (**b**,**d**) of phylogenetic assignments for fragments from 340 different organisms (**Supplementary Methods**). The numbers in parentheses indicate the number of modeled clades for the phylogenetic classifiers.

phylogenetic information, as recent studies have shown[20–22]. In this context, a method based on a self-organizing map (SOM) has been described as being able to cluster fragments of completely sequenced genomes, by organism and higher-level phylotypes, with high accuracy[22].

Here we present PhyloPythia, a method that uses sequence composition to phylogenetically characterize sequence fragments. The method is named after Pythia, the priestess at Apollo's oracle in ancient Delphi. PhyloPythia uses a multiclass support vector machine (SVM) classifier with the oligonucleotide composition of variable-length genome fragments as the input space. It is the first method that uses a state-of-the-art technique for supervised classification of high-dimensional, sparse input data (that is, oligonucleotide counts in relatively short sequence fragments) to solve the problem of phylogenetic assignment to known clades (dominant sample populations or higher level clades). PhyloPythia allows the accurate phylogenetic classification of genomic fragments $\geq$1–3 kb for all taxonomic ranks considered (domain, phylum, class, order and genus). More importantly, PhyloPythia can also achieve this for fragments originating from new organisms. In addition to the generic higher-level classifiers derived from the completed genomes, PhyloPythia allows the inclusion of sample-specific clades trained with marker-gene carrying fragments.

We used PhyloPythia to analyze three metagenomes: the Sargasso Sea sample and two samples of EBPR-sludge used in industrial wastewater processing[11]. For all analyzed samples, our method has substantially improved performance over earlier approaches, accurately assigning fragments $\geq$1 kb to the correct clades. For the dominant populations of the samples, PhyloPythia accurately retrieved fragments even with as little as 100 kb of training sequence. Moreover, it confidently assigned genomic fragments that could not be characterized otherwise. Finally, PhyloPythia automatically characterized large portions of the processed samples at higher taxonomic levels, in agreement with existing marker-gene studies of sample composition.

## RESULTS

### Classifier building

We performed an extensive evaluation with the genome sequences of 340 organisms and identified the optimal sequence sources, parameter settings and model architecture for composition-based taxonomic classification. Coding sequence and genomic fragments carry (slightly different) types of taxonomic information (**Supplementary Note** and **Supplementary Fig. 1** online). As the assignment accuracy depends critically on the amount of available sequence, it is advantageous to use complete genomic fragments for composition-based phylogenetic assignment. This choice also obviates the requirement for prior gene identification.

We also searched extensively for the best oligonucleotide pattern space (**Supplementary Table 1** and **Supplementary Fig. 2** online). The lower-ranking clades from the genus to the class can be discriminated best using 5-mers of contiguous nucleotides. For clades at the ranks of phylum and domain, more complex 6-mer signatures (($w$ =6, $l$ = 4) and ($w$ = 6, $l$ = 6), with $w$ being the pattern width and $l$ the number of literal characters, respectively) are needed to best capture the characteristics of a joint ancestry.

Our evaluation of different kernel functions showed that for feature spaces with more than $w^2$ dimensions, the Gaussian kernel outperforms the linear one (**Supplementary Table 2** online).

We also extensively investigated the relation between fragment length and classification accuracy for training and testing fragments. This is necessary as assembled metagenomes contain fragments whose lengths range from several megabases down to $\leq$700 bp for individual reads, or even $\leq$100 bp if pyrosequencing is used[23] (**Supplementary Note** and **Supplementary Fig. 3** online).

PhyloPythia incorporates the results of these analyses and can process fragments of all lengths. A query fragment is examined by applying a series of classifiers trained with fragments of decreasing lengths. This continues until the query is assigned or we reach a classifier that was trained with shorter fragments than that respective query. If the query cannot be assigned to a known clade, it is assigned to the class 'other'. At the domain level there is no broadly defined 'other' class to which items with unclear signal can be assigned. We thus make an exception to the above scheme for domain assignments and apply a classifier trained with fragments of similar length to the query.

### Assignment accuracy

We evaluated assignment accuracy for the sequences of known and unknown organisms to the classifier. For fragments of unknown organisms, PhyloPythia correctly assigns between 79–96% for all tested lengths and taxonomic ranks (**Fig. 1** and **Supplementary Table 3** online). This specificity is matched by a high sensitivity ('microaccuracy') for all fragments $\geq$5 kb (**Fig. 1**). Only for

fragments shorter than 5 kb, and in particular for those shorter than 3 kb, the sensitivity decreases markedly. It is important to stress that this high accuracy is achieved based on the clade-specific signal that was learned from other organisms of a given clade only. This takes place in a setting where a fragment can be assigned to any one of 31 different clades (**Fig. 2**) and in the presence of considerable 'noise' from fragments of organisms of unknown clades. Notably, the latter fragments are accurately identified as 'unknown' in most cases.

Accuracy increases further when assigning fragments from known organisms. For fragments $\geq$3 kb, the sensitivity and specificity are 90–99% for clades from the rank of domain to order (**Fig. 1** and **Supplementary Table 4** online). For fragments as short as 1 kb, that is, only slightly longer than a single read, specificity reaches 88.7–96.7%, whereas sensitivity ranges from 7.1% at the genus level to 57.7% at the domain level.

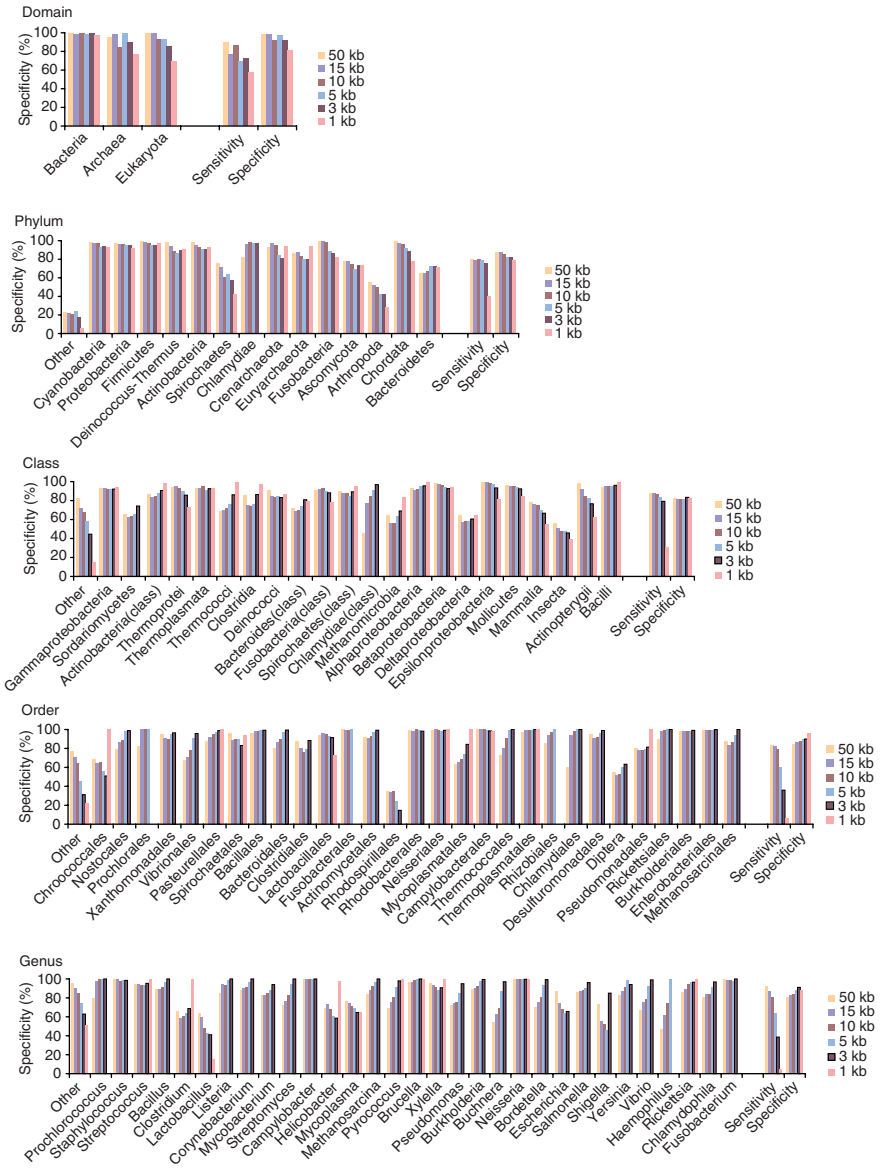


**Figure 2** | Phylogenetic classification accuracy of PhyloPythia by clade for differently sized genomic fragments from unknown organisms. From top to bottom, the clade-specific specificity, class-normalized sensitivity and specificity (**Supplementary Methods**) is shown for fragments of 340 organisms at the ranks of domain, phylum, class, order and genus (rank is given above each graph). The legend specifies the color code for the respective fragment length in the graphs.

## Impact of nontaxonomic signals

Genome sequence composition is shaped by many factors, including translational selection exerted on the synonymous codon usage of the protein-coding sequences[24], lateral gene transfer and the optimal growth temperature of the organism[25]. As our approach uses a suitable supervised classification technique, it is able to learn the characteristics that are relevant for taxonomic classification from an input space that is also shaped by nontaxonomic influences. For instance, PhyloPythia is able to accurately distinguish, in most cases, genomic fragments of different domains for both thermophiles and nonthermophiles (**Supplementary Fig. 4** online).

We also examined the impact of translational selection on classification accuracy by comparing for unknown organisms the accuracy of classifying 3 kb fragments to the accuracy for ribosomal protein–carrying 3-kb fragments (**Supplementary Fig. 5** online): PhyloPythia also achieves a similarly high accuracy for both types of fragments from the genus to the phylum level and a specificity of 83–92%.

## Characterizing the EBPR-sludge metagenomes

We used PhyloPythia to characterize two metagenomic samples of lab-scale EBPR sludge, obtained from Madison, Wisconsin, USA and Brisbane, Australia (referred to as US and OZ, respectively). For these, assemblies between 20 and 28 Mbps of sequence each were generated using the PHRAP (US and OZ data) and JAZZ (US data) assemblers[11].

Both communities were dominated by the uncultured bacterium, *Candidatus Accumulibacter phosphatis* (CAP). A 16*S* rRNA–based analysis revealed that CAP is the only species common to both communities above the detection threshold. However, overlap exists at higher phylogenetic levels[11].

We extended the multiclass, higher-level phylogenetic classifiers with models of the dominant sample populations. In particular, we created an order-level, multi-class model with an additional clade for the Rhodocyclales, based on genomic sequence fragments from *Dechloromonas aromatica* and known CAP sequences, and sample-specific lower-level clades for CAP and a *Thiothrix* sp.–like population abundant in the OZ sample. All three assemblies were characterized with PhyloPythia using this extended model architecture.

To test validity, we examined the consistency (nesting) of assignments for the fragments at the different taxonomic ranks.

PhyloPythia's predictions are consistent: 93–97.7% of all assignments (99.6–99.8% for high-confidence assignments) nest across all ranks (**Table 1**). Additionally, we analyzed the assignments for the 54 rRNA–containing fragments and found all assignments to be correct for fragments >2 kb: for fragments shorter than 2 kb specificity was 86% (95% for high-confidence assignments).

At the phylum level, culture-independent analyses based on 16*S* rRNA indicated that the samples are dominated by CAP and other

**Table 1** | Phylogenetic characterization of two metagenome samples from phosphorus removing sludge

| Sample | US JAZZ | US PHRAP | OZ PHRAP |
|---|---|---|---|
| Fragments | 5,426 | 16,370 | 11,632 |
| Sequence (Mb) | 20.6 | 28.7 | 26.9 |
| Assigned (high confidence) | 3,444 (2,905)[a] | 12,782 (11,459) | 9,908 (9,062) |
| Consistency (%) | 97.7 (99.8) | 96.7 (99.6) | 93 (99) |
| Known CAP fragments | 7 | 665 | 584 |
| $Sn_{fragments} \cdot CAP$ (%) | 100 (85.7) | 86.2 (49.2) | 79.8 (52.7) |
| $Sn_{kb} \cdot CAP$ (%) | 100 (97.1) | 97.5 (80.1) | 94.9 (73.6) |

| Clade | Fragments assigned (kb) | Fragments assigned (%) | Fragments assigned (kb) | Fragments assigned (%) | Fragments assigned (kb) | Fragments assigned (%) |
|---|---|---|---|---|---|---|
| Bacteria | 18,916 | 91.8 | 25,450 | 89 | 24,537 | 91.2 |
| Not assigned (other) | 1,635 | 7.9 | 3,100 | 11 | 1,576 | 5.9 |
| Archaea | 38 | 0.2 | 106 | 0.4 | 691 | 2.6 |
| Eukaryota | 9 | 0 | 75 | 0.3 | 85 | 0.3 |
| Proteobacteria | 15,406 | 74.8 | 17,965 | 63 | 17,797 | 66.2 |
| Not assigned (other) | 3,814 | 18.5 | 9,086 | 32 | 6,283 | 23.4 |
| Actinobacteria | 580 | 2.8 | 663 | 2.3 | 295 | 1.1 |
| Firmicutes | 429 | 2.1 | 456 | 1.6 | 354 | 1.3 |
| Bacteroidetes | 299 | 1.5 | 319 | 1.1 | 568 | 2.1 |
| Euryarchaeota | 21 | 0.1 | 74 | 0.3 | 649 | 2.4 |
| Spirochaetes | 27 | 0.1 | 52 | 0.2 | 695 | 2.6 |
| Cyanobacteria | 8 | 0 | 35 | 0.1 | 111 | 0.4 |
| Deinococcus-Thermus | 3 | 0 | 39 | 0.1 | 107 | 0.4 |
| Betaproteobacteria | 11,129 | 54 | 13,085 | 46 | 9,150 | 34 |
| Not assigned (other) | 5,827 | 28.3 | 13,253 | 46 | 10,254 | 38.1 |
| Gammaproteobacteria | 2,576 | 12.5 | 1,229 | 4.3 | 5,190 | 19.3 |
| Actinobacteria (class) | 460 | 2.2 | 340 | 1.2 | 150 | 0.6 |
| Clostridia | 284 | 1.4 | 196 | 0.7 | 60 | 0.2 |
| Alphaproteobacteria | 245 | 1.2 | 442 | 1.5 | 646 | 2.4 |
| Bacilli | 14 | 0.1 | 28 | 0.1 | 106 | 0.4 |
| Bacteroides (class) | 18 | 0.1 | 17 | 0.1 | 127 | 0.5 |
| Mollicutes | 22 | 0.1 | 60 | 0.2 | 31 | 0.1 |
| Spirochaetes (class) | 14 | 0.1 | 37 | 0.1 | 563 | 2.1 |
| Deinococci | 0 | 0 | 2 | 0 | 38 | 0.1 |
| Deltaproteobacteria | 0 | 0 | 28 | 0.1 | 193 | 0.7 |
| Epsilonproteobacteria | 6 | 0 | 6 | 0 | 94 | 0.3 |
| Methanomicrobia | 0 | 0 | 5 | 0 | 283 | 1.1 |
| Rhodocyclales | 9,948 | 48.3 | 11,020 | 38 | 7,507 | 27.9 |
| Not assigned (other) | 9,233 | 44.8 | 17,351 | 60 | 14,427 | 53.6 |
| Xanthomonadales | 1,218 | 5.9 | 194 | 0.7 | 361 | 1.3 |
| Burkholderiales | 84 | 0.4 | 44 | 0.2 | 76 | 0.3 |
| Actinomycetales | 59 | 0.3 | 28 | 0.1 | 3 | 0 |
| Pseudomonadales | 49 | 0.2 | 65 | 0.2 | 8 | 0 |
| Spirochaetales | 0 | 0 | 9 | 0 | 90 | 0.3 |
| Thiotrichales | | | | | 4318 | 16.1 |
| Not assigned (other) | 10,738 | 52.1 | 18,053 | 63 | 15,771 | 58.6 |
| Accumulibacter | 9,861 | 47.9 | 10,680 | 37 | 6,801 | 25.3 |
| Thiothrix | | | | | 4,318 | 16.1 |

US JAZZ and US PHRAP are assemblies with the JAZZ and PHRAP assembler of the US data set, respectively. OZPHRAP is the PHRAP-assembled OZ data set. All phylogenetic clades assigned ≥50 kb are shown. From top to bottom, the percentage of fragments assigned to clades at the ranks domain, phylum, class, order and genus is shown. $Sn \cdot CAP$ gives the sensitivity per fragment ($Sn_{fragments} \cdot CAP$) and per kilobase ($Sn_{kb} \cdot CAP$) of recovering known CAP sequences in a cross-validation experiment in which the CAP models were trained from the fragments of one sample and applied for classification of the other sample. [a]Values for high confidence assignments ($P \geq 0.85$) are given in brackets.

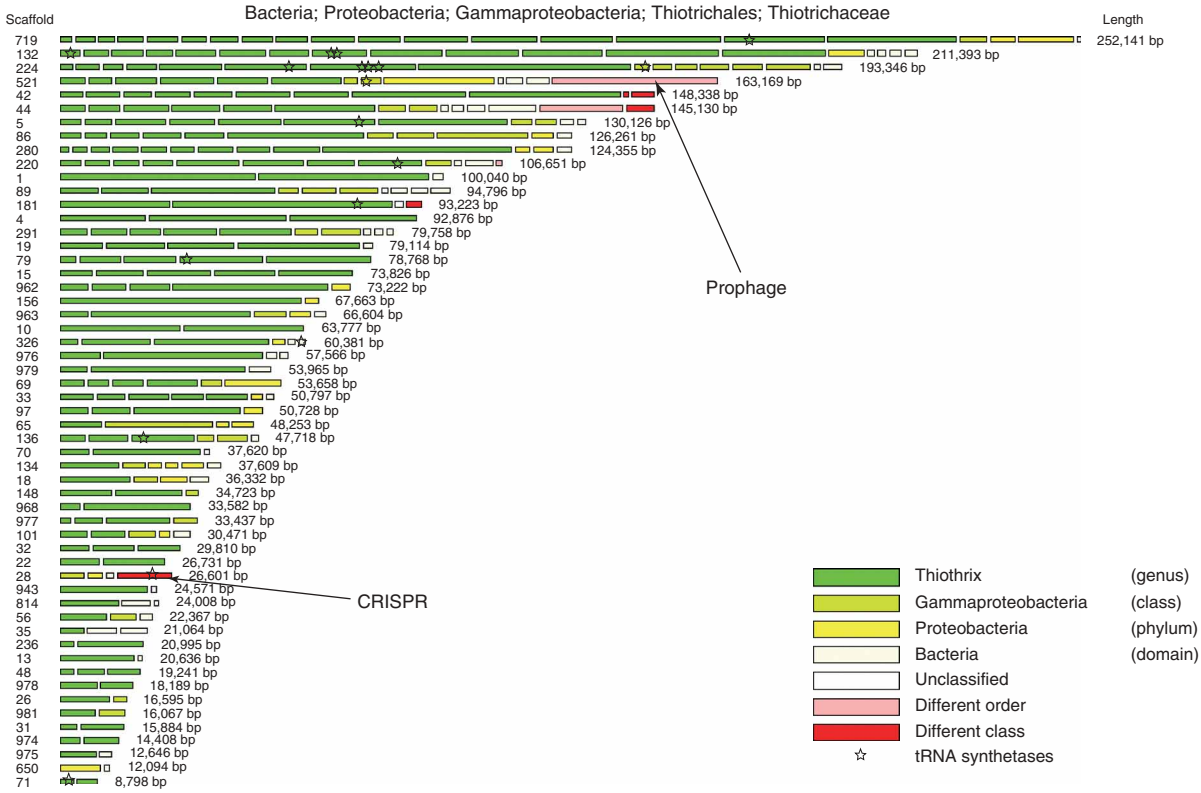


**Figure 3** | Binning accuracy of *Thiothrix* sp. contigs using PhyloPythia. Each line represents a scaffold, which is a collection of contigs (boxes) linked by end pair read information, indicating that those contigs belong to the same genome. The colors indicate the most specific taxonomic rank to which a contig could reliably ($P > 0.85$) be assigned by PhyloPythia. The majority of contigs were identified as belonging to *Thiothrix* (dark green 62%) or a consistent lower level classification (yellow to light green, 35%), with only 3% being unclassified (white) or misclassified (red). Some misclassifications could be correlated with atypical sequence composition due to laterally transferred genes (for example, prophage) or noncoding repeat sequences (for example, CRISPR elements). The number of distinct tRNA synthetases found in the *Thiothrix* scaffold set (indicated by stars) can be used as a proxy for genome completeness. We estimate that 72% of the *Thiothrix* genome has been recovered based on the presence of 13 of 18 tRNA synthetase types on high confidence *Thiothrix* contigs. Only scaffolds longer than 12 kb or containing a tRNA synthetase are shown.

proteobacterial species, flanked by much less abundant species belonging to the Bacteroidetes, and for the OZ sludge only, representatives of the Firmicutes, Verrucomicrobia and Chlorobi phyla[11]. PhyloPythia's assignments correlate well with this community structure (**Table 1**). The majority of fragments of both samples were assigned to the Proteobacteria (63–74.8%). Notably, PhyloPythia also assigned a small fraction of fragments from both samples to the Actinobacteria, which is supported by identification of a partial rRNA gene in the US sludge. Several fragments were assigned to Spirochaetes and Euryarchaeaota that were not found by 16*S* rRNA analysis, providing testable hypotheses about community structure.

At the rank of order, apart from the Rhodocyclales (which comprises CAP), the Xanthomonadales were identified as one of the more frequent clades in both samples, in agreement with the marker gene–based studies (**Table 1**).

For *Accumulibacter phosphatis*, the relative percentage of assigned fragments qualitatively agrees with the rRNA-derived abundance estimates. As a third test, we used the fact that both sludge communities have CAP in common to provide an estimate of the *Accumulibacter* sp. binning accuracy. We constructed CAP-specific classifiers from the known CAP genome fragments of one sample and evaluated the success of recovering known CAP fragments in the other sample. PhyloPythia mainly missed very short fragments, successfully recovering 95–100% of the known fragments for all assemblies (**Table 1**); 74–97% could be assigned with high confidence ($P \geq 0.85$).

The 16*S* rRNA markers also indicate that a *Thiothrix*-like species is relatively well represented in the OZ data set (13.8% of OZ PHRAP contigs containing 16S rRNA genes). The genus *Thiothrix* belongs to the Gammaproteobacteria and, presently, no larger fragments of genomic sequence are available. Based on an initial set of 17 fragments (0.7 Mb), PhyloPythia retrieved an additional 3.7 Mb of *Thiothrix* sequences. We verified these assignments by using the scaffolding information provided by read pairs, that is, contiguous sequence fragments (contigs) linked by end reads of the same cloned insert. Overall, 97% of PhyloPythia's *Thiothrix* assignments were consistent (**Fig. 3**); the remaining 3% of the fragments were either misclassified or not assigned at all, and mainly comprise small contigs with reduced composition signal, larger contigs that contain laterally transferred genes (for example, prophage) or repeat structures like CRISPR[26]. The majority of contigs in these scaffolds are either classified as *Thiothrix* with high confidence (62%), or assigned to consistent higher taxonomic levels with high confidence (35%). Based on the number of distinct tRNA synthetases identified in the *Thiothrix* contigs (**Fig. 3**), we estimate that

**Table 2** | Classification accuracy for contigs from the Sargasso Sea metagenome sample

| | Number of contigs | $Asg_{ctg}$ | $Sp_{ctg}$ | Contig sequence (bp) | $Asg_{nt}$ | $Sp_{nt}$ | Number of contigs | $Asg_{ctg}$ | $Sp_{ctg}$ | Contig sequence (bp) | $Asg_{nt}$ | $Sp_{nt}$ | $\Delta Sp_{ctg}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dominant sample populations | 19 | 68.4 | 100.0 | 361,704 | 97.33 | 100.00 | 69 | 53.62 | 100.00 | 1,994,724 | 96.3 | 100.00 | 0 |
| Genus level | 23 | 52.1 | 91.7 | 362,687 | 92.7 | 98.53 | 79 | 34.18 | 78.12 | 1,940,738 | 94.03 | 97.21 | 13.55 |
| Order level | 33 | 36.3 | 91.7 | 401,613 | 83.74 | 98.53 | 136 | 23.53 | 68.42 | 2,091,355 | 88.28 | 96.38 | 23.25 |
| Class level | 86 | 26.7 | 64.0 | 521,725 | 73.66 | 91.69 | 333 | 27.33 | 62.5 | 2,632.674 | 80.3 | 93.48 | 1.5 |
| Phylum level | 114 | 36.0 | 67.4 | 578,482 | 74.17 | 90.6 | 393 | 52.1 | 47.95 | 2,754,514 | 86.63 | 86.66 | 19.49 |
| Domain level | 140 | 97.9 | 94.9 | 655,980 | 99.43 | 98.2 | 512 | 98.44 | 88.69 | 3,148,484 | 99.55 | 95.55 | 6.2 |

From top to bottom, for the dominant sample populations to higher-level clades, the percentage of assigned sequence (Asg) in terms of contigs (ctg) and per base (bp) is shown, next to the specificity (Sp., percentage of correct assignments). The reference contigs $\geq$1 kb were determined based on the presence of a small subunit rRNA marker (columns 1–6), or also including contigs linked by a scaffold to these contigs (columns 7–12). Column 13 gives the difference in specificity for the first relative to the second set, which (mostly caused by inconsistent assignments for the low coverage sequences) indicates assembly issues.

72% of the *Thiothrix* genome was recovered, and that its genome is ~6 Mb in size. Such estimates of individual population genome size and coverage within metagenomic data sets are useful for estimating the reliability of metabolic reconstruction and for guiding additional sequencing efforts.

In summary, PhyloPythia assigned approximately 90% of the fragments from the two sludge samples at the domain level, 70–81.5% at the phylum level, 61.9–71.1% at the class level, 40–65.5% at the order level, and 37–47.9% to sample-derived *A. phosphatis* and *Thiothrix*-like population bins. Based on the sample-specific models, known fragments of CAP could be retrieved with high accuracy, and 3.7 Mb of additional genomic sequence were assigned to the *Thiothrix*-like organism, a deep-branching Gammaproteobacterium, for which no sequence data are available presently.

### Characterizing the Sargasso Sea metagenome

We used PhyloPythia to characterize the recently reported Sargasso Sea sample[1]. With more than 800,000 contigs totaling >1 Gb of sequence, it is the largest metagenome set available. We used 463 annotated small subunit rRNA genes as a reference for evaluation. Based on taxonomic assignments generated for the corresponding contigs (and contigs linked via a scaffold) of these genes, we quantitatively assessed assignment accuracy on real metagenome data.

As before, we combined models of the dominant sample populations with the generic higher-level models created from the completed genomes. We used between 100 and 162 kb of sequence to train sample-specific population models for each of the four dominant populations (*Prochlorococcus*, unidentified Gammaproteobacteria, *Shewanella* and *Burkholderia* spp.). We used the resulting multilevel model to characterize all 811,371 contigs of the sample.

Notably, we found that the specificity deteriorates slightly when the reference is extended by contigs that are linked to the marker-gene carrying contigs via a common scaffold. Mostly, this is due to false assignments of such scaffold-linked contigs from the low-coverage assembly (**Table 2**). This indicates assembly issues for the data, as has been described earlier for scaffolds containing both archaeal and bacterial rRNA markers[27]. Generally, however, our *in silico* estimates map reasonably to a real metagenome data set. For the more reliable smaller reference set, 68–98% of the contigs $\geq$1 kb (representing 97–99% of the total sequence) could be assigned, and of these, 64–100% of the contig assignments were correct (91–100% per nucleotide). For sequences <1 kb, the accuracy of our method decreased, as for the sludge samples.

For the binning of dominant populations, we compared PhyloPythia to a TETRA-like method. TETRA is specific, allowing the recovery of most large fragments. However, we found that PhyloPythia was substantially more sensitive, in particular for shorter fragments and for fragments of the higher-level Gammaproteobacterial sample population, without sacrificing specificity. PhyloPythia's shortest assignment was a 1.5-kb fragment, whereas TETRA's shortest assignment was 12 kb in length (**Supplementary Tables 5** and **6** online).

We also compared our results to the phylotypes derived from the association of fragments from known organisms with sample fragments on a SOM. The SOM has been described as able to cluster input fragments from completely sequenced genomes with high accuracy by organism and higher-level phylotypes. In our analysis of the high coverage fragments of the Sargasso Sea sample, we found that the SOM correctly assigned only nine fragments of the dominant *Shewanella* and *Burkholderia* species populations, for which nearly identical genomes have been already sequenced. For the extended reference of sequences $\geq$1 kb, the specificity of the SOM assignments decreased to 16%, compared to 48% for PhyloPythia (**Supplementary Tables 5** and **6**). We surmise that the SOM may require large input sets of sequences for a particular organism to place these in a consistent cluster (as in the case of multiple fragments from organisms with completed genomes) and that its generalization ability for less abundant input species is diminished. Generally, a supervised classification procedure like the SVM, which creates clade models in a directed fashion based on phylogenetically relevant features from the high-dimensional feature space of sequence composition, is likely to perform better in a classification problem than an unsupervised procedure that has no knowledge of relevance of features.

The Sargasso Sea sample contains large amounts of a cyanobacterial *Prochlorococcus* population (one of the four dominant populations), one of the most abundant microorganisms found in the oceans. Cyanobacteria evolve rapidly in terms of their sequence composition, and the known *Prochlorococcus* spp. differ by as much as 27% in their average genomic G+C content. For the completely sequenced genomes, PhyloPythia shows considerable accuracy in identifying fragments of unknown Cyanobacteria (**Supplementary Table 3**). PhyloPythia achieves a sensitivity of >75% for all fragments >3 kb and a specificity of more than 93% across all fragment lengths. For Proteobacteria, PhyloPythia's accuracy

**Table 3** | Phylogenetic characterization of sequences from the Sargasso Sea metagenome sample

| Rank | Number of contigs | Contigs assigned (%) | Contig sequence (Mb) | Sequence assigned (%) |
|---|---|---|---|---|
| **Domain** | | | | |
| Bacteria | 112,111 | 83.57 | 231.06 | 85.77 |
| Archaea | 17,591 | 13.11 | 30.77 | 11.42 |
| Not assigned | 3,576 | 2.67 | 5.41 | 2.01 |
| Eukaryota | 871 | 0.65 | 2.15 | 0.80 |
| **Phylum** | | | | |
| Not assigned | 68,695 | 51.21 | 107.01 | 39.72 |
| Proteobacteria | 16,430 | 12.25 | 57.09 | 21.19 |
| Cyanobacteria | 17,755 | 13.24 | 43.84 | 16.27 |
| Firmicutes | 17,967 | 13.39 | 35.03 | 13.01 |
| Euryarchaeota | 9,091 | 6.78 | 18.01 | 6.69 |
| Spirochaetes | 1,316 | 0.98 | 2.31 | 0.86 |
| Fusobacteria | 1,241 | 0.93 | 2.30 | 0.85 |
| Ascomycota | 255 | 0.19 | 1.09 | 0.40 |
| Arthropoda | 392 | 0.29 | 0.73 | 0.27 |
| Actinobacteria | 349 | 0.26 | 0.61 | 0.23 |
| Bacteroidetes | 185 | 0.14 | 0.34 | 0.13 |
| Chordata | 224 | 0.17 | 0.33 | 0.12 |
| Chlamydiae | 78 | 0.06 | 0.33 | 0.12 |
| Deinococcus-Thermus | 81 | 0.06 | 0.18 | 0.07 |
| Crenarchaeota | 90 | 0.07 | 0.16 | 0.06 |
| **Class** | | | | |
| Not assigned | 107,513 | 80.14 | 182.98 | 67.93 |
| Gammaproteobacteria | 6,084 | 4.54 | 25.09 | 9.31 |
| Betaproteobacteria | 2661 | 1.98 | 16.47 | 6.11 |
| Mollicutes | 9,340 | 6.96 | 15.58 | 5.78 |
| Cyanobacteria | 1741 | 1.3 | 11.16 | 4.14 |
| Methanomicrobia | 1,401 | 1.04 | 4.74 | 1.76 |
| Bacilli | 1374 | 1.02 | 3.96 | 1.47 |
| Deltaproteobacteria | 672 | 0.5 | 2.21 | 0.82 |
| Spirochaetes (class) | 1,108 | 0.83 | 1.88 | 0.70 |
| Alphaproteobacteria | 632 | 0.47 | 1.47 | 0.55 |
| Epsilonproteobacteria | 563 | 0.42 | 0.77 | 0.28 |
| Clostridia | 198 | 0.15 | 0.57 | 0.21 |
| Sordariomycetes | 99 | 0.07 | 0.49 | 0.18 |
| Thermoplasmata | 133 | 0.10 | 0.46 | 0.17 |
| Fusobacteria (class) | 137 | 0.10 | 0.35 | 0.13 |
| Chlamydiae (class) | 61 | 0.05 | 0.27 | 0.10 |
| Actinobacteria (class) | 124 | 0.09 | 0.26 | 0.10 |
| Insecta | 101 | 0.08 | 0.26 | 0.10 |
| Deinococci | 31 | 0.02 | 0.10 | 0.04 |
| **Order** | | | | |
| Not assigned | 126,987 | 94.66 | 226.53 | 84.09 |
| Burkholderiales | 851 | 0.63 | 13.06 | 4.85 |
| Prochlorales | 1,607 | 1.20 | 10.59 | 3.93 |
| Alteromonadales | 207 | 0.15 | 8.21 | 3.05 |
| Mycoplasmatales | 2,831 | 2.11 | 4.59 | 1.70 |
| Vibrionales | 206 | 0.15 | 2.01 | 0.74 |
| Lactobacillales | 638 | 0.48 | 1.91 | 0.71 |
| Chroococcales | 134 | 0.10 | 0.57 | 0.21 |
| Enterobacteriales | 86 | 0.06 | 0.37 | 0.14 |
| Spirochaetales | 164 | 0.12 | 0.34 | 0.13 |
| Campylobacterales | 253 | 0.19 | 0.33 | 0.12 |
| Pseudomonadales | 45 | 0.03 | 0.29 | 0.11 |
| Methanosarcinales | 31 | 0.02 | 0.21 | 0.08 |
| Thermoplasmatales | 66 | 0.05 | 0.17 | 0.06 |
| **Genus** | | | | |
| Not assigned | 131,913 | 98.33 | 238.43 | 88.51 |
| Burkholderia | 667 | 0.50 | 12.46 | 4.63 |
| Prochlorococcus | 1,235 | 0.92 | 8.54 | 3.17 |
| Shewanella | 207 | 0.15 | 8.21 | 3.05 |
| Vibrio | 49 | 0.04 | 1.15 | 0.43 |
| Streptococcus | 38 | 0.03 | 0.24 | 0.09 |
| Pseudomonas | 4 | 0.00 | 0.15 | 0.06 |
| Mycoplasma | 26 | 0.02 | 0.12 | 0.05 |
| **Dominant populations** | | | | |
| Not assigned | 132,703 | 98.92 | 246.86 | 91.64 |
| Burkholderia | 103 | 0.08 | 9.07 | 3.37 |
| Shewanella | 207 | 0.15 | 8.21 | 3.05 |
| Gammaproteobacteria | 673 | 0.5 | 3.31 | 1.23 |
| Prochlorococcus | 463 | 0.35 | 1.94 | 0.72 |

From top to bottom, for contigs ≥1 kb, the percentage of sequence assigned to clades at the ranks domain, phylum, class, order and genus is shown. All phylogenetic clades assigned ≥100 kb are shown.

exceeds 90% for fragments >3 kb, with a specificity of more than 95% across all fragment lengths. As a particularly difficult case, we investigated the accuracy of finding (G+C)-rich *Prochlorococcus* sp. fragments using a low-(G+C) model and vice versa (**Supplementary Table 7** online). This allowed the accurate retrieval of fragments of half of the organisms. Notably, fragments of the (G+C)-richest genome were retrieved with high accuracy by our method with the low-(G+C) model. We improved the model by including fragments of all genomes except the one being tested, which resulted in good accuracy for five of the six genomes. Fragments of the last outlier genome (*Prochlorococcus marinus* str. MIT 9313) could be accurately retrieved only by building a model based on fragments of all genomes. This experiment nicely demonstrates the advanced capabilities of our method as well as the difficulties one may encounter.

In the Sargasso Sea sample, PhyloPythia identified 463 contigs or 1.94 Mb of sequence for the dominant *Prochlorococcus* sp. population (*Prochlorococcus* SAR-1; **Table 3**); 87% of the high coverage contigs among these contain regions most similar to *Prochlorococcus* sp. genes in RefSeq, demonstrating the reliability of these assignments. In the original annotation, 370 kb are annotated for this population with 222 kb overlapping our assignments. Most of the additional 1.7 Mb of assigned sequence comes from low-coverage contigs that were not annotated in the original release.

For the dominant *Shewanella* and *Burkholderia* species populations, 8.2 Mb and 9.07 Mb of sequence, respectively, were retrieved. Both populations show high consistency in the assignments for the marker gene–characterized reference set, indicating that these assignments are largely correct. Of these, 6.9 Mb and 8.4 Mb, respectively, are in the original annotation for the high-coverage contigs. Based on assignment sensitivity and specificity for the reference set, we estimate that the sample in total contains ~9.27 Mb and ~8.3 Mb of sequence from the dominant *Burkholderia* and *Shewanella* species populations.

Most importantly, our analysis shows that the Sargasso Sea sample represents a treasure trove of information, with many phyla awaiting a more detailed investigation. The compilation of more population models, which can be built using as little as 100 kb of training sequence, should allow the fragment retrieval and population-based binning for more populations from this sample leading to new insights and understanding of the microbial populations inhabiting this marine environment.

## DISCUSSION

We demonstrated that sequence composition allows the accurate characterization of genomic fragments from the complete phylogenetic spectrum that has been sampled to date by genome projects. For several metagenome samples of considerable complexity (two phosphorus-removing sludge samples and the Sargasso Sea sample), our technique permitted a comprehensive phylogenetic characterization well beyond what has been possible before. Fragments were assigned to either well characterized higher-ranking clades or to clades that could be modeled for sample-populations based on marker gene–carrying contigs and similar means. For the dominant genera, several more megabases of sequence could be assigned. Additionally, large parts of the samples were characterized at higher taxonomic levels. Specific assignments were possible even for fragments as short as 1 kb, and in the presence of considerable noise from organisms of unknown clades.

The analysis of the Sargasso Sea sample demonstrated the gain in classification accuracy in comparison to the unsupervised clustering of fragments with sequences of known organisms. Owing to their capability for data exploration and hypothesis generation, unsupervised techniques represent a valuable complement to supervised methods such as the one we presented. The results for PhyloPythia and TETRA in binning of the dominant sample populations showed that both methods can be used to accurately identify most of the larger fragments of the dominant sample populations. PhyloPythia additionally excels in its ability to assign short fragments of the dominant sample populations and fragments of higher-level clades that are best described by more complex shapes in the feature space (see below).

What type of metagenome fragments can be reliably assigned is a complex function, depending on the complexity of the organismal population, the amount of available training data for the relevant clades, the parameter space of sequence composition, the fragment length and the type of available training data (organismal or only related higher-level clades). Our evaluation suggests that highly complex and mostly unassembled samples (such as soil samples) will remain largely uncharacterizable. Assignment of short fragments has been reported for organisms from mixtures of lower complexity, based on the availability of near-complete genomes of these organisms for training. Conceivably, similar approaches to ours that are optimized for fragments $<1$ kb might be able to characterize unassembled reads of unknown organisms for moderately complex samples. If the soil problem could be reduced to separate analyses of distinct partly enriched subpopulations, this would likely permit considerable progress in the subsequent *in silico* characterization of such samples. If the objective is to recover the genome of a particular low-abundance population, then some form of pre-enrichment, such as flow sorting[28] or single-cell amplification[29], will likely ensure a better sequencing cost benefit.

An advantage of the SVM technique is its ability to learn the relevant class-specific characteristics, even in a space where considerable variation is caused by other influences. For example, sequence composition is dominated by influences such as global G+C content and thermophily at the domain level[25], and is quite varied for the archaeal genomes[30]. Nonetheless, PhyloPythia can accurately discriminate among fragments of the different domains. One intriguing observation here is the higher complexity of the best feature space at domain level compared to lower phylogenetic ranks. Although the accuracy differences for hexamers and shorter patterns are not dramatic, they indicate the complex shape of the structure that is required for effective discrimination among the domains. Evolutionary relationships at this level cannot be described by simple, unifying patterns that summarize the variations of lower-ranking clades.

Phenomena such as horizontal gene transfer can complicate composition-based phylogenetic classification, as evidenced by some misplaced fragments of the *Thiothrix* population (see above). In contrast, such composition-based 'misplacement' can provide interesting indications of the potential origin[18,31]. For instance, our *in silico* evaluation at the domain level showed that fragments of the *Thermotoga maritima* genome, which is known to contain archaeal regions[32], were assigned to the Archaea (**Supplementary Fig. 4**).

Assignment accuracy increases with the number of clades common to both the classifier and the analyzed sample. Although our present knowledge of the phylogenetic space is far from complete, for higher-ranking clades there is already sufficient coverage to allow a partial characterization of the samples from most environments. At the phylum level, we were able to model 11 of the approximately 53 existing prokaryotic phyla[4]. The class-level model contains 18 prokaryotic clades. Organisms from several unexplored phyla are currently being sequenced and will soon allow the addition of new clades to composition-based models.

We believe that composition-based taxonomic classification can have an important role and complement comparative sequence analysis. Composition-based analysis evaluates global, clade-specific characteristics of sequence fragments. Comparative analysis can provide the *a priori* knowledge and initial data sets for composition-based classification, to allow the characterization of a large fraction of a sample at higher phylogenetic levels and to identify further sequences of specific sample populations. For best results, initial collections of training sequences can be compiled based on phylogenetic markers and similar means, analogous to our analyses. For very diverse communities, a viable strategy may be the isolation and sequencing of fosmid-sized fragments bearing these marker genes.

Additional advantages of composition-based classification include automation and speed: tens of thousands of fragments can be classified across all ranks with a few days of computation on a single processor. The advent of such techniques for the analysis of metagenome samples will likely increase our understanding of the uncultured microbial world.

## METHODS

**Compositional sequence patterns.** For compositional feature analysis, we map a given piece of DNA sequence $s$ to a higher-dimensional space of nucleotide patterns $o = \{o_1, o_2, \ldots, o_q\}$, where

$o$ is defined by the pattern length $w$ and the number of literals $l$[33]. In this space, $s$ is represented by the compositional input vector $v = (a_1, a_2, \ldots, a_q)$; where $a_i$ is the frequency of pattern $o_i$ in $s$. Input vectors are normalized by the total number of patterns for each sequence.

**Phylogenetic modeling.** Phylogenetic classification is a multiclass problem, where at any given rank an organism belongs to exactly one of all existing clades (that is, classes). Our overall classification framework includes multiclass classifiers trained with genomic fragments of different length at the rank of domain, phylum, class, order and genus as well as sample-specific classifiers. In the models, each adequately sampled clade at a given rank ($\geq 3$ distinct species of the clade have been sequenced) is represented by a class. A class 'other-unknown' is trained with sequences from all poorly sampled clades in our data set.

We implemented a phylogenetic framework across multiple taxonomic ranks. Every rank includes several multiclass SVM classifiers that are trained with fragments of a certain length. The input items for the SVM are the compositional vectors derived from DNA sequence fragments. Intrinsically, the SVM is a binary classifier. For multiclass classification, we apply the 'all-versus-all' technique, where $N(N-1)/2$ distinct binary classifiers are trained, one for each possible pair of classes. Fragments are assigned to a class with a voting mechanism. In a post-processing step, every assignment to a known clade is re-evaluated with a classifier trained to discriminate between fragments of this clade and all others (one-versus-all approach); at this point false positive assignments to known clades are frequently discarded.

We used more than 1 Gb of genomic sequence from 340 organisms[34] for training. The input data represent all 3 domains, 14 different phyla, 22 classes, 29 orders and 31 genera with 3 or more species (**Supplementary Fig. 6** online). We obtained the taxonomic information on the organisms and their phylogenetic relationships from the US National Center for Biotechnology Information (NCBI) Taxonomy database[35].

**Availability.** A web server with different models is available for the processing of smaller sequence sets online (http://cbcsrv.watson.ibm.com/phylopythia.html). Please contact the authors for the analysis of larger samples and creation of new models with other sample-specific clades.

**Additional methods.** Further information on the methods for pattern discovery, accuracy evalutation, construction of the meta-genome classifier used for the analysis of the Sargasso Sea sample is available in **Supplementary Methods** online.

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

**ACKNOWLEDGMENTS**

We thank N. Ivanova, V. Kunin and F. Warnecke for help with selection of CAP and *Thiothrix*-specific training sets and for validation analyses of the metagenomic data-set binning, L. Krause for providing the SEED data, T. Huynh for implementing the web interface, and S. Polonsky for comments and discussion. The work of H.G.M. and P.H. was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program; the University of California, Lawrence Livermore National Laboratory, under contract W-7405-Eng-48; Lawrence Berkeley National Laboratory under contract DE-AC03-76SF00098; and Los Alamos National Laboratory under contract W-7405-ENG-36. PhyloPythia's results were incorporated in the US Department of Energy Joint Genome Institute Integrated Microbial Genomes & Metagenomes (IMG/M) experimental system (http://www.jgi.doe.gov).

**AUTHOR CONTRIBUTIONS**

A.C.M. developed and evaluated the method, A.T. contributed codes for pattern discovery and discussion, P.H. and H.G.M. helped with discussions and the evaluation of the results for the EBPR sludges, A.C.M., I.R., H.G.M. and P.H. contributed to the writing of the manuscript, and A.C.M. and I.R. designed and planned the project.

**COMPETING INTERESTS STATEMENT**

The authors declare that they have no competing financial interests.

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

1. Venter, J.C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
2. Tringe, S.G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).
3. Tyson, G.W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
4. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, **3**, REVIEWS0003 (2002).
5. Woese, C.R. & Fox, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**, 5088–5090 (1977).
6. Woese, C.R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
7. Graham, D.E., Overbeek, R., Olsen, G.J. & Woese, C.R. An archaeal genomic signature. *Proc. Natl. Acad. Sci. USA* **97**, 3304–3308 (2000).
8. Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L. & Koonin, E.V. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1**, 8 (2001).
9. Ciccarelli, F.D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
10. Cole, J.R. *et al.* The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* **33**, D294–D296 (2005).
11. Garcìa Martin, H. *et al.* Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.* **24**, 1263–1269 (2006).
12. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glockner, F.O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6**, 938–947 (2004).
13. Gans, J., Wolinsky, M. & Dunbar, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**, 1387–1390 (2005).
14. Karlin, S. & Burge, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**, 283–290 (1995).
15. Karlin, S. & Mrazek, J. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* **94**, 10227–10232 (1997).
16. Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G. & Fertil, B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* **16**, 1391–1399 (1999).
17. Nakashima, H., Ota, M., Nishikawa, K. & Ooi, T. Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res.* **5**, 251–259 (1998).
18. Sandberg, R. *et al.* Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res.* **11**, 1404–1409 (2001).
19. Abe, T. *et al.* A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency. *Genome Inform. Ser. Workshop Genome Inform.* **13**, 12–20 (2002).
20. Pride, D.T., Meinersmann, R.J., Wassenaar, T.M. & Blaser, M.J. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* **13**, 145–158 (2003).
21. Chapus, C. *et al.* Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol. Biol.* **5**, 63 (2005).
22. Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S. & Ikemura, T. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res.* **12**, 281–290 (2005).
23. Edwards, R.A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, 57 (2006).
24. Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F. & Sockett, R.E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33**, 1141–1153 (2005).

25. Lynn, D.J., Singer, G.A. & Hickey, D.A. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* **30**, 4272–4277 (2002).
26. Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. & Koonin, E.V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**, 7 (2006).
27. DeLong, E.F. Microbial community genomics in the ocean. *Nat. Rev. Microbiol.* **3**, 459–469 (2005).
28. Kalyuzhnaya, M.G. *et al.* Fluorescence *in situ* hybridization-flow cytometry-cell sorting-based method for separation and enrichment of type I and type II methanotroph populations. *Appl. Environ. Microbiol.* **72**, 4293–4301 (2006).
29. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680–686 (2006).
30. Campbell, A., Mrazek, J. & Karlin, S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **96**, 9184–9189 (1999).
31. McHardy, A.C. *Gene finding and the evaluation of synonymous codon usage features in microbial genomes.*. Thesis, Bielefeld Univ., (2004).
32. Nelson, K.E. *et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* **399**, 323–329 (1999).
33. Tsirigos, A. & Rigoutsos, I. A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.* **33**, 922–933 (2005).
34. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
35. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **29**, 11–16 (2001).