Supplementary Methods

Compositional sequence patterns

For compositional feature analysis, a given piece of DNA sequence s is mapped to a higher-dimensional space of nucleotide patterns $\pi = \{\pi_1, \pi_2,..., \pi_q\}$, where π is defined by the pattern length w and the number of literals l[1]. In this space, s is represented by the compositional input vector $\phi = (a_1, a_2,..., a_q)$; where a_i is the frequency of pattern π_i in s. The method is a generalization of conventional compositional approaches and exhibits a number of desirable characteristics. First, nucleotide patterns of arbitrary lengths and densities can be computed, thus allowing the selection of the parameters with the most discriminatory power. Second, the method extends composition-based schemes in that it is able to 'ignore' certain nucleotide positions: this is achieved through the use of generating templates that include 'gaps' i.e. "wild-cards." Third, optionally the periodicity of the genetic code is taken into account: in particular, when collecting the instances of a pattern, the constraint can be imposed that a pattern be position-specific. The input vectors are subsequently normalized by the total number of patterns for each sequence.

Support Vector Machine (SVM)

The SVM[2, 3] is a maximum margin classifier that during training learns to optimally discriminate between the items of two classes. In the context of our work, the items are the compositional input vectors derived from DNA sequences, and the classes represent different phylogenetic clades. In the feature space, the algorithm implicitly learns a hyperplane which optimally separates the members of the two classes. Based on its position relative to the hyperplane an item is assigned to a class during the classification process. The confidence of the assignment is determined by the distance of the item from the hyperplane. The feature space can be different from the input space, which is determined by the employed kernel function. By use of a non-linear kernel such as the Gaussian kernel, a decision function can be learned that can accurately discriminate among items that are not linearly separable in the input space.

For multi-class classification, we apply the ‘all-vs.-all’ technique, where $N\cdot(N\text{-}1)/2$ distinct binary classifiers, one for each possible pair of classes, are used to assign a sequence fragment to a class[4]. The predicted class is the one that receives the most ‘votes’ from the internal classifiers, and is assigned randomly in the case of a tie. During a second classification step with a binary ‘One-vs.-all’ SVM classifier, these assignments are either corroborated or rejected. Rejection of false positive assignments of sequences that truly belong to an unknown clade occur frequently, as the model has been better trained to identify these using data from all organisms (except from those belonging to the clade of interest) instead of only those from poorly sampled clades. For our prototype, we used the multi-class SVM implementation of the LIBSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm).

SVM training

The compositional input vectors for the training of the SVM are created by mapping the input sequences s to the feature space π that is defined by the pattern span w and the number of literals l (in this context, a literal is a character from the DNA alphabet). The input vectors are normalized per row and scaled across columns in the range [0, 1]. Similar numbers of sequences are used for each class in model training. If it was not possible to include exactly the same number of sequences for each class, the misclassification cost C was scaled by the number of items, such that the overall misclassification cost for every class was the same. This was necessary, as, depending on the number of genomes in a given clade, it is not always possible to sample each genome equally to obtain the specified number of sequences for a class, or to generate the number of necessary fragments from each genomic sequence. A Gaussian kernel defined by the parameter γ is used with the SVM. The optimal values for C and γ are determined prior to the training in a grid search of the parameter space with 5-fold cross-validation on a subset of the training data. For the phylum, class, order and genus levels, 200 sequences per class are used for the grid search and 1000 for the training of the classifier. For the domain level, 1000 sequences per clade are used for the grid search and 3000 for training of the classifier.

Combined metagenome classifier

The *Phylopythia* framework includes several classifiers for each level that are built on fragments of different lengths. Based on results of the extensive evaluation, we decided to include classifiers trained on 5, 10, 15 and 50 kb fragments into the framework for the level of the phylum and class, and classifiers trained on 50 and 15 kb fragments for the lower levels, to maintain a high specificity level. Even though the inclusion of additional classifiers would lead to increased sensitivity, we opted for a setting with higher specificity. Beginning with the classifier that was trained on 50 kb fragments, a query sequence is tested with classifiers trained on successively shorter fragments, until an assignment is made, all classifiers have been applied, or a classifier is reached that has been trained with fragments shorter than the query. For the domain level, classifiers trained on 1, 3, 5, 10, 15, and 50 kb fragments are used, each for fragments with a similar size to the respective training fragments. For the classification of fragments from known organisms, 10 kb models are also included at the order and genus levels. For the metagenome sludge samples, additional 10, 15 and 50 kb (15 and 10 kb) models for the *Accumulibacter (Thiothrix)* genus are included in the framework, depending on the amount of available training data from the two sludge samples. Assignments at the different phylogenetic levels are checked for inconsistencies, which are resolved by choice of the lower level prediction.

Sequences

We used more than 1 billion bases worth of genomic sequence from 340 organisms with complete or nearly complete genome sequences. Genomic fragments were created by splitting the sequence of each organism into non-overlapping fragments of lengths 1, 3, 5, 10, 15 and 50 kb. Fragmented draft genomes (available as multiple contigs) were joined together in arbitrary order. For initial explorations of the best sequence source for our technique, a set of 1,028,017 reliable organism-specific genes was used. In this set, only genes with homologs either within the set or in RefSeq[5] were included, which also did not show the atypical sequence composition that is characteristic for certain types of laterally acquired sequence, which was determined using a previously described method[6].

Evaluation procedures

Each of the described experiments was evaluated by cross-validation with data that was withheld from the training procedures. To allow estimation of classification accuracy for genomic fragments of novel organisms, models were built using sequences from only some of the organisms, while others were withheld for evaluation. More specifically, the set of 340 organisms was split at random into 3 approximately equally-sized sets. Each of these sets in turn was set aside, while the other two were used to train the phylogenetic classifier. For nearly all of the available organisms, a model could be created that had not used any of the organism's sequences in training. To estimate accuracy for fragments of known organisms, for any given fragment length, a section of the genomic sequence was set aside for evaluation, while the rest was used to create genomic fragments for the training of the classifier. The models for this test were created with sequences from all 340 organisms and are also the ones used for the classification of the metagenome sequence samples.

For classification, composition vectors were derived from the original sequence fragments, which were normalized per row and scaled across columns in the range [0, 1]. For the evaluation with genomic fragments, tests were run with 100 genomic sequence fragments from every genome, if that many were available.

Measures of accuracy are the class-normalized sensitivity, or "micro-accuracy," $Sn = \left(\sum_{i=1}^{N} \frac{tp_i}{t_i} + \frac{tp_{other}}{t_{other}} \right) \cdot \frac{1}{N+1}$, and the specificity $Sp = \sum_{i=1}^{N} \frac{tp_i}{p_i} \cdot \frac{1}{N}$, where tp_i is the number of correctly assigned items to clade *i*, p_i is the total number of items assigned to clade *i*, and t_i is the number of items of clade *i*. The specificity is averaged over the *N* clades whereas the sensitivity is averaged over (*N* + 1) clades as the latter set includes the class 'Other'. The overall accuracy measures the success of assignment per item; the class-normalized sensitivity (micro-accuracy) measures the accuracy per class, which gives a performance measure that is not influenced by a skewed composition of a data set in terms of the contained classes. As metagenome samples are likely to differ in their class composition from the sequenced genomes, the class-normalized sensitivity gives a more generalized estimate of performance than the overall assignment accuracy on this test data set.

The Sargasso Sea sample

The metagenome sample of the Sargasso Sea[7] (Accession no. AACY00000000) was downloaded from Genbank. It comprises 811372 contigs (Accession no. AACY01000001 - AACY01811372); 2224 of these are part of an assembly of scaffolds of 3-fold coverage or more (CH004436-CH004736); 496417 are part of an assembly of low coverage scaffolds (CH004737-CH0236877); 312731 are individual reads or contigs not part of a scaffold.

463 contigs with small subunit rRNAs have been annotated for the sample; these were used to generate a reference for the evaluation. The rRNA genes were assigned to clades according to *Bergey's Manual of Systematic Bacteriology* with the 16S rRNA classifier of the ribosomal database project[8], using a ≥90% confidence threshold. To extend the reference set, contigs located on the same scaffold were added to the reference, yielding a total of 982 taxonomically assigned contigs, described with varying degrees of specificity by clades between, and including, the levels of domain and species.

For the four dominant sample populations of the high-coverage sequences, sample-specific *PhyloPythia* models (containing 5 classes – the dominant sample populations and the class 'Other') were generated using the marker-gene carrying contigs (*Shewanella*, *Burkholderia*, unidentified Gammaproteobacteria, and *Prochlorococcus*). In total, we used between 100 and 166 kb of training sequence to generate sequence fragments for each of the populations in model training. This procedure leaves many contigs of the dominant populations for evaluation (in addition to the training contigs, of which individual fragments – as opposed to the complete unit – were used for model creation). For *Prochlorococcus*, the available contigs were extended by the contigs located on a common scaffold, and several additional contigs identified based on blast homologies to known *Prochlorocci* genes. Similarly, for the unidentified Gammaproteobacteria, the training set was extended by contigs located on a common scaffold with the reference contigs. Training fragments for 3, 5, 10, 15 and 50 kb-trained *PhyloPythia* models were generated using a sliding window with a step size of 1/10th of the generated fragment size (e.g. 5 kb for 50 kb fragments). The class 'Other' was created from fragments of the 340 completed genomes, minus *Prochlorococcus*, *Shewanella* and

Burkholderia genomes. Sequence fragments were extended by their reverse complement for the sample-specific models, as only a small part of the organisms' genome was available for training (for the higher level models built with complete genomes, we found that this is not necessary, as presumably a shape can be learned in the feature space that accommodates both leading and lagging strand fragments for the organisms where this type of feature is apparent).

Comparison to available techniques

To compare *PhyloPythia* results with phylotype assignments generated with a Self-Organizing map (SOM)[9], the available results for the Sargasso sample (for the 1 kb fragments of all contigs ≥ 1 kb) were downloaded. The SOM is an unsupervised classification technique (i.e. a clustering technique), which was used by Abe *et al.* to identify Sargasso Sea sequences that cluster with the sequences of known organisms (37596 of 134149 contigs). The results describe the known organisms in terms of 25 higher-level phylotypes, which were manually chosen (3 order-level clades, 11 class- and 11 phylum-level clades by NCBI taxonomy). The data gives for each tested 1 kb fragment the number of fragments of each phylotype that co-localize with the tested fragment on the SOM. As each order and class level assignment indicates a phylum level assignment, we were able to perform a consistent comparison of our assignments with all the SOM assignments at the phylum level (note that this under- not overestimates the error rate, as false assignments at class level to a clade within the same phylum is thus considered correct).

We tried two different methods of inferring contig assignments to ensure generating the best overall assignment per contig from the available predictions for 1 kb fragments. Either a contig was assigned to a phylotype based on the highest overall count for all fragments of a contig, or we tried assigning individual 1 kb fragments based on counts and using majority vote to decide the overall contig assignment – ties were broken through random selection among the possibilities. The first approach had a higher accuracy on the marker-gene carrying reference contigs of the Sargasso sample (**Supplementary Table 5**) and was thus used subsequently for comparison with *PhyloPythia*.

For retrieval of fragments from the dominant sample populations, we compared *PhyloPythia* with TETRA[10], a method that provides the user with pairwise correlations coefficients of the input sequences relatedness in terms of their tetranucleotide usage. TETRA computes a quadratic result matrix for all-versus-all pairwise comparisons of all input fragments, which quickly becomes intractable for larger sets of sequences. We adapted the method so that query sequences could be compared to reference sequences of interest only, using as a reference in the analysis of the Sargasso sea sample the same sequences that were used to generate the *PhyloPythia* models of the dominant sample populations. This allows a direct comparison between *PhyloPythia* and TETRA in terms of the binning accuracy for sample-derived models for the dominant sample populations. For analysis with the TETRA method, the sequences for each population were concatenated and reverse complemented to generate one reference per sample population. As we slightly deviated from the original procedure, we performed a micro-evaluation on the Sargasso reference contigs, to ensure that we used the best cut-off setting for discrimination between significant assignments and noise (**Supplementary Table 5**).

1. Tsirigos, A. & Rigoutsos, I. A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res* **33**, 922-933 (2005).
2. Vapnik, V.N. The Nature of Statistical Learning Theory. (Springer, 1995).
3. Boser, B., Guyon, I. & Vapnik, V.N. in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. (ed. D. Haussler) 144--152 (ACM Press, 1992).
4. Hsu, C.-W., Lin, C-J A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* **13**, 415-425 (2002).
5. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**, D501-504 (2005).
6. Tsirigos, A. & Rigoutsos, I. A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res* **33**, 3699-3707 (2005).
7. Venter, J.C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74 (2004).
8. Cole, J.R. et al. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* **33**, D294-296 (2005).
9. 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).

10. 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).