Accurate phylogenetic classification of DNA fragments based on sequence composition

Alice C. McHardy[1], Héctor García Martín[2], Aristotelis Tsirigos[1], Philip Hugenholtz[2], and Isidore Rigoutsos[1†]

[1] Bioinformatics and Pattern Discovery Group, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

[2] US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA.

[†] To whom correspondence should be addressed.

Abstract

Metagenome studies have retrieved vast amounts of sequence out of a variety of environments, leading to novel discoveries and great insights into the uncultured microbial world. Except for very simple communities, diversity makes sequence assembly and analysis a very challenging problem. To understand the structure and function of microbial communities, a taxonomic characterization of the obtained sequence fragments is highly desirable, yet currently limited mostly to those sequences that contain phylogenetic marker genes. We show that for clades at the rank of domain down to genus, sequence composition allows the very accurate phylogenetic characterization of genomic sequence. We developed a composition-based classifier, PhyloPythia, for *de novo* phylogenetic sequence characterization and have trained it on a data set of 340 genomes. By extensive evaluation experiments we show that the method is accurate across all taxonomic ranks considered, even for sequences that originate from novel organisms and are as short as 1kb. Application to two metagenome datasets obtained from samples of phosphorus-removing sludge showed that the method allows the accurate classification at genus level of most sequence fragments from the dominant populations, while at the same time correctly characterizing even larger parts of the samples at higher taxonomic levels.

Introduction

The emerging field of metagenomics is dedicated to the study of sequences obtained directly by high-throughput sequencing of DNA samples from microbial communities. The approach has already delivered exciting insights into the lifestyle, evolution and characteristics of microbial organisms[3-5] that could not have been obtained otherwise, as the vast majority of microbes resist cultivation[6]. From a technical standpoint, the field has created novel computational challenges, such as a need for assembly and gene finding programs tailored for highly diverse sequence collections of organisms sampled with different abundances, and tools for the accurate phylogenetic characterization of the short sequences that are created in vast amounts.

One approach to classifying metagenomic sequence fragments is the use of 'marker genes', such as ribosomal RNAs, as phylogenetic anchors for identification of the source organism of a fragment. Ribosomal RNAs are highly conserved and allow the most accurate placement of an organism or sequence harboring the respective gene within the tree of life. The original observation and the framework for the quantification of evolutionary relationships resulted from the pioneering work of Carl Woese and colleagues[7, 8]. Even though the original marker gene collection has been expanding through inclusion of ubiquitous and slowly evolving or clade-specific proteins[9-11], the approach permits the characterization of only a limited number of fragments. For a very low complexity community found in acid mine drainage[5], an organism-specific 'binning' of fragments based on GC-content and read coverage retrieved near-complete genomes for the dominant species, which displayed significant differences in genomic GC-content. The use of tetranucleotide signatures has also shown promise in the characterization of low-complexity communities or the dominant organisms of more complex populations[12]. However, these schemes are not able to characterize the more diverse (and thus more challenging) metagenomes such as those from the very complex communities found in soil, which are estimated to contain millions of distinct taxa[13]. In the case of an extremely complex Minnesota soil community, a gene-centric characterization of the sample was undertaken, because less than 1% of reads could be assembled[4]. Gene-centric analyses, while useful for determining genes important for overall community function, in most

cases do not allow identification of the species harboring these genes. It is thus imperative that fast and accurate tools be developed that will allow the taxonomic characterization of short genomic sequence fragments and enable more comprehensive metagenome analyses.

In what follows, we present a method which uses clade-specific characteristics in sequence composition to phylogenetically characterize sequence fragments. Genomic sequence composition is well known to reflect organism-specific characteristics, which has been dubbed the 'genome signature'[14-19]. The phenomenon is sufficiently pronounced to allow the simultaneous supervised or unsupervised discrimination between several different species for even relatively short sequence fragments[12, 20-22]. Furthermore, these signatures also carry phylogenetic information, as recent studies on smaller data sets have shown[23, 24]. Based on draft or high-quality genomic sequences of 340 organisms from all domains of life, we constructed, optimized and extensively evaluated a composition-based phylogenetic classifier. For classification we use a multi-class Support Vector Machine classifier with the oligonucleotide composition of genome fragments as the input space. Due to the multitude of influences on sequence composition, the SVM is a good technique for this problem, as it is able to learn the relevant clade-specific characteristics of sequence composition also in spaces that are dominated by other influences. Across the complete sequence space considered, our method allows the accurate phylogenetic classification of genomic fragments. This is true for all taxonomic ranks considered (domain, phylum, class, order and genus), and, more importantly, for previously unseen fragments which originate from novel organisms. We named our new method for phylogenetic sequence characterization *PhyloPythia*[1].

We applied the method for the phylogenetic characterization of two metagenome samples of biological phosphorus removing sludge as used in the industrial processing of wastewater[25]. Our technique was able to automatically assign the phylogenetically characterized fragments of these samples (based on marker genes, population overlap between communities and scaffolding of contigs by read pair information) to the correct clades with high accuracy. Furthermore, additional genomic fragments could be assigned that could not be characterized by any other method, and larger parts of the sample could

[1] After the prophetess of the oracle of Delphi, who was commonly known as the Pythia.

be characterized at higher taxonomic levels in agreement with marker-gene based studies of sample composition.

Materials and 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[19, 26] (Supplementary Figure S10). In this space, s is represented by the compositional input vector $\phi(s) = (a_1, a_2, ..., a_q)$; where a_i is the frequency of pattern π_i in s. The method we propose is a generalization of conventional compositional approaches and exhibits several desirable properties. First, nucleotide patterns of arbitrary lengths and densities can be computed, which in turn allows to select the parameters with the most discriminatory power. Second, the method extends straightforward composition-based schemes in that is able to 'ignore' certain nucleotide positions: this is achieved through the use of generating templates that include 'gaps' and thus do not comprise continuous nucleotides. One such example template is *A.G,* which will match any of *AAG*, *ACG*, *AGG* or *ATG*, while ignoring the identity of the nucleotide that occupies the middle position. 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.

Multi-class classification

Phylogenetic classification is a multi-class problem, where at any given rank, such as e.g. the domain, an organism belongs to exactly one out of all existing clades. In the multi-class model, each adequately sampled clade for a particular rank is represented by an individual class. Adequately sampled here means a representation by at least 3 or more different species in the genomic data set. This allows us to estimate the classification accuracy for sequences from novel, unknown organisms of this clade by setting aside the sequence from at least one organism for testing, and using those of the other two for learning the clade-specific properties in the model (see Results). A class 'unknown' is

used to model all other existing clades, and is trained with the sequences of organisms from poorly sampled clades in our data set. To improve discrimination between the known and unknown clades, an additional classification step is needed. In this, every assignment is re-evaluated with a second classifier that has been specifically trained to discriminate between the sequences of a particular class and all others (One-versus-all approach).

We implemented this multi-class framework using Support Vector Machine classifiers. The Support Vector machine (SVM) is a high-performing machine learning technique that has been applied with improvements in classification accuracy to many biological problems and has a strong theoretical foundation[1, 2]. SVMs are intrinsically binary classifiers, but recent advances have extended their applicability to multi-class problems with considerable success. The SVM is a maximum margin classifier that during training learns to optimally discriminate between the items of two classes. In 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 implicitely learns a hyper-plane which optimally separates the items of the two classes. Based on its' position relative to this plane an item is assigned a class during the classification process. Hereby, the confidence of the assignment is determined by the distance of the item from the plane. The feature space can be different from the input space, which is determined by the utilized kernel function. By use of a non-linear kernel such as the Gaussian kernel, a decision function can be learned that can accurately discriminate between items that are not linearly separable in the input space.

For multi-class classification, we apply the 'all-vs.-all' technique, where $N \cdot (N-1)/2$ distinct binary classifiers, one for each possible pair of classes, are used to assign a piece of sequence. 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 the second classification step with a binary 'One-vs.-all' SVM classifier, these assignments are either confirmed 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 implementation, we

used the multi-class SVM algorithm of the LIBSVM package (`http://www.csie.ntu.edu.tw/~cjlin/libsvm`). See the supplementary material for more details on the complete procedure.

Materials

The genome sequences and annotation for 340 completely or nearly completely sequenced organisms was obtained from the SEED comparative genomics repository[27] (Supplementary Table S1). To create genomic sequence fragments, the sequence of each organism was split into non-overlapping fragments of lengths 1, 3, 5, 10, 15 and 50kb. Fragmented draft genomes were joined together in arbitrary order. For initial explorations of suitable sequence sources, also a set of 1.028.017 reliable organism-specific genes was used, which contains genes that have homologs either within this set or in RefSeq[28], and do not show the atypical sequence composition that is characteristic for certain types of laterally acquired sequence[26]. The taxonomic information for all organisms was obtained from the NCBI Taxonomy database[29].

Results

We used more than a gigabase of genomic sequence from 340 organisms (Supplementary Figure S1) to develop a composition-based phylogenetic classifier. In the set, all three domains of life, fourteen different phyla, 22 classes, 29 orders, and 31 genera are represented by 3 or more different species (Fig. 1). From a modeling point of view, the problem of phylogenetic classification can be broken down into distinct multi-class problems at the different taxonomic ranks: At any given rank, an organism belongs to exactly one out of all possible taxonomic classes. We modeled this using a multi-class SVM approach (see Methods).

The input space for the phylogenetic classifier is defined by the sequence composition vectors of the sequence fragments from different phylogenetic clades. From the training items the classifier learns in the feature space the clade-specific characteristics which allow their optimal discrimination. To allow discrimination between known (modeled) clades and others, a class ‘Other’ for all other, not modeled clades is included (see Methods).

The accuracy of phylogenetic classification was evaluated on withheld data in a blindfolded manner. In particular, we evaluated with the sequences of novel organisms from the perspective of the classifier, meaning that no sequence material of these organisms was included in the training data sets for model creation. This gives an estimate of the classification accuracy for yet undiscovered organisms, such as one is likely to predominantly encounter among the retrieved sequence fragments from an environment. To include the sequences of all 340 organisms in this procedure, the data set was randomly divided into three approximately equally sized sets of organisms. Each of these sets in turn was withheld for evaluation, while a phylogenetic classifier was trained with sequence fragments from the remaining organisms. To determine the classification accuracy for fragments from organisms for which already some genome sequence is known ('known' organisms), phylogenetic classifiers for the different taxonomic ranks were built with genomic fragments from all 340 genomes. Other fragments from these 340 genomes were then used to estimate the classification accuracy for organisms where some genomic sequence is already known. The models created with the sequence data from all 340 organisms were also applied for the characterization of the metagenome sequence samples.

Optimal oligonucleotide input space, sequence type, and other parameters

An extensive evaluation was performed to identify the optimal parameter settings, model architecture and sequence sources. As sequence sources, both genomic fragments and coding sequences were evaluated. Both types carry a strong phylogenetic signal in sequence composition across all taxonomic ranks. The clade-specific signal that is learnt from genomic fragments is not quite the same as the one learned from the genic regions, as the mutually decreasing performance of genomic fragments or genes with the other model demonstrates (Supplementary Figure S1). The direct assignment of genomic fragments eliminates the need to perform the intermediate step of gene identification. Additionally, the complete sequence, as opposed to only the coding parts thereof, can be used for classification.

To determine the oligonucleotide pattern space best suited for phylogenetic classification, an extensive search was performed for pattern lengths w of 2 – 6, allowing

for w-l = 0, 1, 2 unspecified positions in the composition template (Supplementary Table S2, Figure S2). The analysis showed that the lower ranking clades from the level genus to the class can be optimally discriminated based on literal nucleotide 5-mers of the genomic sequence fragments. For clades at the ranks of phylum and domain, more complex 6-mer patterns are necessary to optimally capture the characteristics of a joint ancestry.[2]

The SVM during training is guaranteed to learn a function that optimally separates the input items. By use of a Gaussian kernel, a classifier can be learnt that can separate items which are linearly inseparable in the input space of sequence composition. We found in extensive evaluation experiments that the Gaussian kernel outperforms the linear one in composition-based phylogenetic sequence assignment.

Fragment length and classification accuracy

Assembled metagenome sequence samples contain fragments of different sizes, which usually are at least 700bp or more in length. In a series of tests we evaluated the classification accuracy for fragments of different lengths. Because the classification accuracy of the SVM is influenced by the lengths of the fragments that are used for training, we evaluated this relation with 1kb, 3kb, 5kb, 10kb, 15kb and 50kb genomic fragments. For these tests, the sequences to be classified were obtained from organisms that are unknown to the classifier. For each length, a classifier was trained with approximately equal numbers of fragments for every clade (see 'Multi-class Support Vector machine training' and 'Evaluation procedures' in the Supplement) and evaluated with up to one hundred fragments of each organism, if that much sequence was available.

The analysis showed that from the genus to the phylum level, across all fragment lengths tested, the classifiers that were trained with longer fragments generally exhibit a higher specificity. The sensitivity of classification increases for all lengths with the use of shorter fragment-trained classifiers, but only significantly so for classifiers trained with similar-sized or longer fragments than the one being tested (Supplementary Tables S4-8,

[2] For CDSs, starting with oligomers of length 3 or longer, in-frame patterns are generally more informative than not in-frame patterns. The most informative patterns at the level of the genus are in-frame literal hexamers. At the higher levels, less specific hexamers with two non-literal positions are more informative (Table S3).

Fig S3-7). These convenient relations result from the use of the second, binary classifier, which more accurately rejects false positive assignments with the less noisy sequence composition vectors that are generated from longer sequence fragments.

We implemented these observations in a framework to achieve optimal classification accuracy for fragments of all lengths. In this framework a fragment to be classified is sequentially tested with classifiers that have been trained with decreasing length fragments, until the fragment is assigned, or a classifier trained on similar fragment lengths to the tested fragment is reached. If a fragment cannot be assigned to a known clade, it is assigned to the class 'Other' (see Supplement 'Combined metagenome classifier').

At the level of domain, the best classification accuracy is achieved using a classifier that was trained with fragments of similar size to the fragment being evaluated. The reason for this different behavior is that there is no broadly defined 'Other / Unknown' class to which items with unclear signal get assigned, and can then be retested with another model trained on shorter sequence fragments.

Accuracy of phylogenetic characterization with a composition-based classifier

In our evaluation, from the perspective of the classifier the tested fragments are always new, meaning that none of them was included in the training data sets. They can come from either 'Known' or 'Unknown' organisms. 'Known' means that some genomic sequence was already known and could be included in the training data for model construction. 'Unknown' means that no sequence of an organism was available for inclusion in the training data sets.

For novel organisms, 79-96% of all assignments are correct for all fragment lengths and across all taxonomic ranks evaluated (Figure 2B). The specificity is matched by a similarly high sensitivity for all fragments of length 5kb or longer (Figure 2A). Only for fragments shorter than 5kb we observe significant loss in sensitivity. It is important to stress that this high classification accuracy is achieved for sequences that originate from unknown organisms, and in a setting where a fragment can be assigned to any of up to 31 different known clades (for the genus level) and in the presence of considerable 'noise' of fragments from organisms of unknown clades, which are accurately identified as such in

most cases. Figure 3 shows the specificity of the assignments for the items of every clade that can be identified by the phylogenetic classifier.

The classification accuracy further increases for genomic fragments that originate from known organisms. For all fragments of lengths 3kb or more, both the sensitivity and specificity are 90-99% for clades from domain to order (Figure 2C/D, Supplementary Table S11). Even for fragments as short as 1kb, i.e. only slightly longer than a single read, 88.7-96.7% of all fragments are correctly assigned, whereby the sensitivity is 7.1-57.7%.

Genome sequence composition is shaped by many factors, including global genomic GC-content and the average temperature of the organisms' habitat as evidenced by multi-species comparisons of synonymous codon usage[30]. Nevertheless, our method learned the clade-specific characteristics in this space in a manner that allowed to accurately distinguish genomic fragments of the different domains; both thermophiles and non-thermophiles are correctly assigned in most cases (Supplementary Figure S8). We also investigated how translational selection, responsible for considerable intragenomic variation in the synonymous codon usage of some organisms[31], impacts on the classification accuracy. For this, the accuracy attainable for 3kb genomic fragments of unknown organisms was compared to that for 3kb fragments which carry ribosomal proteins. We found that a high accuracy, similar to the values obtained overall, was also achieved for the ribosomal protein-carrying fragments for clades from the genus to the phylum level, with 83-92% of all assignments being correct (Supplementary Table S10, Figure S9).

Composition-based classification of two metagenome samples from phosphorus removing sludge communities

PhyloPythia was applied for the characterization of two metagenome sequence samples of microbial communities responsible for the removal of biological phosphorus during industrial treatment of wastewater[25]. These microbial communities are characterized by considerable complexity in terms of the organisms they comprise.

Samples of lab-scale enhanced biological phosphorus removing (EBPR) sludge were obtained from two locations: one from Madison, Wisconsin (US) and the other from

Brisbane, Australia (OZ). From these, three sequence data sets were generated, each comprising 20-28 megabases of sequence; assemblies of the US and OZ data sets with the assembler PHRAP (USPHRAP and OZPHRAP), and a control assembly of the US data set with the JAZZ assembler (USJAZZ). The community of both samples is dominated by the uncultured bacterium, "*Candidatus* Accumulibacter phosphatis" (CAP). A 16S rRNA-based analysis revealed that CAP is the only species common to the two communities above the detection threshold, although overlap exists at higher phylogenetic levels[25].

Based on fragments that were already characterized by other means (harboring phylogenetic marker genes, read coverage, and subtraction binning, see below), this enabled us to assess the accuracy of classification for real metagenome samples of considerable complexity. Furthermore, the amount of characterized sequence for the samples could be substantially extended by classification of large parts of the uncharacterized sequences at higher taxonomic levels in agreement with marker-gene based study of sample composition, and by identification of additional fragments for the more abundant populations.

To extend the fraction of taxonomically characterized sequences, we (i) applied the multi-class phylogenetic classifiers of PhyloPythia that were described in the previous sections for the different ranks to these samples, and (ii) created additional classifiers for the lower ranks, using known sequences of the dominant organisms. In particular, we created a new multi-class model for the rank of order with an additional clade for the *Rhodocyclales,* based on genomic sequence fragments from *Dechloromonas aromatica* and known CAP sequences (identified by phylogenetic marker genes, read coverage, and overlap between the US and OZ sludge datasets) from both samples, as well as sample-specific models for CAP and a high abundance *Thiothrix*-like organism found only in the OZ sample, respectively. All three assemblies were characterized by PhyloPythia using these classifiers for different taxonomic ranks. See the section 'Combined metagenome classifier' of the supplementary material for details on the applied procedure.

As a first test of validity, the consistency (nesting) of assignments for the fragments at the different taxonomic ranks was checked (Table 1). The predictions are very consistent: 93-97.7% of all assignments (99.6 - 99.8% for high confidence

assignments) nest across all ranks (Table 1). As a second test, we analyzed the assignments for the 54 rRNA containing fragments of the samples and found all assignments for fragments >2kb to be correct, whereas 14 % (5% high confidence) were falsely assigned of the fragments shorter than 2kb.

Characterization at phylum level

Figure 4 shows the taxonomic assignments from the rank of domain to genus for the three assemblies. Culture-independent analyses of the US and OZ samples based on 16S rRNA indicated that both samples are dominated by CAP and other species belonging to the phylum *Proteobacteria*, flanked by much lower abundance species belonging to the *Bacteroidetes*, and for the OZ sludge only, representatives of the *Firmicutes*, *Verrucomicrobia* and *Chlorobi* phyla[25]. The assignments of PhyloPythia correlated well with this observed community structure (Table 1), with the majority of fragments of both samples being assigned to the *Proteobacteria* (63-74.8%). Interestingly, 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. A certain number of fragments were also assigned to phyla not found in 16S rRNA analysis, such as *Spirochaetes* and *Euryarchaeaota,* providing testable hypotheses about community structure.

Characterization at order level

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 the US and OZ sludge, in agreement with the marker-gene based studies (Table 1).

Characterization at genus level: Accumulibacter phosphatis

The relative percentage of fragments assigned to *Accumulibacter* qualitatively agrees with the rRNA-derived estimates of the relative abundance of these organisms in the sample, similarly for the *Thiothrix*-like organism present in the OZ sample. As a third test, we were able to leverage the fact that the two sludge communities share CAP to provide an estimate of the *Accumulibacter* binning accuracy of PhyloPythia. To this end,

CAP-specific classifiers were constructed from the known CAP genome fragments of one sample, applied for the identification of CAP fragments in the other sample, and the success of recovering known fragments was evaluated. Here, Phylopythia mainly missed very short fragments and successfully recovered 95-100% of the known fragments for all assemblies (Table 1). Of these, 74-97% could be assigned with high confidence (p-value ≥ 0.85).

Characterization at genus level: Thiothrix

16S rRNA phylogenetic markers indicated a *Thiothrix*-like species to be relatively well represented (13.8% of reads in OZ phrap contigs containing 16S rRNA genes) in the OZ data set. The genus *Thiothrix* belongs to the *Gammaproteobacteria*, and so far no genomic sequences besides ribosomal rRNA-carrying fragments are available. Starting with a training set of 17 characterized fragments (or 0.7Mbp of sequence), PhyloPythia was able to retrieve an additional 3.7Mbp of sequence for this organism. We were able to verify these assignments by using the scaffolding information provided by read pairs, i.e. contiguous sequence fragments (contigs) can be linked in some cases by end reads of the same cloned insert. Fig. 5 shows a series of contigs independently classified by PhyloPythia with high confidence, and linked together by read pair information. As can be seen from this Figure, 97% of the assignments are consistent. The remaining 3% of the fragments were either misclassified or not assigned at all, and are mainly small contigs with less composition signal or larger contigs that contain laterally transferred genes with atypical sequence composition (Figure 5). The majority of contigs (62%) in these scaffolds are classified as *Thiothrix* with high confidence (green), or are assigned with high confidence to consistent higher taxonomic levels (35%, light green, yellow). Based on the number of distinct tRNA synthetases identified in the *Thiothrix* contig set (Fig. 5 shown as stars), we estimate that 72% of the *Thiothrix* genome has been recovered. Therefore, we estimate the size of this genome to be ~6 Mbp. Estimates of individual population genome size and coverage within metagenomic datasets is useful for inferring the reliability of metabolic reconstruction and for guiding additional sequencing efforts.

Summary

Our extensive evaluation shows that the phylogenetic characterization of metagenome sequence samples with our composition-based technique is accurate. Without the use of prior knowledge, approximately 90% of the fragments from two samples of phosphorus removing sludge could be assigned at the domain level, 70-81.5% at the level of phylum, and 61.9-71.1% of the fragments could be assigned to known clades at class level. Based on the sample-specific models, known fragments of CAP could be retrieved with high accuracy, and 3.7Mbp of additional genomic sequence could be retrieved for *Thiothrix*, which is a deep-branching *Gammaproteobacterium*, for which no sequence has been available so far.

Conclusions

The presented results show that sequence composition allows the accurate characterization of genomic fragments from the complete phylogenetic spectrum that has been sampled by genome projects. Our composition-based technique allows a comprehensive phylogenetic characterization of complex metagenome samples, well beyond what has been possible to date, by accurately assigning short fragments to either well characterized higher-ranking clades or to sample-specific clades that can be modeled by data retrieved from the sample with other means.

Application of our method for the characterization of two metagenome samples from phosphorus removing sludge allowed the retrieval of several additional megabases of sequence from the dominating genera, and characterization of even larger parts of the samples at higher taxonomic levels. Additionally, the extensive *in silico* evaluation with 340 organisms showed the high accuracy of assignments, where high specificity was attained even for fragments as short as one 1kb, and in the presence of considerable noise from organisms of unknown clades. We want to stress that these values relate to the multi-class problem of *de novo* phylogenetic characterization, which is a more difficult test than a binary problem such as the discrimination between fragments from one particular clade and others.

An advantage of the applied SVM technique compared to mean-based classifiers is that it is able to learn the relevant class-specific characteristics, even in a space where considerable variation is caused by other influences. For instance, for the domain level,

previous studies showed that sequence composition is dominated by other influences such as global genomic GC-content and thermophily[30], and is quite heterogeneous for the archaeal genomes[32]. Despite this, our phylogenetic classifier achieved high separation accuracy in the placement of genomic fragments at the domain level. An intriguing observation in this context is the higher complexity of the feature space allowing the optimal separation of genomic fragments, compared to the lower phylogenetic levels. Although the accuracy differences for the optimal hexamers patterns and shorter patterns are not dramatic, they indicate the complex shape of the structure that is needed for discrimination between the different clades at the domain level, and that evolutionary relationships at this level cannot be described by simple unifying patterns which summarize the variations of lower-ranking clades.

The number of assignable fragments and the accuracy of these assignments generally increases with the number of clades that are represented both in a sequence sample and the phylogenetic classifier. Although our current 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, 14 clades could be modeled, 11 of which belong to the approximately 53 existing prokaryotic phyla[6]. At the class level the model contains 18 prokaryotic clades. Several organisms from the currently unexplored phyla are the focus of ongoing genome projects (http://www.genomesonline.org/), which will allow the addition of new clades to composition-based models soon.

Below the rank of class, the number of existing prokaryotic clades becomes large and the fraction of representatively samples clades comparably small. At this point, it becomes important to have prior knowledge about the phylogenetic characteristics and optimally also some labeled sequence material for the creation of classifiers which include the most relevant clades of a sample. In a manner analogous to our study of the metagenome sludge communities, initial collections of training sequences can be compiled based on phylogenetic markers and similar means. In the case of very diverse communities where this will deliver mainly short fragments which are too short for the construction of a reliable classifier, a viable strategy might be the isolation and

sequencing of fosmid-sized fragments bearing these marker genes, to increase the amount of initially available sequence.

We believe that composition-based classification can play a very important role and complement traditional comparative analysis. Composition-based analysis can evaluate global, clade-specific characteristics of genome sequence composition and is automatable. 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 an environmental sample at higher phylogenetic levels and to identify the sequences of specific organisms. An additional advantage of composition-based classification is its' speed: once the time-consuming step of classifier building has been completed, tens of thousands of fragments can be classified across all ranks within a few days of computation on a single processor. The advent of such computational techniques for the analysis of metagenome sequence samples will allow us to shed more light on and to increase our understanding of the uncultured microbial world.

Acknowledgements

We thank S. Polonsky for comments and discussion, and Natalia Ivanova, Victor Kunin and Falk Warnecke for help with selection of CAP and *Thiothrix*-specific training sets and for validation analyses of the metagenomic dataset binning. The generation of the metagenomic datasets and subsequent validation of the PhyloPythia binning was performed under the auspices of the DOE's Office of Science, Biological and Environmental Research Program; the University of California, Lawrence Livermore National Laboratory, under contract no. W-7405-Eng-48; Lawrence Berkeley National Laboratory under contract no. DE-AC03-76SF00098; and Los Alamos National Laboratory under contract no. W-7405-ENG-36. The assembled metagenomic data with PhyloPythia binning has been incorporated into the U.S. Department of Energy Joint Genome Institute Integrated Microbial Genomes & Metagenomes (IMG/M) experimental system (www.jgi.doe.gov/).

References

1. Vapnik, V.N. The Nature of Statistical Learning Theory. (Springer, 1995).
2. 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).
3. Venter, J.C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74 (2004).
4. Tringe, S.G. et al. Comparative metagenomics of microbial communities. *Science* **308**, 554-557 (2005).
5. Tyson, G.W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).
6. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3**, REVIEWS0003 (2002).
7. Woese, C.R. & Fox, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74**, 5088-5090 (1977).
8. Woese, C.R. Bacterial evolution. *Microbiol Rev* **51**, 221-271 (1987).
9. Graham, D.E., Overbeek, R., Olsen, G.J. & Woese, C.R. An archaeal genomic signature. *Proc Natl Acad Sci U S A* **97**, 3304-3308 (2000).
10. Griffiths, E., Ventresca, M.S. & Gupta, R.S. BLAST screening of chlamydial genomes to identify signature proteins that are unique for the Chlamydiales, Chlamydiaceae, Chlamydophila and Chlamydia groups of species. *BMC Genomics* **7**, 14 (2006).
11. Griffiths, E. & Gupta, R.S. Molecular signatures in protein sequences that are characteristics of the phylum Aquificae. *Int J Syst Evol Microbiol* **56**, 99-107 (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. 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).
15. Karlin, S. & Burge, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**, 283-290 (1995).
16. Karlin, S., Ladunga, I. & Blaisdell, B.E. Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci U S A* **91**, 12837-12841 (1994).
17. Karlin, S. & Mrazek, J. Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci U S A* **94**, 10227-10232 (1997).
18. Karlin, S., Mrazek, J. & Campbell, A.M. Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**, 3899-3913 (1997).
19. Tsirigos, A. & Rigoutsos, I. A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res* **33**, 922-933 (2005).
20. 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).
21. Sandberg, R. et al. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res* **11**, 1404-1409 (2001).
22. 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).
23. Chapus, C. et al. Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol Biol* **5**, 63 (2005).
24. Doolittle, W.F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129 (1999).
25. Garcia Martín, H. et al. Metagenomic analysis of phosphorus removing sludge communities. (in preparation).
26. 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).

27. 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).
28. 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).
29. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **29**, 11-16 (2001).
30. 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).
31. 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).
32. Campbell, A., Mrazek, J. & Karlin, S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci U S A* **96**, 9184-9189 (1999).

Figures and legends

Fig 1. Clades at different depths of the phylogenetic tree that are sufficiently represented by genomes of the 340 organisms for composition-based modeling.

Fig 2. Accuracy of phylogenetic characterization for differently sized genomic fragments from 340 organisms. The plots show the sensitivity (A/C) and specificity (B/D) of phylogenetic assignments by PhyloPythia. The legend gives in brackets the number of modeled clades for the phylogenetic classifiers. On the left side (A and B), the classification accuracy achievable for fragments from novel organisms (of which no genomic sequence was included in the training sets for model creation is shown). The right side (C and D) shows the achievable accuracy for organisms of which some genomic sequence is already known (other genomic fragments than the ones that were tested were included in the training data sets for the classifier). For D, the value for the

50kb fragments is undetermined, as evaluation of the specificity at the genus level is somewhat restricted by the fact that with the exception of organisms belonging to the 'Other' class, for most clades there is very little sequence that can be used for evaluation for fragments longer than 3kb, as most of it was included in the training data sets.

Fig 3. Phylogenetic classification accuracy by clade for differently sized fragments of 340 unknown organisms (for which no sequence fragments were included in the training data for the classifiers). From top to bottom, the clade-specific specificity, overall sensitivity (*Sn.*) and specificity (*Sp.*) is shown for clades at the rank of domain, phylum, class, order, and genus.

Fig 4. Dominant clades predicted for two metagenome samples of phosphorus removing sludge. USJAZZ and USPHRAP 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 $\geq 50kb$ are shown.

Fig 5. Binning accuracy of *Thiothrix* 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 (see legend). 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 (e.g. prophage) or non-coding repeat sequences (e.g. 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 13/18 tRNA synthetase types on high confidence *Thiothrix* contigs. Only scaffolds longer than 12kb or containing a tRNA synthetase are shown.

Table 1. Phylogenetic characterization of two metagenome samples from phosphorus removing sludge[25]. USJAZZ and USPHRAP 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 $\geq 50kb$ are shown. CAP = Candidatus *A. phosphatis*. From top to bottom, the percentage of sequence assigned to clades at the ranks domain, phylum, class, order and genus is shown.

Sample	US JAZZ		US PHRAP		OZ PHRAP	
# Fragments	5426		16370		11632	
Sequence (Mbp)	20.6		28.7		26.9	
Assigned (high conf.)	3444 (2905)		12782 (11459)		9908 (9062)	
Consistency (%)	97.7 (99.8)		96.7 (99.6)		93 (99)	
# known CAP fragments	7		665		584	
Sn. CAP (% Fragments)	100 (85.7)		86.2 (49.2)		79.8 (52.7)	
*Sn.*CAP (% kbp)	100 (97.1)		97.5 (80.1)		94.9 (73.6)	
CLADE	kb	%	kb	%	kb	%
Bacteria	18916	91.8	25450	89	24537	91.2
Unknown	1635	7.9	3100	11	1576	5.9
Archaea	38	0.2	106	0.4	691	2.6
Eukaryota	9	0	75	0.3	85	0.3
Proteobacteria	15406	74.8	17965	63	17797	66.2
Unknown / Other	3814	18.5	9086	32	6283	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	11129	54	13085	46	9150	34
Unknown / Other	5827	28.3	13253	46	10254	38.1
Gammaproteobacteria	2576	12.5	1229	4.3	5190	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	9948	48.3	11020	38	7507	27.9
Unknown / Other	9233	44.8	17351	60	14427	53.6
Xanthomonadales	1218	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
Unknown / Other	10738	52.1	18053	63	15771	58.6
Accumulibacter	9861	47.9	10680	37	6801	25.3
Thiothrix					4318	16.1

Fig 1: Clades at different depths of the phylogenetic tree that are sufficiently represented among the genomes of 340 organisms for composition-based modeling.

Archaea (23)

Crenarchaeota (4)	*Thermoprotei (4)*		
Euryarchaeota (18)	*Thermoplasmata (4)*	*Thermoplasmatales (4)*	*Thermococci (3)*
		Thermococcales (3)	*Pyrococcus (3)*
	Methanomicrobia (4)	*Methanosarcinales (4)*	*Methanosarcina (3)*

Bacteria (291)

Proteobacteria (147)	*Gammaproteobacteria (76)*	*Vibrionales (5)*	*Vibrio (4)*
		Xanthomonadales (7)	*Xylella (4)*
		Pasteurellales (12)	*Haemophilus (5)*
		Pseudomonadales (10)	*Pseudomonas (7)*
		Enterobacteriales (37)	*Buchnera (3)*
			Escherichia (7)
			Salmonella (8)
			Shigella (4)
			Yersinia (7)
	Alphaproteobacteria (28)	*Rhodospirillales (3)*	
		Rhodobacterales (3)	
		Rhizobiales (11)	*Brucella (3)*
		Rickettsiales (9)	*Rickettsia (5)*
			Wolbachia (3)
	Betaproteobacteria (24)	*Neisseriales (6)*	*Neisseria (5)*
		Burkholderiales (15)	*Burkholderia (7)*
			Bordetella (4)
	Deltaproteobacteria (8)	*Desulfuromonadales (3)*	
	Epsilonproteobacteria (10)	*Campylobacterales (10)*	*Campylobacter (5)*
			Helicobacter (4)
Cyanobacteria (14)	*n. a. (13)*	*Chroococcales (5)*	
		Nostocales (3)	
		Prochlorales (4)	*Prochlorococcus (4)*
Deinococcus-Thermus (3)	*Deinococci (3)*		
Actinobacteria (25)	*Actinobacteria (class) (25)*	*Actinomycetales (21)*	*Corynebacterium (3)*
			Mycobacterium (8)
			Streptomyces (3)
Spirochaetes (6)	*Spirochaetes (class) (6)*	*Spirochaetales (6)*	
Chlamydiae (9)	*Chlamydiae (class) (9)*	*Chlamydiales (9)*	*Chlamydophila (6)*
Fusobacteria (3)	*Fusobacteria (class) (3)*	*Fusobacterales (3)*	*Fusobacterium (3)*
Bacteroidetes (5)	*Bacteroides (class) (4)*	*Bacteroidales (4)*	*Bacteroides (3)*
Firmicutes (72)	*Clostridia (9)*	*Clostridiales (7)*	*Clostridium (6)*
	Mollicutes (9)	*Mycoplasmatales (6)*	*Mycoplasma (5)*
	Bacilli (54)	*Bacillales (26)*	*Staphylococcus (7)*
			Bacillus (10)
			Listeria (5)
		Lactobacillales (28)	*Streptococcus (17)*
			Lactobacillus (4)

Eukaryota (26)

Ascomycota (6)	*Sordariomycetes (3)*		
Arthropoda (4)	*Insecta (4)*	*Diptera (3)*	
Chordata (9)	*Mammalia (5)*		
	Actinopterygii (3)		

Figure 2

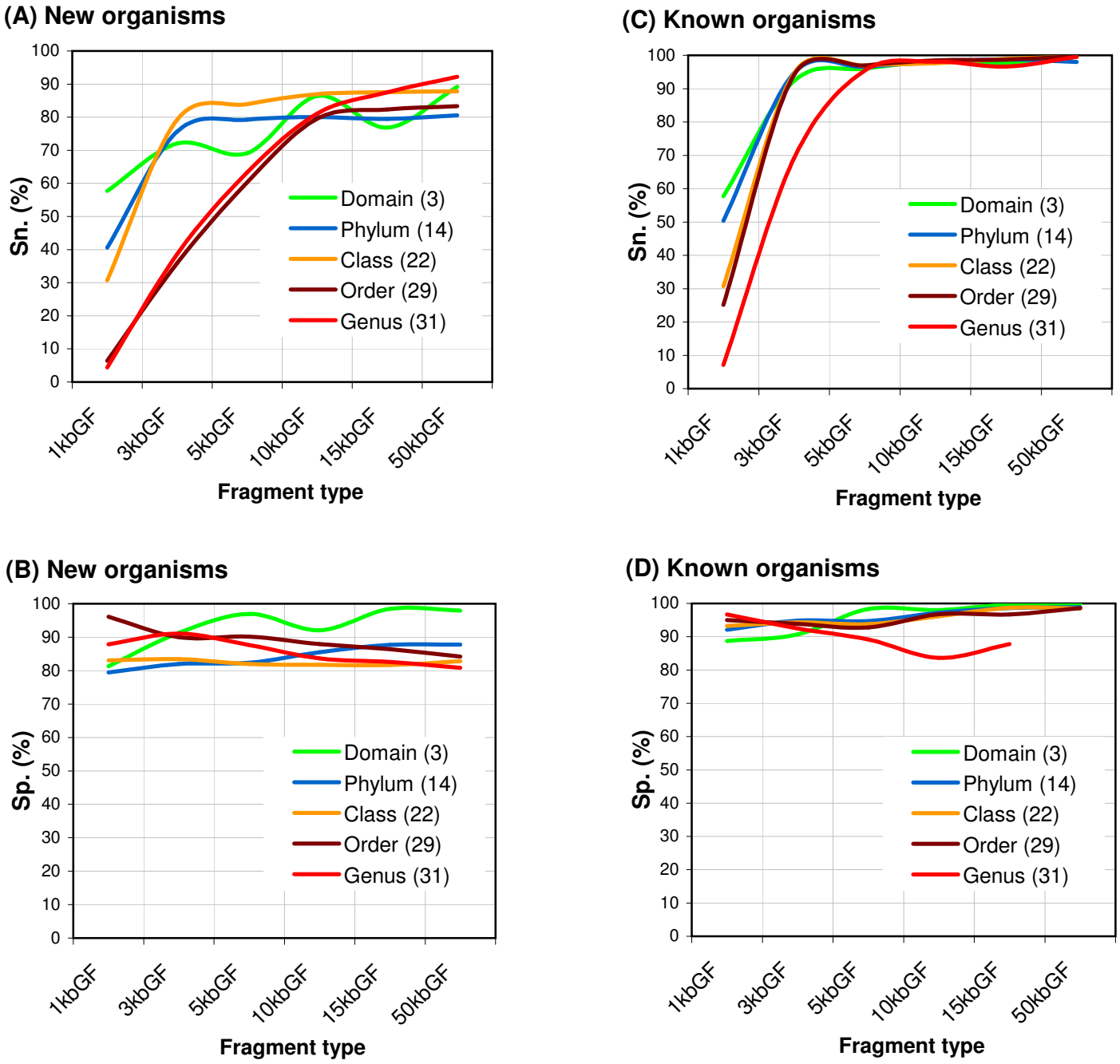


Fig 2. Accuracy of phylogenetic characterization for differently sized genomic fragments from 340 organisms. The plots show the sensitivity (A/C) and specificity (B/D) of phylogenetic assignments by PhyloPythia. The legend gives in brackets the number of modeled clades for the phylogenetic classifiers. On the left side (A and B), the classification accuracy achievable for fragments from novel organisms is shown. The right side (C and D) shows the classification accuracy for fragments from organisms for which some genomic sequence is known (some genomic fragments of the organism could be included in the training sets for the classifier).

Figure 3

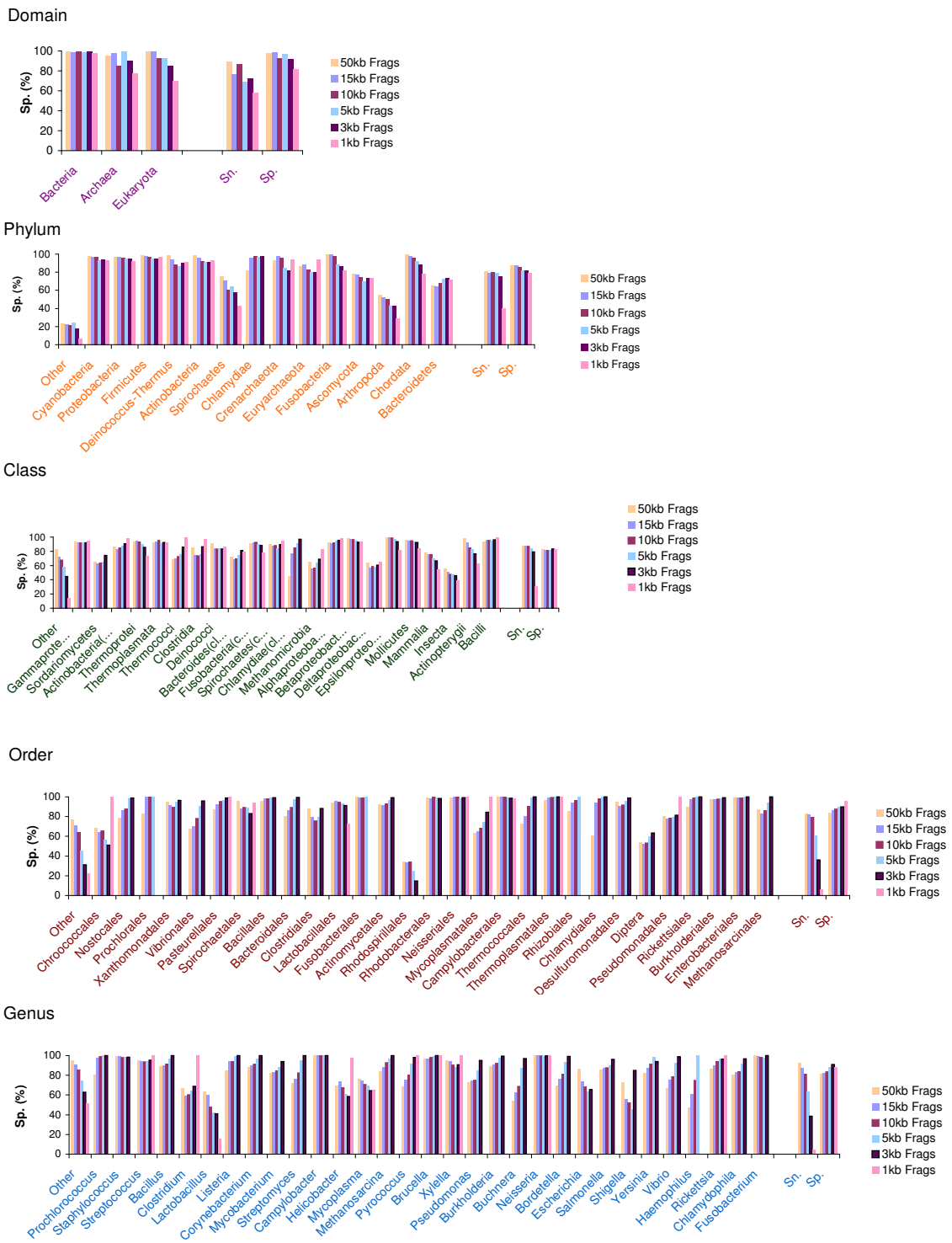


Fig 3. Phylogenetic classification accuracy by clade for differently sized fragments of 340 unknown organisms (for which no sequence fragments were included in the training data for the classifiers). From top to bottom, the clade-specific specificity, overall sensitivity (Sn.) and specificity (Sp.) is shown for clades at the rank of domain, phylum, class, order, and genus.

Figure 4

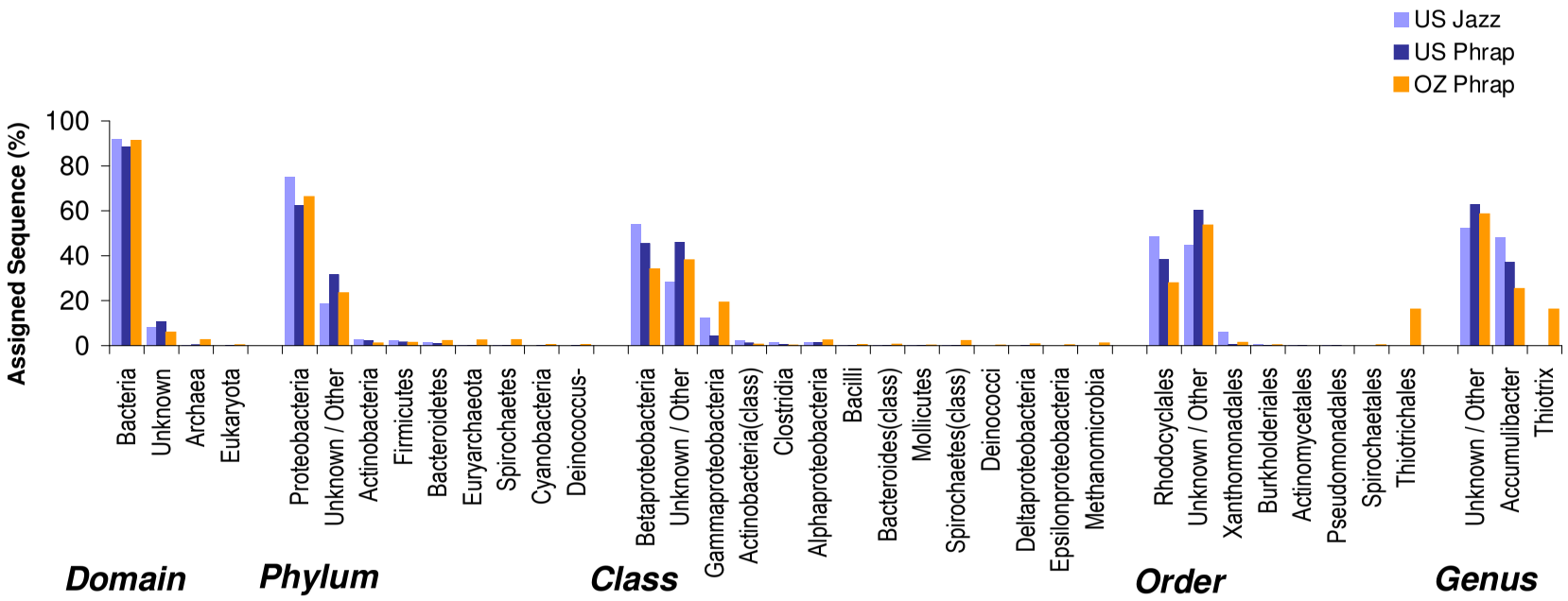
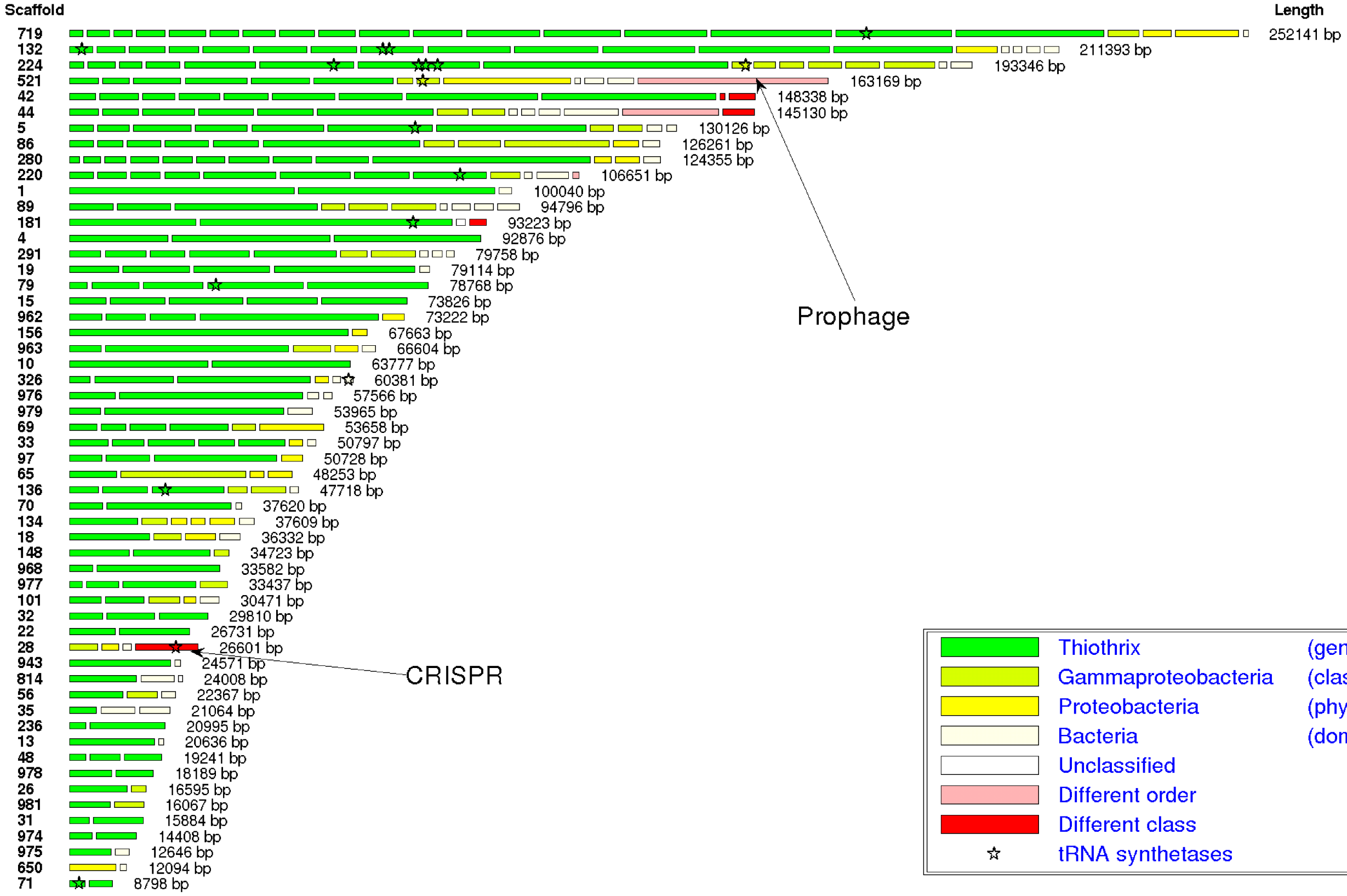


Fig 4. Dominant clades predicted for two metagenome samples of phosphorus removing sludge. USJAZZ and USPHRAP are assemblies with the JAZZ and PHRAP assembler of the US data set, respectively. OZPHRAP is the PHRAP-assembled OZ data set. All phylogenetic clades assigned ≥ 50kb are shown.

Bacteria; Proteobacteria; Gammaproteobacteria; Thiotrichales; Thiotrichaceae



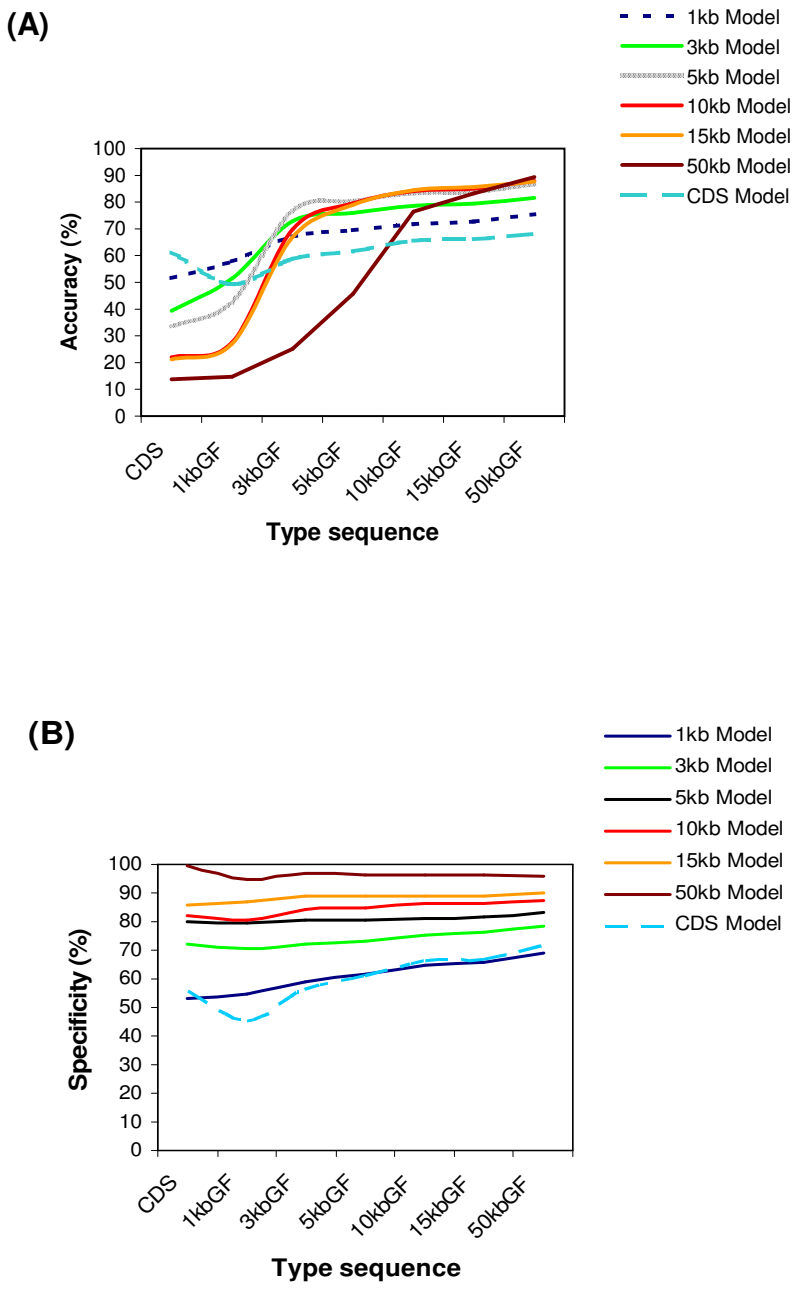
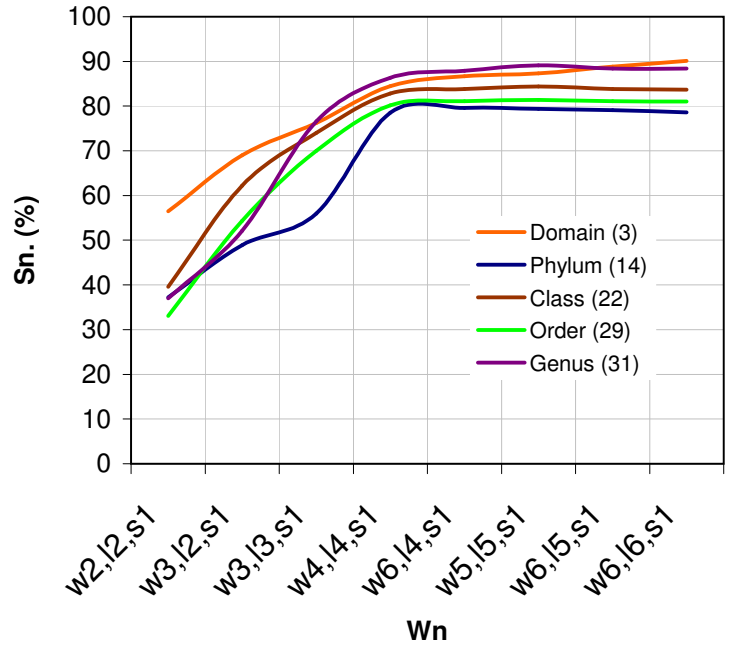


Fig. S1. Assignment accuracy for differently sized genomic fragments from unknown organisms at the level of the class (including 22 phylogenetic clades). 7 class-level classifiers were trained with tetranucleotide patterns (w4, l4) derived from genomic fragments of length 1, 3, 5, 10, 15 and 50kb as well as protein encoding sequences (CDS). For each classifier, the sensitivity (percentage of data classified correctly) (A) and specificity of assignments (B) with fragments of different lengths is shown for organisms that are unknown to the model (of which no sequences were used in training).



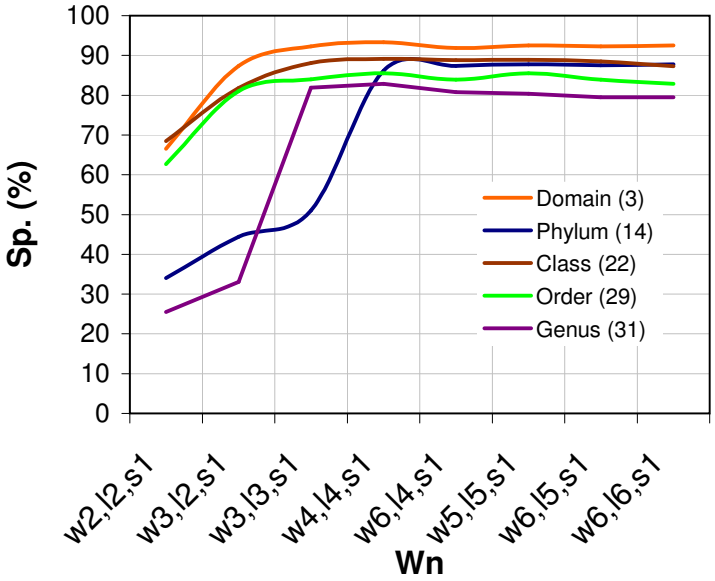


Fig. S2. Wn parameter search for the sequence composition space with the highest classification accuracy for genomic sequence fragments from unknown organisms (of which no sequence fragments were included in the model training data sets) at different phylogenetic levels. The legend gives in brackets the number of clades for each of the analyzed levels. The composition space is defined by the word length w, the number of literal characters l, and the step size s. Plot (A) shows the sensitivity and plot (B) the specificity, that is attainable in a given space with 15kb genomic fragments of organisms unknown to the classifier.

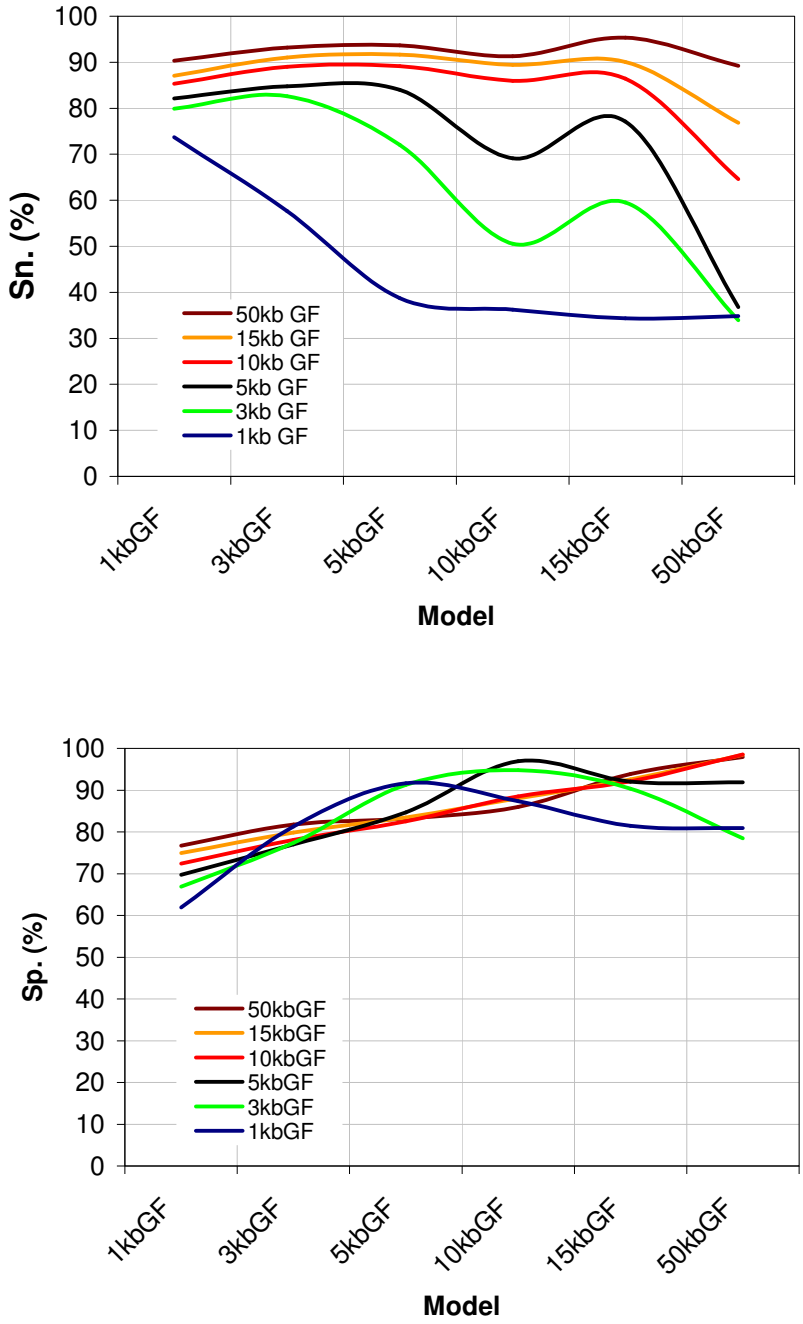


Fig. S3. Evaluation of the relation of genomic fragment length used for model creation and classification accuracy for genomic fragments of unknown organisms and different lengths. For the level domain, the sensitivity (Sn.) and specificity (Sp.) of classification is displayed for genomic fragments of different lengths with classifiers trained on fragments of different lengths.

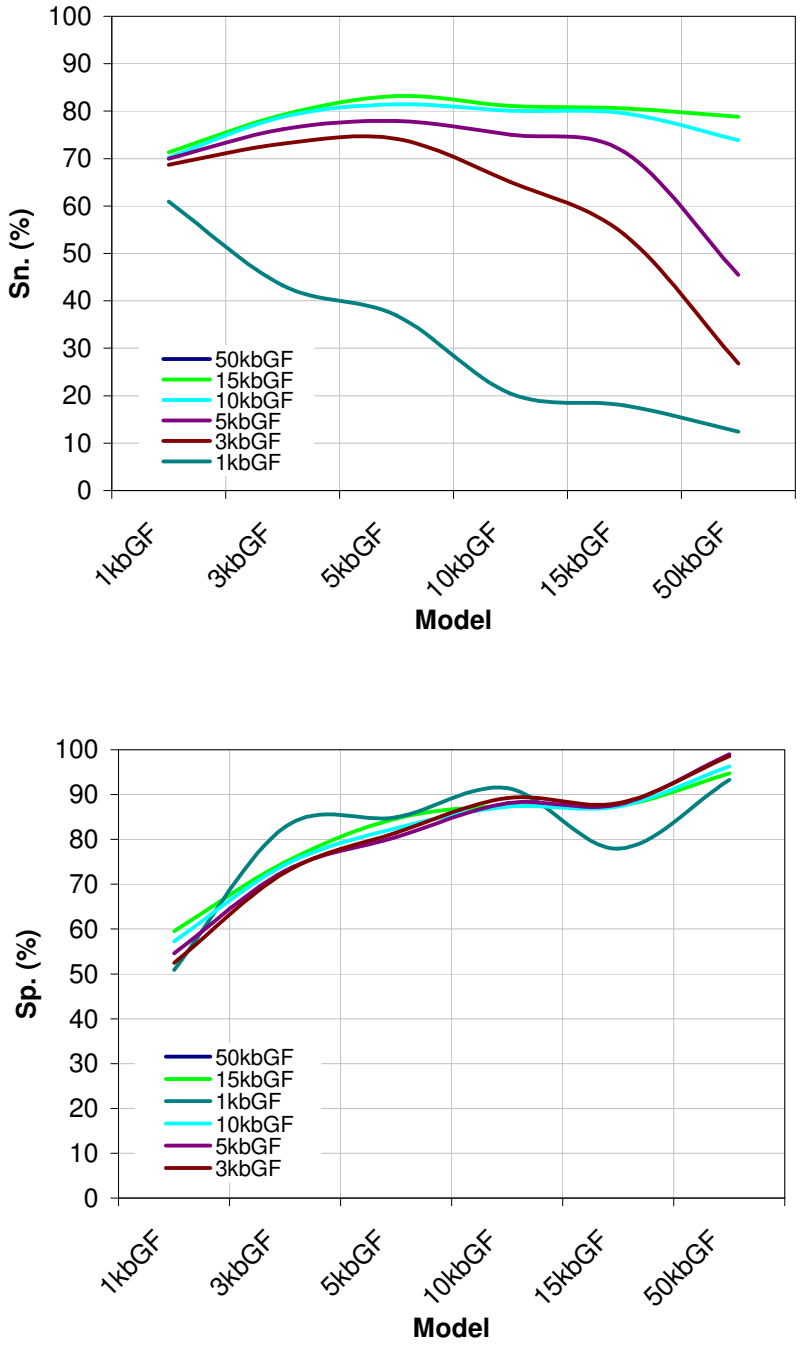
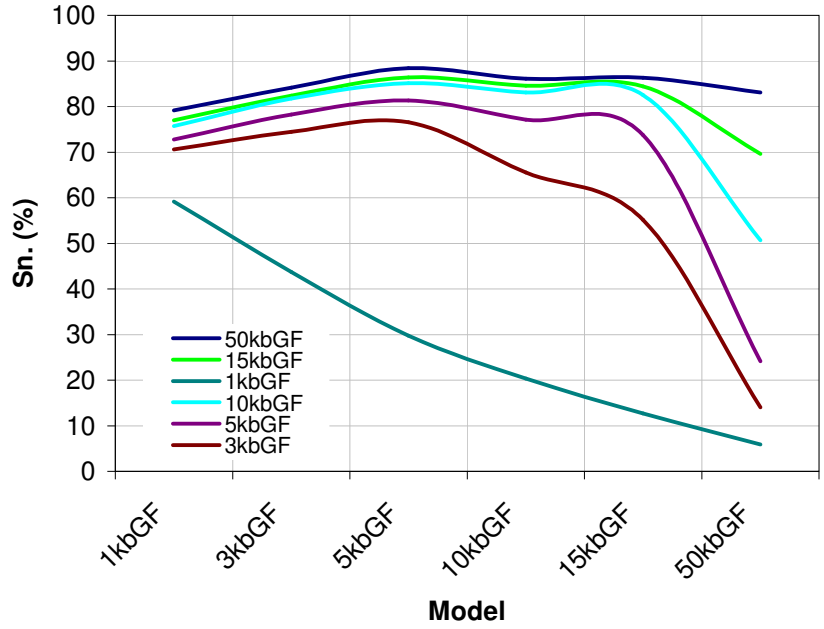


Fig. S4. Evaluation of the relation of genomic fragment length used for model creation and classification accuracy for genomic fragments of unknown organisms and different lengths. For the level phylum, the sensitivity (Sn.) and specificity (Sp.) of classification is displayed for genomic fragments of different lengths with classifiers trained on fragments of different lengths.



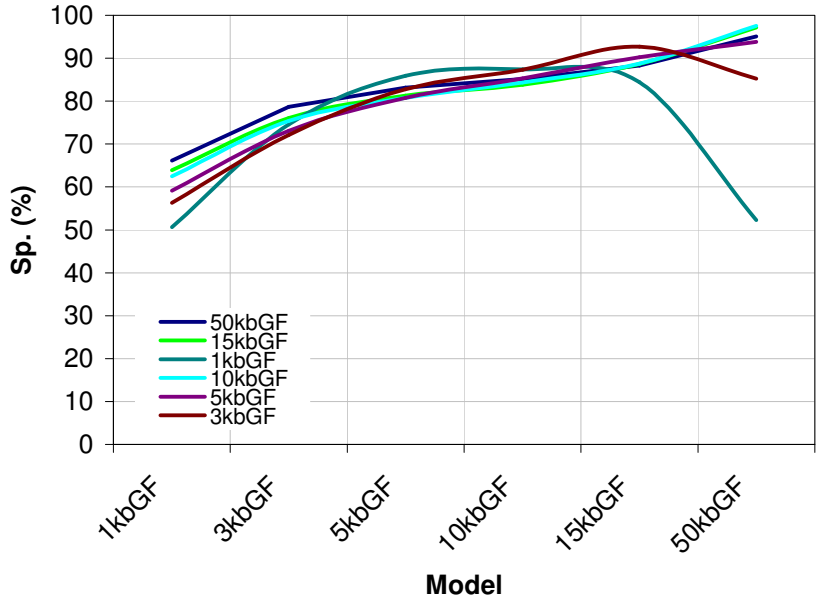


Fig. S5. Evaluation of the relation of genomic fragment length used for model creation and classification accuracy for genomic fragments of unknown organisms and different lengths. For the level class, the sensitivity (Sn.) and specificity (Sp.) of classification is displayed for genomic fragments of different lengths with classifiers trained on fragments of different lengths.

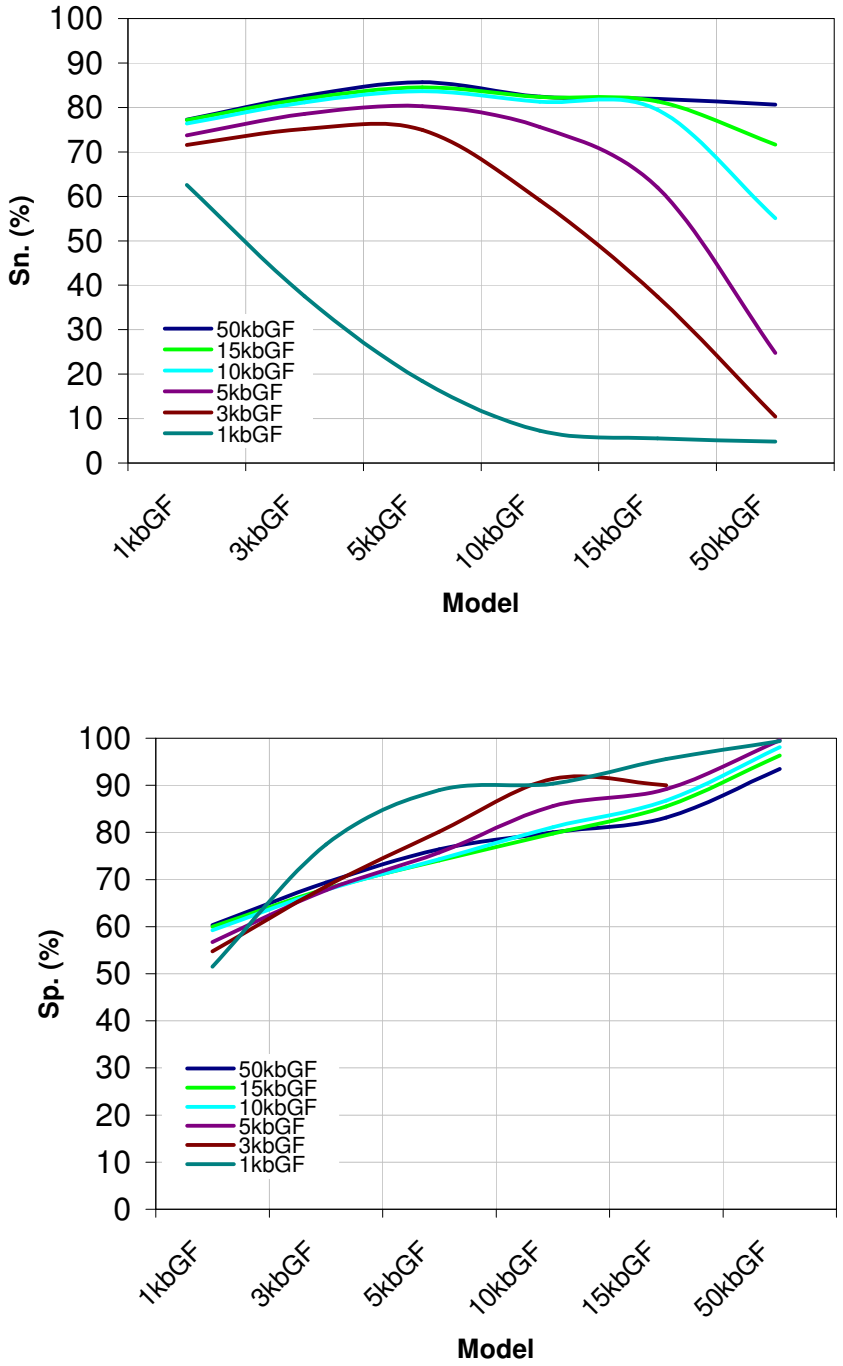


Fig. S6. Evaluation of the relation of genomic fragment length used for model creation and classification accuracy for genomic fragments of unknown organisms and different lengths. For the level order, the sensitivity (Sn.) and specificity (Sp.) of classification is displayed for genomic fragments of different lengths with classifiers trained on fragments of different lengths.

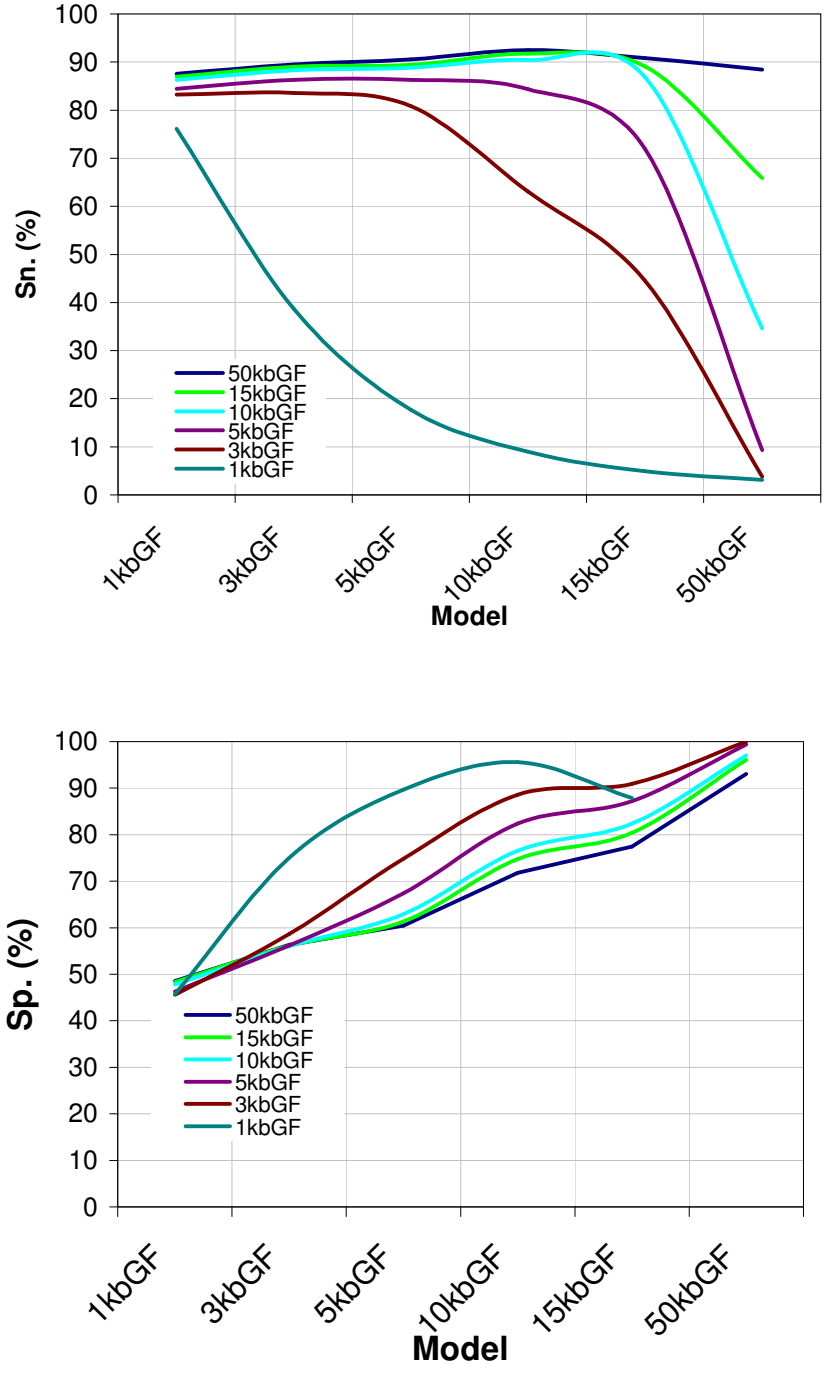


Fig. S7. Evaluation of the relation of genomic fragment length used for model creation and classification accuracy for genomic fragments of unknown organisms and different lengths. For the level genus, the sensitivity (A) and specificity (B) of classification is displayed for genomic fragments of different lengths with classifiers trained on fragments of different lengths.

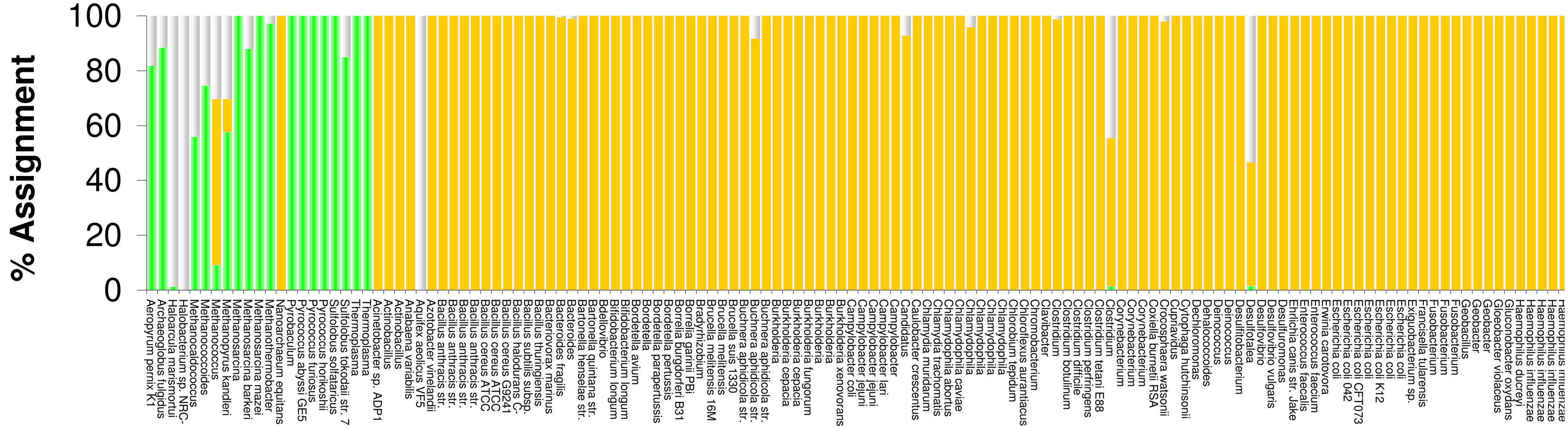
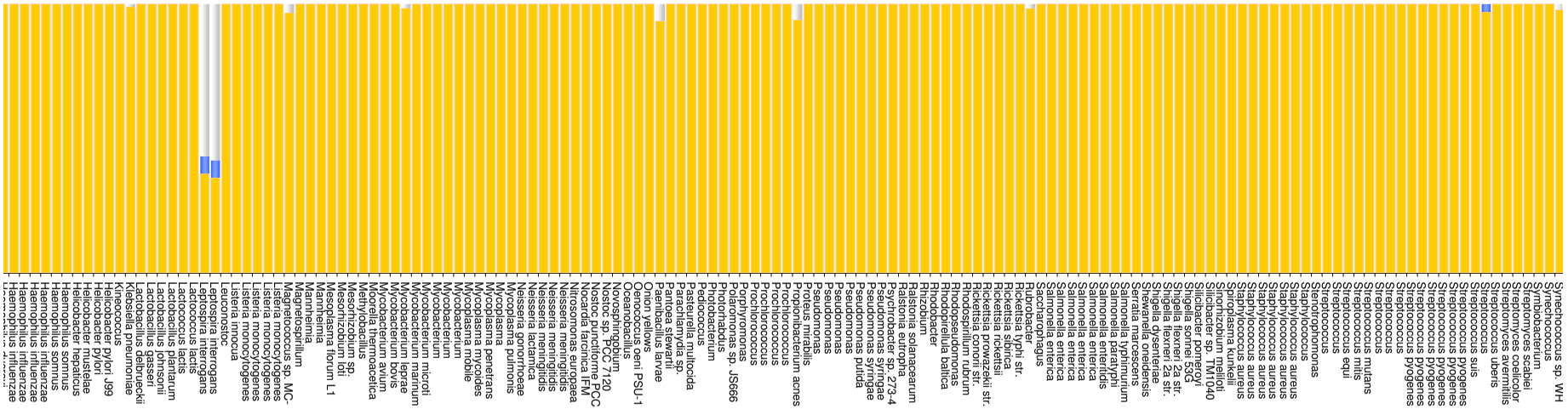
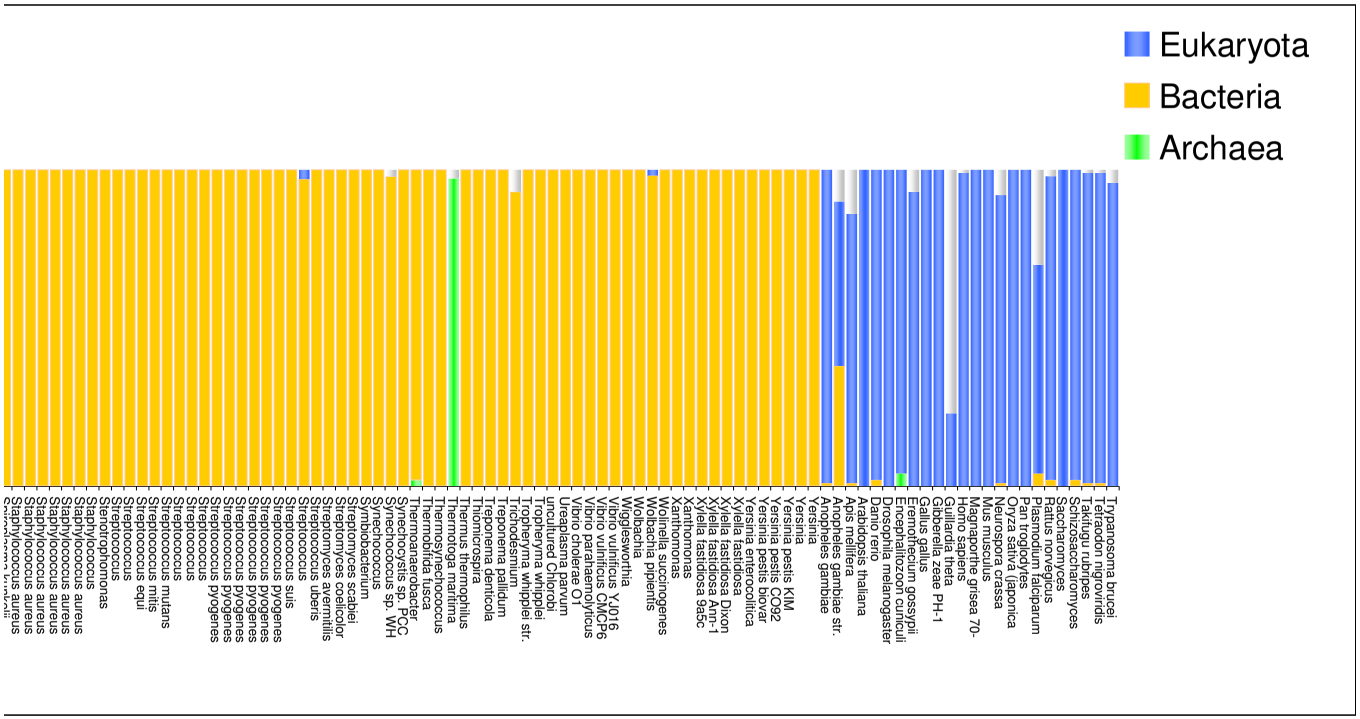


Fig. S8. Assignments at the domain level with Phylopythia for 50kb genomic fragments from unknown organisms (of which no sequence was used in training). Displayed is for every organism the percentage of fragments assigned to the archaea (green), bacteria (orange) and eukaryota (blue). Grey indicates an assignment to ‘origin unknown’.

Supplementary Figure S8



Supplementary Figure S8



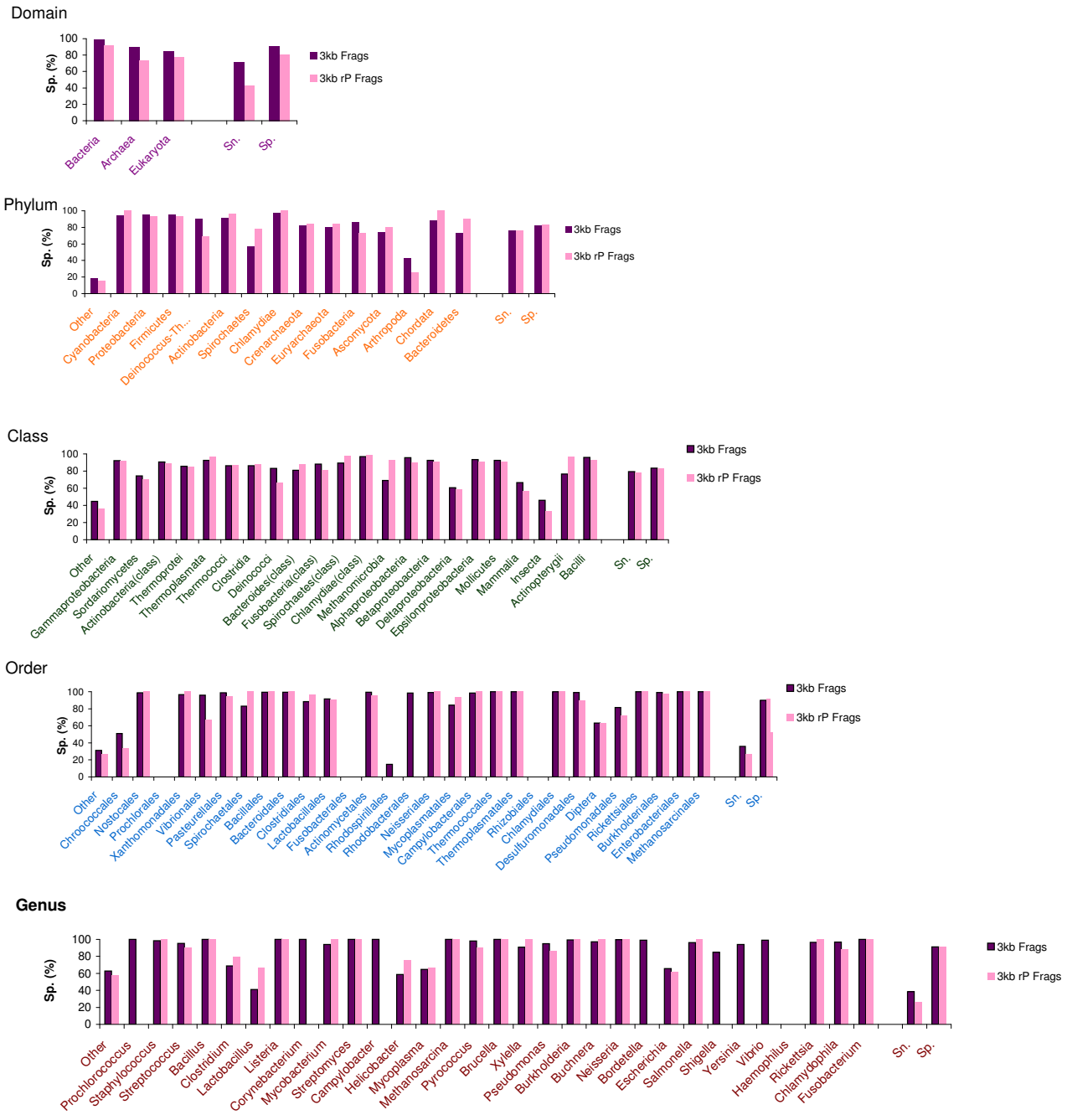


Fig. S9. Comparison of classification accuracy for 3kb fragments and 3kb fragments carrying ribosomal proteins with Phylopythia. For the clades at the domain, phylum, class, order, and genus level, the specificity of assignment for fragments from organisms unknown to the classifier is displayed, as well as the overall sensitivity and specificity of classification.

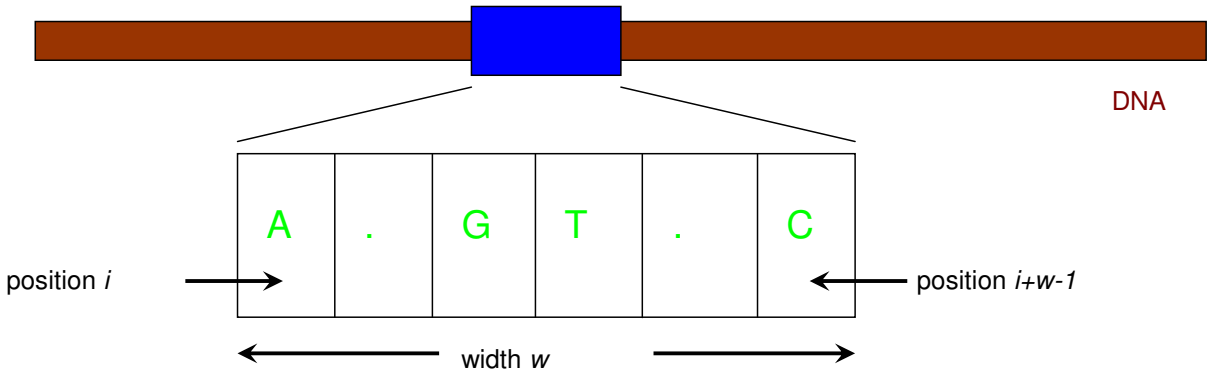


Fig. S10. Example of a (*w*, *l*)-pattern, where *w* is 6 and the number of literals *l* is 4. In this context, literals correspond to specified characters of the DNA alphabet, whereas the dot symbols (wildcards) match any of these characters.

Table S1. Species with (near-) complete genome sequences included in this study. The right column gives the number of organisms for every species.

Species	NCBI ID	# Organisms
Acinetobacter sp. ADP1	62977	1
Actinobacillus actinomycetemcomitans	714	1
Actinobacillus pleuropneumoniae	715	1
Aeropyrum pernix	56636	1
Anabaena variabilis	1172	1
Anopheles gambiae	7165	2
Apis mellifera	7460	1
Aquifex aeolicus	63363	1
Arabidopsis thaliana	3702	1
Archaeoglobus fulgidus	2234	1
Azotobacter vinelandii	354	1
Bacillus anthracis	1392	4
Bacillus cereus	1396	3
Bacillus halodurans	86665	1
Bacillus subtilis	1423	1
Bacillus thuringiensis	1428	1
Bacteriovorax marinus	97084	1
Bacteroides fragilis	817	2
Bacteroides thetaiotaomicron	818	1
Bartonella henselae	38323	1
Bartonella quintana	803	1
Bdellovibrio bacteriovorus	959	1
Bifidobacterium longum	216816	2
Bordetella avium	521	1
Bordetella bronchiseptica	518	1
Bordetella parapertussis	519	1
Bordetella pertussis	520	1
Borrelia burgdorferi	139	1
Borrelia garinii	29519	1
Bradyrhizobium japonicum	375	1
Brucella melitensis	29459	3
Buchnera aphidicola	9	3
Burkholderia cenocepacia	95486	1
Burkholderia cepacia	292	2
Burkholderia fungorum	134537	1
Burkholderia pseudomallei	28450	1
Burkholderia vietnamiensis	60552	1
Burkholderia xenovorans	36873	1
Caenorhabditis elegans	6239	1
Campylobacter coli	195	1
Campylobacter jejuni	197	2
Campylobacter lari	201	1
Campylobacter upsaliensis	28080	1
Candidatus Blochmannia floridanus	203907	1
Canis familiaris	9615	1
Caulobacter vibrioides	155892	1
Chlamydia muridarum	83560	1
Chlamydia trachomatis	813	1
Chlamydophila abortus	83555	1

Supplementary Table S1

Chlamydophila caviae	83557	1
Chlamydophila pneumoniae	83558	4
Chlorobaculum tepidum	1097	1
Chloroflexus aurantiacus	1108	1
Chromobacterium violaceum	536	1
Clavibacter michiganensis	28447	1
Clostridium acetobutylicum	1488	1
Clostridium botulinum	1491	1
Clostridium difficile	1496	1
Clostridium perfringens	1502	1
Clostridium tetani	1513	1
Clostridium thermocellum	1515	1
Corynebacterium diphtheriae	1717	1
Corynebacterium efficiens	152794	1
Corynebacterium glutamicum	1718	1
Coxiella burnetii	777	1
Crocosphaera watsonii	263511	1
Cupriavidus metallidurans	119219	1
Cupriavidus necator	106590	1
Cytophaga hutchinsonii	985	1
Danio rerio	7955	1
Dechloromonas aromatica	259537	1
Dehalococcoides ethenogenes	61435	1
Deinococcus geothermalis	68909	1
Deinococcus radiodurans	1299	1
Desulfitobacterium hafniense	49338	1
Desulfotalea psychrophila	84980	1
Desulfovibrio desulfuricans	876	1
Desulfovibrio vulgaris	881	1
Desulfuromonas acetoxidans	891	1
Drosophila melanogaster	7227	1
Ehrlichia canis	944	1
Encephalitozoon cuniculi	6035	1
Enterococcus faecalis	1351	1
Enterococcus faecium	1352	1
Eremothecium gossypii	33169	1
Escherichia coli	562	7
Exiguobacterium sp. 255-15	262543	1
Ferroplasma acidarmanus	97393	1
Francisella tularensis	263	1
Fusobacterium nucleatum	851	3
Gallus gallus	9031	1
Geobacillus stearothermophilus	1422	1
Geobacter metallireducens	28232	1
Geobacter sulfurreducens	35554	1
Gibberella zeae	5518	1
Gloeobacter violaceus	33072	1
Gluconobacter oxydans	442	1
Guillardia theta	55529	1
Haemophilus ducreyi	730	1
Haemophilus influenzae	727	4
Haloarcula marismortui	2238	1
Halobacterium salinarum	2242	1
Helicobacter hepaticus	32025	1
Helicobacter mustelae	217	1

Supplementary Table S1

Helicobacter pylori	210	2
Histophilus somni	731	2
Homo sapiens	9606	1
Kineococcus radiotolerans	131568	1
Klebsiella pneumoniae	573	1
Lactobacillus delbrueckii	1584	1
Lactobacillus gasseri	1596	1
Lactobacillus johnsonii	33959	1
Lactobacillus plantarum	1590	1
Lactococcus lactis	1358	2
Leptospira interrogans	173	2
Leuconostoc mesenteroides	1245	1
Listeria innocua	1642	1
Listeria monocytogenes	1639	4
Magnaporthe grisea	148305	1
Magnetococcus sp. MC-1	156889	1
Magnetospirillum magnetotacticum	188	1
Mannheimia haemolytica	75985	1
Mannheimia succiniciproducens	157673	1
Mesoplasma florum	2151	1
Mesorhizobium loti	381	1
Mesorhizobium sp. BNC1	266779	1
Methanocaldococcus jannaschii	2190	1
Methanococcoides burtonii	29291	1
Methanococcus maripaludis	39152	1
Methanopyrus kandleri	2320	1
Methanosarcina acetivorans	2214	1
Methanosarcina barkeri	2208	1
Methanosarcina mazei	2209	1
Methanothermobacter thermautotrophicus	145262	1
Methylobacillus flagellatus	405	1
Moorella thermoacetica	1525	1
Mus musculus	10090	1
Mycobacterium avium	1764	1
Mycobacterium bovis	1765	1
Mycobacterium leprae	1769	1
Mycobacterium marinum	1781	1
Mycobacterium microti	1806	1
Mycobacterium smegmatis	1772	1
Mycobacterium tuberculosis	1773	2
Mycoplasma mobile	2118	1
Mycoplasma mycoides	2102	1
Mycoplasma penetrans	28227	1
Mycoplasma pneumoniae	2104	1
Mycoplasma pulmonis	2107	1
Nanoarchaeum equitans	160232	1
Neisseria gonorrhoeae	485	1
Neisseria lactamica	486	1
Neisseria meningitidis	487	3
Neurospora crassa	5141	1
Nitrosomonas europaea	915	1
Nocardia farcinica	37329	1
Nostoc punctiforme	272131	1
Nostoc sp. PCC 7120	103690	1
Novosphingobium aromaticivorans	48935	1

Supplementary Table S1

Oceanobacillus iheyensis	182710	1
Oenococcus oeni	1247	1
Onion yellows phytoplasma	100379	1
Oryza sativa	4530	1
Paenibacillus larvae	1464	1
Pan troglodytes	9598	1
Pantoea stewartii	66269	1
Parachlamydia sp. UWE25	264201	1
Pasteurella multocida	747	1
Pectobacterium atrosepticum	29471	1
Pediococcus pentosaceus	1255	1
Photobacterium profundum	74109	1
Photorhabdus luminescens	29488	1
Picrophilus torridus	82076	1
Plasmodium falciparum	5833	1
Polaromonas sp. JS666	296591	1
Porphyromonas gingivalis	837	1
Prochlorococcus marinus	1219	4
Propionibacterium acnes	1747	1
Proteus mirabilis	584	1
Pseudomonas aeruginosa	287	2
Pseudomonas fluorescens	294	2
Pseudomonas putida	303	1
Pseudomonas syringae	317	1
Pseudomonas syringae group genomosp. 3	251701	1
Psychrobacter sp. 273-4	259536	1
Pyrobaculum aerophilum	13773	1
Pyrococcus abyssi	29292	1
Pyrococcus furiosus	2261	1
Pyrococcus horikoshii	53953	1
Ralstonia solanacearum	305	1
Rattus norvegicus	10116	1
Rhizobium leguminosarum	384	1
Rhodobacter sphaeroides	1063	1
Rhodopirellula baltica	265606	1
Rhodopseudomonas palustris	1076	1
Rhodospirillum rubrum	1085	1
Rickettsia conorii	781	1
Rickettsia prowazekii	782	1
Rickettsia rickettsii	783	1
Rickettsia sibirica	35793	1
Rickettsia typhi	785	1
Rubrobacter xylanophilus	49319	1
Saccharomyces cerevisiae	4932	1
Saccharophagus degradans	86304	1
Salmonella enterica	28901	5
Salmonella enteritidis	592	1
Salmonella paratyphi	54388	1
Salmonella typhimurium	602	1
Schizosaccharomyces pombe	4896	1
Serratia marcescens	615	1
Shewanella oneidensis	70863	1
Shigella dysenteriae	622	1
Shigella flexneri	623	2
Shigella sonnei	624	1

Supplementary Table S1

Silicibacter pomeroyi	89184	1
Silicibacter sp. TM1040	292414	1
Sinorhizobium meliloti	382	1
Spiroplasma kunkelii	47834	1
Staphylococcus aureus	1280	6
Staphylococcus epidermidis	1282	1
Stenotrophomonas maltophilia	40324	1
Streptococcus agalactiae	1311	2
Streptococcus equi	1336	1
Streptococcus mitis	28037	1
Streptococcus mutans	1309	1
Streptococcus pneumoniae	1313	3
Streptococcus pyogenes	1314	6
Streptococcus suis	1307	1
Streptococcus thermophilus	1308	1
Streptococcus uberis	1349	1
Streptomyces avermitilis	33903	1
Streptomyces coelicolor	1902	1
Streptomyces scabiei	1930	1
Sulfolobus solfataricus	2287	1
Sulfolobus tokodaii	111955	1
Symbiobacterium thermophilum	2734	1
Synechococcus elongatus	32046	1
Synechococcus sp. WH 8102	84588	1
Synechocystis sp. PCC 6803	1148	1
Takifugu rubripes	31033	1
Tetraodon nigroviridis	99883	1
Thermoanaerobacter tengcongensis	119072	1
Thermobifida fusca	2021	1
Thermoplasma acidophilum	2303	1
Thermoplasma volcanium	50339	1
Thermosynechococcus elongatus	146786	1
Thermotoga maritima	2336	1
Thermus thermophilus	274	1
Thiomicrospira crunogena	39765	1
Treponema denticola	158	1
Treponema pallidum	160	1
Trichodesmium erythraeum	1206	1
Tropheryma whipplei	2039	2
Trypanosoma brucei	5691	1
Ureaplasma parvum	134821	1
Vibrio cholerae	666	1
Vibrio parahaemolyticus	670	1
Vibrio vulnificus	672	2
Wigglesworthia glossinidia	51229	1
Wolbachia endosymbiont of *D. melanogaster*	163164	1
Wolbachia pipientis	955	2
Wolinella succinogenes	844	1
Xanthomonas axonopodis	53413	1
Xanthomonas campestris	339	1
Xylella fastidiosa	2371	4
Yersinia enterocolitica	630	1
Yersinia pestis	632	3
Yersinia pseudotuberculosis	633	3
uncultured Chlorobi bacterium	156405	1

Table S2. Wn parameter search for the sequence composition space with the highest classification accuracy for genomic fragments of 340 unknown organisms (to the classifier) at different phylogenetic levels. The composition space is defined by the word length *w*, the number of literal characters *l*, and the step size *s* . *Acc.* denotes the percentage of correctly assignments for all tested fragments, *Sn.* and *Sp.* are the sensitivity and specificity of the classification (Supplementary material, 'Evaluation procedures').

Rank	kernel	Wn	*Acc.* (%)	*Sn.* (%)	*Sp.* (%)
Genus	gaussian	w2,l2,s1	38.3	37	25.5
Genus	gaussian	w3,l2,s1	41.7	52.1	33.1
Genus	gaussian	w3,l3,s1	84.3	76.6	81.9
Genus	gaussian	w4,l4,s1	88.1	86.3	82.9
Genus	gaussian	w6,l4,s1	87.2	87.9	80.8
Genus	gaussian	**w5,l5,s1**	**87.6**	**89.1**	**80.4**
Genus	gaussian	w6,l5,s1	86.6	88.4	79.5
Genus	gaussian	w6,l6,s1	86.6	88.4	79.5
Order	gaussian	w2,l2,s1	43.3	33.1	62.7
Order	gaussian	w3,l2,s1	61.3	54.3	81
Order	gaussian	w3,l3,s1	77.1	70	84
Order	gaussian	w4,l4,s1	84.2	80.2	85.5
Order	gaussian	w6,l4,s1	84.9	81.1	83.9
Order	gaussian	**w5,l5,s1**	**85.1**	**81.4**	**85.5**
Order	gaussian	w6,l5,s1	84.9	81.1	83.9
Order	gaussian	w6,l6,s1	84.4	81	82.9
Class	gaussian	w2,l2,s1	43.2	39.6	68.5
Class	gaussian	w3,l2,s1	65.4	62.2	81.8
Class	gaussian	w3,l3,s1	77.9	74	88.1
Class	gaussian	w4,l4,s1	85.7	82.8	89.1
Class	gaussian	w6,l4,s1	87	83.8	88.8
Class	gaussian	**w5,l5,s1**	**87.1**	**84.4**	**88.9**
Class	gaussian	w6,l5,s1	87.2	83.8	88.5
Class	gaussian	w6,l6,s1	86.7	83.7	87.3
Phylum	gaussian	w2,l2,s1	43.1	37.2	34
Phylum	gaussian	w3,l2,s1	53.3	48.9	44.4
Phylum	gaussian	w3,l3,s1	59.6	56	51
Phylum	gaussian	w4,l4,s1	85.7	78.5	86.2
Phylum	gaussian	**w6,l4,s1**	**87.5**	**79.6**	**87.4**
Phylum	gaussian	w5,l5,s1	87.2	79.4	87.8
Phylum	gaussian	w6,l5,s1	87.4	79.1	87.5
Phylum	gaussian	w6,l6,s1	87	78.6	87.8
Domain	gaussian	w2,l2,s1	74.5	56.5	66.5
Domain	gaussian	w3,l2,s1	87.7	69	87.4
Domain	gaussian	w3,l3,s1	91	76.2	92.3
Domain	gaussian	w4,l4,s1	94.7	84.5	93.3
Domain	gaussian	w6,l4,s1	95.4	86.7	91.9
Domain	gaussian	w5,l5,s1	95.1	87.3	92.5
Domain	gaussian	w6,l5,s1	95.8	88.8	92.3
Domain	gaussian	**w6,l6,s1**	**95.7**	**90.1**	**92.5**

Table S3. Wn parameter search for the sequence composition space with the highest classification accuracy for protein encoding genes of 340 unknown organisms at different phylogenetic levels. The composition space is defined by the word length w, the number of literal characters l, and the step size s. Results are shown for the multi-class classification without post-processing by the binary classifiers. *Acc.* denotes the percentage of correctly assignments for all tested fragments, *Sn.* and *Sp.* are the sensitivity and specificity of the classification (Supplementary material, 'Evaluation procedures').

Rank	kernel	Wn	*Acc.* (%)	*Sn.* (%)	*Sp.* (%)
Genus	gaussian	w2,l2,s1	*43.1*	*38.1*	*28.7*
Genus	gaussian	w2,l2,s3	26.9	28.2	15.6
Genus	gaussian	w3,l2,s3	40	57.8	31.7
Genus	gaussian	w3,l3,s1	49.7	67.4	39.4
Genus	gaussian	w3,l3,s3	49	67.1	38.5
Genus	gaussian	w4,l4,s1	50.7	74.2	41.5
Genus	gaussian	w4,l4,s3	51.7	75.3	42.4
Genus	gaussian	w6,l4,s3	57	80.9	47.1
Genus	gaussian	w5,l5,s1	52.7	76.5	43.4
Genus	gaussian	w5,l5,s3	55.1	78	45.3
Genus	gaussian	w6,l5,s3	57.1	81.2	47.2
Genus	gaussian	w6,l6,s1	54.5	77.6	44.7
Genus	**gaussian**	**w6,l6,s3**	**58.2**	**80**	**47.9**
Class	gaussian	w2,l2,s1	34.3	36.2	35.7
Class	gaussian	w2,l2,s3	19.6	27.5	26.2
Class	gaussian	w3,l2,s3	40.2	50	40.6
Class	gaussian	w3,l3,s1	51.1	57.5	49.2
Class	gaussian	w3,l3,s3	53.9	60.9	51.3
Class	gaussian	w4,l4,s1	57.1	65.3	52.5
Class	gaussian	w4,l4,s3	58.3	67.9	54.5
Class	**gaussian**	**w6,l4,s3**	**63.3**	**72.7**	**59.2**
Class	gaussian	w5,l5,s1	57.6	66.9	53
Class	gaussian	w5,l5,s3	59.4	69.8	56.6
Class	gaussian	w6,l5,s3	62.3	71.9	58.5
Class	gaussian	w6,l6,s1	57.8	67.7	54.2
Class	gaussian	w6,l6,s3	60.6	70.2	57.5
Phylum	gaussian	w2,l2,s1	40.2	37.4	35.5
Phylum	gaussian	w2,l2,s3	26.3	37.1	30.3
Phylum	gaussian	w3,l2,s3	42.4	53.9	40.5
Phylum	gaussian	w3,l3,s1	57.2	57	49.1
Phylum	gaussian	w3,l3,s3	58.1	58.4	50.8
Phylum	gaussian	w4,l4,s1	60.2	63.2	50.8
Phylum	gaussian	w4,l4,s3	58.9	65.1	50.8
Phylum	**gaussian**	**w6,l4,s3**	**63**	**69**	**54.6**
Phylum	gaussian	w5,l5,s1	57.5	63.3	48.9
Phylum	gaussian	w5,l5,s3	57.7	66	51.4
Phylum	gaussian	w6,l5,s3	61.8	68.2	54.3
Phylum	gaussian	w6,l6,s1	56.6	63.5	49.2
Phylum	gaussian	w6,l6,s3	59.9	66.6	52.7
Domain	gaussian	w2,l2,s1	72.3	56.1	57.9
Domain	gaussian	w2,l2,s3	61.8	61.5	54.7
Domain	gaussian	w3,l2,s3	76.8	71.6	65.7
Domain	gaussian	w3,l3,s1	76.6	69.4	62.9
Domain	gaussian	w3,l3,s3	82.9	78.5	71.7
Domain	gaussian	w4,l4,s1	83.4	74.2	71.8
Domain	gaussian	w4,l4,s3	86	80.4	75
Domain	**gaussian**	**w6,l4,s3**	**86.6**	**82.2**	**75.5**
Domain	gaussian	w5,l5,s1	80	75.6	66.3
Domain	gaussian	w5,l5,s3	84.1	79.7	72.1
Domain	gaussian	w6,l5,s3	81.5	80	69.5
Domain	gaussian	w6,l6,s1	n.d.	n.d.	n.d.
Domain	gaussian	w6,l6,s3	84.4	81.1	74.1

Table S4: Evaluation at domain level of the relation of the fragment length used for model creation to the accuracy of assignments for genomic fragments from unknown organisms. For every clade, the clade-specific sensitivity (class *Sn*.), specificity (class *Sp.*) of assignments for genomic fragments of different lengths from the organisms of this clade among the 340 organisms is shown. The two rightmost columns display the overall sensitivity (*Sn.*) and specificity (*Sp.*) of de novo sequence classification (Supplementary Material, 'Evaluation procedures').

Eval With	Built On	Bacteria	Archaea	Eukaryota	Bacteria	Archaea	Eukaryota	*Sn.* (%)	*Sp.* (%)
		Class *Sn.* (%)			Class *Sp.* (%)				
1kbGF	1kbGF	77.51	68.76	74.96	99.49	43.19	42.99	**73.74**	61.89
1kbGF	3kbGF	85.22	34.14	53.67	97.38	77.01	69.47	57.68	81.29
1kbGF	5kbGF	93.29	0	22.83	94.4	-	88.96	38.71	**91.68**
1kbGF	10kbGF	98.19	0	10.5	89.22	-	85.42	36.23	87.32
1kbGF	15kbGF	98.47	0	4.5	88.87	-	73.97	34.32	81.42
1kbGF	50kbGF	99.87	0	4.75	86.89	-	75	34.87	80.94
3kbGF	1kbGF	84.57	73.24	81.79	99.23	47.59	53.85	79.87	66.89
3kbGF	3kbGF	87.49	75.1	85.17	99.26	65.22	67.64	**82.58**	77.37
3kbGF	5kbGF	92.39	48.86	74.88	98.86	89.92	85.13	72.04	91.3
3kbGF	10kbGF	96.49	0	55.17	97.01	-	92.52	50.55	**94.77**
3kbGF	15kbGF	97.23	26.52	54.75	96.34	79.57	95.01	59.5	90.31
3kbGF	50kbGF	99.93	0	1.96	86.83	-	70.15	33.96	78.49
5kbGF	1kbGF	86.42	75.74	84.12	99.14	49.95	60	82.1	69.7
5kbGF	3kbGF	87.59	79.08	87.75	99.35	62.89	68.82	**84.8**	77.02
5kbGF	5kbGF	91.9	75.41	84.62	99.21	71.55	83.31	83.98	84.69
5kbGF	10kbGF	95.9	33.22	78.33	98.57	99.71	92.47	69.15	**96.92**
5kbGF	15kbGF	97.07	54.05	80.29	98.37	83.63	94.09	77.14	92.03
5kbGF	50kbGF	99.94	0	10.38	89.09	-	94.68	36.77	91.88
10kbGF	1kbGF	88.07	81.02	86.95	99.39	52.28	65.5	85.34	72.39
10kbGF	3kbGF	88.37	85.36	93.28	99.73	62.24	72.27	89	78.08
10kbGF	5kbGF	91.81	85.16	90.41	99.59	65.31	82.66	**89.13**	82.52
10kbGF	10kbGF	94.98	72.77	90.03	99.42	72.1	94.04	85.93	88.52
10kbGF	15kbGF	96.9	71.84	90.49	99.24	84.65	92.33	86.41	92.07
10kbGF	50kbGF	99.25	33.24	61.39	97.3	99.71	98.71	64.62	**98.57**
15kbGF	1kbGF	88.99	83.32	88.86	99.56	55.18	69.98	87.05	74.91
15kbGF	3kbGF	89.28	88.93	94.84	99.8	63.87	75.84	91.01	79.84
15kbGF	5kbGF	92.53	89.32	93.12	99.68	66.43	84.2	**91.65**	83.44
15kbGF	10kbGF	94.84	79.53	94.02	99.61	69.15	95.46	89.46	88.07
15kbGF	15kbGF	97.09	79.38	93.8	99.44	86.07	92.1	90.09	92.54
15kbGF	50kbGF	98.98	51.87	79.73	98.56	97.68	99.09	76.86	**98.44**
50kbGF	1kbGF	89.53	87.39	94.05	99.73	51.68	78.62	90.32	76.68
50kbGF	3kbGF	91	90.93	97.54	99.84	58.8	86.59	93.16	81.74
50kbGF	5kbGF	92.66	91.69	96.75	99.78	61.84	88.05	93.7	83.23
50kbGF	10kbGF	94.89	82.53	96.52	99.77	58.83	99.29	91.31	85.97
50kbGF	15kbGF	97.19	91.96	96.89	99.66	90.06	91.93	**95.35**	93.88

Table S5: Evaluation at phylum level of the relation of the fragment length used for model creation to the accuracy of assignments for genomic fragments from unknown organisms. For every clade, the clade-specific sensitivity (class *Sn.*), specificity (class S*p*.) of assignments for genomic fragments of different lengths from the organisms of this clade among the 340 organisms is shown. The two rightmost columns display the overall sensitivity (*Sn.*) and specificity (*Sp.*) of de novo sequence classification (Supplementary Material, 'Evaluation procedures').

Eval With	Built On	Other Class *Sn.* (%)	yanobacter	roteobacter	Firmicutes	coccus-Th	ctinobacter	Spirochaete	Chlamydiae	renarchaeo	Euryarchae	Fusobacter	Ascomycot	Arthropoda	Chordata	Bacteroide
1kbGF	1kbGF	34.27	66.58	64.94	58.01	48.67	80.62	13.33	87.56	60	55	95.33	44.67	48	84.11	72.67
1kbGF	3kbGF	64.33	23.83	70.4	32.85	38.67	48.96	11.33	36.33	56.33	49.11	51.33	8	66	81.33	10.25
1kbGF	5kbGF	79.8	0.17	64.48	12.69	32.67	19.67	29.83	0	36.33	20.33	79.67	2.83	52	83.89	39.25
1kbGF	10kbGF	99.13	0	5.57	0.9	16	50.04	20.67	0	1.67	0.22	0	0	1	87.89	23.67
1kbGF	15kbGF	94	0	0.49	0	3.33	0	22.67	0	19.33	0	0	0	22.33	84.22	23.33
1kbGF	50kbGF	100	10.17	0	0	0	9.88	0	0	0	0	0	0	0	66.22	0
3kbGF	1kbGF	20.33	81	71.81	61.81	61	89.42	13	90.22	81.67	55.11	99	61.17	59.67	93.11	92.33
3kbGF	3kbGF	42.87	76.83	83.22	75.64	59	92.25	20.67	88.22	76.33	61.83	98	69.83	76	92.56	83.75
3kbGF	5kbGF	51.4	78.5	83.2	74.56	56.67	88.75	57.17	88.33	73.33	66.67	98	50.83	67.33	94.89	82.25
3kbGF	10kbGF	77.33	64	80.45	71	51	86.96	24.33	64.44	63.33	57.44	73.67	31.83	62.67	96.56	70.33
3kbGF	15kbGF	83.67	38.08	80.28	51.62	45.67	74.12	25.5	27.89	65.33	57.44	22.33	8.5	58.67	94.56	76.33
3kbGF	50kbGF	97.13	54.25	37.6	28.1	0	70.12	4.33	0	1	7.5	1	0	8.67	92.11	0
5kbGF	1kbGF	15.82	83.33	73.77	61.43	64.67	90.04	18	89.89	86	54.06	99.33	62	58.33	95.89	97.33
5kbGF	3kbGF	34.58	83.25	85.32	79.74	63.67	94.67	25.17	88.78	84.33	63.11	99.33	79.83	75.67	94.89	90.5
5kbGF	5kbGF	43.66	89.25	86.53	79.54	64	91.08	60.83	88.67	78.67	69.17	99.67	64.33	66.33	97.33	90
5kbGF	10kbGF	68.36	81.83	86.54	80.68	60.33	93.17	27.5	87.22	67	68.11	95.33	65	68.67	98.33	77.67
5kbGF	15kbGF	76.7	67.58	88.48	75.76	55.67	86.21	29.67	83.11	68	68.56	79.67	43.67	68	98.11	82
5kbGF	50kbGF	91.52	68.75	74.52	67.08	6.33	82.79	19.33	5.33	27	35.17	35.33	8.5	34	96	30.33
10kbGF	1kbGF	11.44	84.25	76.88	61.66	66.67	91.07	15.38	89.56	88.67	51.83	99.67	63.17	60	97.78	98.33
10kbGF	3kbGF	24.08	88.5	87.55	82.2	66.67	96.81	25.42	88.89	92.33	65.06	99.67	89.5	79.67	97.22	94
10kbGF	5kbGF	30.93	93.83	90.01	83.3	66	94.76	64.05	88.67	84.33	73.22	99.67	78.17	68.33	99.67	91.75
10kbGF	10kbGF	51.55	88.25	90.59	84.99	65	96.73	29.77	88.56	77	78.33	99.33	81.5	76.67	99.67	81.33
10kbGF	15kbGF	65.82	81.5	92.55	83.18	64.67	90.65	32.27	88.44	77.67	77.56	98	70.33	78.33	99.56	82.33
10kbGF	50kbGF	84.96	79.33	88.61	83.72	30.33	89.77	28.6	73.33	57.67	57	84.33	54.83	59.33	99.22	62.67
15kbGF	1kbGF	10.27	85.75	79.11	63.12	66.67	91.69	13.89	86.69	91.67	51.22	100	64.83	51.67	98.44	98.67
15kbGF	3kbGF	20.62	90.92	88.81	83.16	67	97.03	22.96	86.27	93.33	65.17	100	93.83	78	98.44	96.75
15kbGF	5kbGF	25.94	95.25	91.47	85.12	66.33	95.61	64.26	86.27	85.67	75.28	100	86.17	70	99.67	94.75
15kbGF	10kbGF	46.49	89.83	92.19	86.33	66	96.73	26.48	85.85	81	81.72	99.33	91.33	76.67	99.78	82
15kbGF	15kbGF	61.27	84.08	93.79	85.95	66	91.47	30	85.85	79.67	81	99	75.83	79.67	99.78	80.33
15kbGF	50kbGF	81.74	84.33	91.68	86.15	43.33	91.65	26.11	83.5	58.67	65.83	94.67	70.83	63.33	99.89	66.33
50kbGF	1kbGF	9.39	85.35	87.12	60.2	74.85	93.85	6.41	79.65	98.46	58.73	100	63.83	54.33	98.44	99.65
50kbGF	3kbGF	15.38	97.06	92.32	82.49	74.85	98.74	11.22	79.22	97.24	66.55	100	95	80.33	98.44	99.42
50kbGF	5kbGF	16.29	98	94.62	85.3	74.85	96.77	73.08	79.22	90.34	81.41	100	91.17	71.67	99.67	96.13
50kbGF	10kbGF	29.52	89.59	95.29	86.28	74.85	99.08	13.14	78.79	92.41	91.27	100	95.83	83.67	99.67	86.81
50kbGF	15kbGF	54.64	84.56	96.85	85.76	74.85	94.45	14.74	78.79	88.28	86.96	100	80.5	86.67	99.67	81.6
50kbGF	50kbGF	62.13	87.91	96.25	84.82	72.46	95.37	7.51	78.79	77.24	81.92	100	87.5	75.33	99.62	75

Supplementary Table S5

Other Class	Cyanobacti	Proteobact	Firmicutes	Deinococci	Actinobacte	Spirochaet	Chlamydiae	Crenarchae	Euryarchae	Fusobacter	Ascomycot	Arthropoda	Chordata	Bacteroidetes	*Sn.* (%)	*Sp.* (%)
Sp. (%)																
10.29	50.44	91.75	78.46	35.18	71.22	21.86	64.38	42.55	64.54	38.24	33.5	13.56	75.32	31.1	60.92	50.86
7.08	92.86	92.41	96.06	83.45	97.75	90.67	97.03	66.8	71.64	85.08	97.96	14.84	77.05	95.35	43.27	82.78
6.33	100	92.49	97.23	91.59	97.93	94.21	-	97.32	93.85	81.29	73.91	26.53	84.93	73.02	36.91	84.95
4.99	-	98.91	100	94.12	93.39	56.11	-	100	100	-	-	100	84.24	87.65	20.45	91.44
4.47	-	100	-	100	-	30.91	-	90.62	-	-	-	52.34	79.79	92.11	17.98	77.97
4.65	93.13	-	-	-	99.58	-	-	-	-	-	-	-	87.26	-	12.42	93.32
13.34	54.21	94.65	82.88	34.72	69.45	20.21	33.71	56.58	74.92	35.91	31.34	25.83	89.63	30.64	68.71	52.48
15.05	81.59	95.29	93.14	77.63	84.09	54.87	86.59	75.83	69	59.51	55.42	35.35	88.15	60.58	73.13	72.65
13.87	95.83	95.31	95.09	88.54	90.95	73.61	95.78	80	78.79	77.37	66.02	36.33	93.13	74.77	74.13	81.54
13.54	97.83	96.89	97.71	90.53	92.8	60.08	99.66	94.06	90.94	99.55	95.02	67.38	91.86	75.09	65.02	89.24
10.84	99.56	99.09	99.31	96.48	97.96	39.53	100	90.32	93.66	100	96.23	53.33	89.58	78.16	54	88.09
6.56	95.45	99.17	99.36	-	97.45	100	-	100	100	100	-	100	94.42	-	26.79	98.59
12.27	58.07	94.91	85.57	38.34	68.73	29.27	28.47	61.58	78.09	36.12	30.34	30.65	94.42	29.55	69.99	54.58
16.18	83.18	95.77	93.46	78.28	80.48	54.32	82.54	82.41	71.4	56.02	51.07	41.58	93.64	58.58	76.19	73.05
16.22	92.89	95.55	95.31	87.67	88.07	74.34	94.55	78.93	78.95	75.7	60.22	37.9	96.26	70.87	77.94	80.51
18.51	96.75	96.5	97.12	89.6	93.24	65.48	98.74	93.06	87.95	98.62	84.23	68.9	93.95	69.35	75.05	88.11
17.67	98.78	98.46	98.52	93.3	96.01	50	99.6	94.88	89.88	100	90.34	52.04	93.84	73.43	71.41	87.79
11.18	97.4	98.59	98.75	100	97.4	100	100	100	99.37	100	100	99.03	96.64	98.91	45.47	99.01
10.72	61.61	95.64	88.65	40.49	68.4	29.02	25.16	73.48	79.61	37.52	31.24	34.29	96.49	30.16	70.42	56.55
14.72	84.96	96.27	95.12	82.99	76.96	53.15	80.89	93.58	71.4	55.06	50.66	49.79	95.84	56.97	78.5	74.55
15.76	91.77	95.76	96.18	88.79	86.92	77.69	95.8	88.15	78.83	76.67	58.41	43.25	96.76	71.82	80.45	81.91
19.26	98.6	96.67	97.89	89.45	93.51	71.2	98.52	94.67	85.35	97.39	72.55	63.36	95.43	68.16	79.28	87.34
22.02	98	98.18	98.33	91.51	94.66	57.61	98.51	95.88	88.58	98.33	80.69	53.29	96.24	72.43	78.86	87.3
18.32	99.06	98.26	98.49	100	97.01	99.42	100	100	98.09	100	99.1	95.7	97.92	83.56	68.91	97.61
10.76	64.19	96.22	91.1	41.32	67.18	27.68	21.29	76.82	80.66	38.46	31.7	36.47	98.12	29.9	70.25	57.22
14.76	85.84	96.67	95.98	85.9	75.38	50.82	74.22	95.56	70.62	52.63	52.52	51.43	96.83	56.91	78.82	74.38
15.63	91.95	95.96	96.88	90.05	86.52	75.27	95.55	93.12	79.52	75.95	59.36	46.15	97.82	71.24	81.45	82.52
19.7	98.63	96.79	98.54	91.24	93.62	68.42	98.1	98.38	84.64	96.13	72.49	62.84	96.25	65.95	80.12	87.29
23.16	97.49	98.22	98.68	94.29	95.37	61.13	95.82	97.95	88.36	99	76.73	53.23	97.5	69.25	79.58	87.36
21.42	99.12	98.2	98.69	100	97.17	98.6	100	100	97.93	100	95.51	90.48	98.79	73.43	73.87	96.28
9.9	73.28	96.72	95.23	43.1	69.43	28.57	14.64	72.73	81.7	27.72	44.74	46.97	99.66	39.21	71.35	59.55
14.05	91.88	97.42	98.31	87.41	76.87	42.17	69.32	93.38	69.06	44.01	63.83	54.4	99.44	61.76	79.22	74.95
14.49	93.37	96.56	98.33	93.98	87.96	83.21	97.86	92.25	83.29	69.44	66.46	46.54	99.67	76.59	83.23	84.68
17.89	99.75	97.28	99.61	96.9	96.57	67.21	97.85	94.37	76.96	96.15	74.19	59.2	99.78	68.87	81.08	87.48
26.31	97.74	98.32	99.34	97.66	97.63	63.01	82.35	93.43	88.06	100	76.79	57.91	99.78	71.87	80.55	87.42
20.24	100	98	99.33	100	99.44	91.67	100	100	96.54	100	80.09	88.98	99.87	71.76	78.79	94.69

Table S6: Evaluation at class level of the relation of the fragment length used for model creation to the accuracy of assignments for genomic fragments from unknown organisms. For every clade, the clade-specific sensitivity (class *Sn.*), specificity (class *Sp.*) of assignments for genomic fragments of different lengths from the organisms of this clade among the 340 organisms is shown. The two rightmost columns display the overall sensitivity (*Sn.*) and specificity (*Sp.*) of de novo sequence classification (Supplementary Material, 'Evaluation procedures').

Eval With	Built On	Other	Gammapro	Sordariom	Actinobact	Thermopro	Thermopla	Thermococ	Clostridia	Deinococci	Bacteroide	Fusobacte	Spirochaet	Chlamydia	Methanomi	Alphaprote
		Class *Sn.* (%)														
1kbGF	1kbGF	37.55	59.56	59.33	69.88	40.33	59.33	88	37.11	48.33	73.33	93	9.33	84.78	40.67	60.22
1kbGF	3kbGF	82.1	23.24	30.33	44.5	54.67	23.33	75.33	19.33	23.67	16.75	81.67	10.5	2.11	23.67	48.85
1kbGF	5kbGF	92.9	26.67	0	17.58	18.67	29.33	5	3.78	26	48	27.67	0	0	2	7.3
1kbGF	10kbGF	98.29	0.64	0	0.42	36.33	8.33	9.33	0	9.67	0	55.33	1.5	0	1.67	0
1kbGF	15kbGF	99.36	0.17	0	0.08	8	0.33	0	0	0	3.67	0	2.5	0	0	0
1kbGF	50kbGF	99.88	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3kbGF	1kbGF	16.21	74.07	87	79.04	51.67	89.67	97	52.78	61.67	89.33	97.67	8.67	91.44	58	76.15
3kbGF	3kbGF	53.86	76.65	78.67	82.54	72	87.33	96	58.67	57	86.75	99.67	21.5	86.44	61	79
3kbGF	5kbGF	74.55	80.37	76.67	81.04	75.67	86	95.33	50.11	55.33	92.67	96.33	59.17	84.56	46.33	80.59
3kbGF	10kbGF	86.79	73.43	55.33	71.54	78.33	68.67	87	52.11	48	82.33	90.33	21.5	28.22	31	57.52
3kbGF	15kbGF	92.62	63.92	42.33	66.04	76	62	76.33	40.56	28.33	89	49.33	23.67	5.33	24.67	32.07
3kbGF	50kbGF	99.79	6.24	0	20.71	0	0	0	0	0	0	0	3.83	0	0	0
5kbGF	1kbGF	12.32	76.92	91.67	79.92	58	91.67	98	56	64.33	91.67	98.33	11.83	90.67	63.33	80.56
5kbGF	3kbGF	46.16	81.4	89.67	84.79	72	93.67	98	67.33	63.33	90.5	99.67	28.17	88.44	65.67	82.15
5kbGF	5kbGF	64.58	86.21	88.33	84.92	82	91.33	97.67	63.33	62.67	97.67	98.67	64.83	88.89	55	86.11
5kbGF	10kbGF	77.97	84.83	78	84.62	85.67	80.33	94.33	67.11	57.67	97.33	97	26.17	83.56	42.67	84.78
5kbGF	15kbGF	86.87	85.59	66	83.12	80.33	80.67	90.67	58.89	51	96.33	88.33	29.17	70.67	58.33	75.41
5kbGF	50kbGF	99.24	52.67	3.67	66.08	17.67	0	2.67	0	13	22.67	0	19.83	0	0	0
10kbGF	1kbGF	10.3	80.39	94.33	82.34	61.33	94	99	57.78	68.33	95	99.67	11.2	89.89	80.33	84.15
10kbGF	3kbGF	39.65	85.15	96	88.21	75	98.67	98.33	72.22	66	94.75	99.67	28.93	88.89	80.33	85.41
10kbGF	5kbGF	56.29	91.18	94.67	88.76	85	98	99	75	65.33	99	99.67	68.9	88.89	72	89.15
10kbGF	10kbGF	68.27	91.11	91.33	90.35	92.33	94.33	98.33	73.56	64.33	99.67	99.67	30.1	88.67	62.67	90.37
10kbGF	15kbGF	80.83	92.84	85.67	89.6	88	96.67	97.67	65.78	63.67	99.33	99.67	31.77	88.44	77.67	83.04
10kbGF	50kbGF	96.31	88.12	65.33	86.41	55	0	72.67	29.33	42.33	92	0.67	29.1	0	0	25.74
15kbGF	1kbGF	9.5	82.11	97	83.55	59	97.33	99.67	59.33	68.33	96.67	100	11.11	86.69	86.33	84.74
15kbGF	3kbGF	37.13	86.58	96.67	88.98	71.67	99.33	99.67	71.56	66.33	98.25	100	28.89	86.27	86	86.28
15kbGF	5kbGF	53.71	93.43	97	88.98	88.33	99.33	99.33	77.56	66	99.67	99.33	68.89	86.27	80.67	88.61
15kbGF	10kbGF	63.36	93.38	96.33	91.17	95	97.67	99.33	75.11	66	100	100	26.48	86.13	76.33	90.67
15kbGF	15kbGF	79.35	94.79	94	90.35	90.67	99	99.33	66.22	66	100	99.67	28.89	85.99	87.33	84.27
15kbGF	50kbGF	93.56	93.52	83.67	89.41	59	35.33	93.33	63.44	55.33	98.67	49.33	26.3	36.2	34.33	76.35
50kbGF	1kbGF	10.28	83.42	97	88.79	70.77	100	100	60.41	77.25	99.59	100	5.45	79.65	95.73	92.13
50kbGF	3kbGF	36.45	87.59	96.67	93.28	75.86	100	100	69.48	74.85	100	100	14.42	79.22	95.73	92.95
50kbGF	5kbGF	54.36	97.2	96.67	91.71	89.66	100	100	75.59	74.85	100	100	75	79.65	94.42	93.58
50kbGF	10kbGF	60.28	97.23	96.33	94.35	99.31	100	100	73.55	74.85	100	100	13.14	78.79	91.38	95.93
50kbGF	15kbGF	76.42	97.83	95.33	93.02	97.93	100	100	65.57	74.85	100	100	14.74	78.79	95.73	92.25
50kbGF	50kbGF	85.43	98.22	95	95.05	75.86	68.69	100	67.92	74.85	100	100	13.14	78.79	79.74	93.92

Supplementary Table S6

Betaprotec	Deltaprotec	Epsilonpro	Mollicutes	Mammalia	Insecta	Actinopter	Bacilli	Other Class *Sp.* (%)	Gammaprc	Sordariom	Actinobact	Thermopro	Thermopla	Thermoco	Clostridia	Deinococci	teroides(cl
76.12	23.33	87.11	80.56	82.67	43.67	46.67	59.81	29.03	83.75	25.32	72.44	42.46	58.75	53.55	48.2	38.87	43.39
71.08	18.5	86	67.22	85.33	51.33	43	22.93	17.86	95.82	59.87	94.01	68.33	97.22	74.34	92.06	94.67	93.06
33.38	9.83	86.33	67.56	77.67	38.67	65	0.52	15.11	94.56	-	98.37	93.33	91.67	100	97.14	86.67	80
1.12	0	70.11	51	55.33	12.33	56	0.04	13.58	100	-	100	73.15	100	100	-	93.55	-
0	0	62.67	29.78	56.67	5	24	0	13.32	100	-	100	75	100	-	-	-	100
0	0	8.89	26.67	0	0	0.67	0	13.02	-	-	-	-	-	-	-	-	-
85.25	26.17	88	86.22	93.33	62.33	77.67	74.67	44.34	87.77	20.83	69.49	54.96	70.05	52.62	52.03	39.61	44.22
86.33	38.67	95	80.89	96	64.67	76	78.7	39.79	88.62	51.87	83.98	85.38	92.58	61.94	63.16	86.36	80.51
79.62	39.83	94.89	85.67	95.67	67.33	88.33	75.17	39.72	91.51	73.95	90.42	83.46	91.49	86.93	85.74	86.91	80.58
74.5	33	80.67	84.56	91	61.67	88.67	60.69	28.47	96.7	82.18	94.7	90.04	96.26	93.21	95.52	87.8	94.27
57.21	17.67	90.78	78	89.33	44	70.67	48.7	23.23	99.15	94.07	98.63	92.68	100	92.71	95.05	98.84	95.7
0	0	76	72.56	0	0	43.67	0	13.89	100	-	99.8	-	-	-	-	-	-
88.08	24.5	89.44	86.89	96.33	60.33	87.67	76.04	47.47	89.51	20.88	68.65	63.04	75.14	55.79	55.02	42.98	42.77
89.5	42	96.67	85.67	94	68.33	87	84.87	46.44	88.72	51.43	81.37	91.14	95.25	59.76	63.39	82.25	74.18
83.88	42.33	96.67	89.56	94	72	94.33	86.33	48.61	91.31	64.63	87.69	88.17	91.64	76.9	78.84	84.3	75.71
84.67	38.83	81.33	90.22	93	66.33	92.67	83.89	42.29	95.17	76.22	92.61	95.19	89.93	86.54	91.93	86.93	82.95
78.79	27.83	94.44	83.89	90	57.67	82	81.02	39.03	98.75	88	95.27	95.63	98.78	81.93	83.73	99.35	93.53
0	6.5	79.44	80.44	4.67	15	71	0	16.65	99.77	100	98.75	100	-	100	-	100	100
91.04	24.17	88.22	87.78	97.67	72.67	93	79.43	56.16	90	22.57	67.95	72.73	84.18	59.64	57.71	42.01	43.18
93	46.83	99	88.04	96	77	93.33	89.11	52.39	89.05	51.25	78.73	97.4	98.34	58.88	64.68	85.71	72.47
88.88	45.67	98.33	92.32	97	75.67	98.33	91.5	58.63	91.66	58.44	84.37	92.39	92.16	73.7	77.41	83.05	75.77
90.62	42.17	83.22	93.32	96.67	81.33	97	91.02	52.62	94.44	74.86	88.97	96.52	94.97	79.3	91.82	84.65	73.65
88.21	35.33	96.44	85.89	95.33	73.33	91.33	91.39	51.81	98.75	88.62	93.68	97.42	99.66	77.72	82.45	96.46	89.76
28.33	30.17	81.89	84.13	79.67	55.67	89.33	33.94	23.81	99.17	99.49	95.15	100	-	100	100	100	97.18
92.33	26.83	87.67	90.64	99	73	96.67	82.29	61.4	90.08	24.11	66.78	68.6	89.02	58.4	59.27	44.86	44.62
94.46	49.33	99.11	90.47	97.67	74	96.33	90.71	54.1	89.55	52.73	76.97	97.73	99.33	55.06	64.92	86.15	74.01
90.58	47.67	98.33	94.65	99	77.33	99.33	92.88	62.18	91.82	56.29	82.71	94.98	91.41	71.98	78.78	84.62	79.31
91.92	45.17	83.44	93.65	99	83.33	99.33	92.8	55.75	94.46	71.89	87.55	98.28	91.85	73.4	91.47	84.26	69.93
89.92	39.17	97.89	89.46	99	72.33	95.33	92.97	56.24	99.04	90.38	92.96	98.19	99	75.44	83.83	95.19	90.36
74	32.83	83.44	86.29	92	68	94.67	82.12	40.46	98.53	95.44	92.76	100	100	100	99.65	100	94.57
96.23	30.82	86.36	91.01	100	74.67	96.33	85.49	69.19	91.63	35.66	71.23	65.25	89.19	61.49	68.32	39.94	46.85
96.99	60.3	100	91.01	99.67	80.33	96.33	93.43	57.81	92.92	60.04	78.94	94.02	100	51.2	69.48	91.24	72.23
95.16	47.75	99.36	96.63	99	79	99.33	94.61	73.44	93.53	60.8	86.38	93.53	96.84	72.79	87.5	80.13	80.65
96.03	48.11	82.68	96.07	99	88	99.33	96.47	66.1	95.51	70.15	91.3	94.74	90.83	69.48	94.38	84.46	70.59
93.79	48.76	99.35	90.45	99.67	79.33	96	94.9	66.57	99.59	91.96	94.59	95.95	97.03	77.54	91.09	85.62	90.91
93.48	42.44	84.31	89.33	99.33	86	96	93.4	60.53	97.2	79.83	95.71	100	100	100	99.77	100	86.71

Supplementary Table S6

ɔbacteria(cl	ochaetes(cl	amydiae(cl	thanomicro	aproteobac	proteobacta	proteobacɔn	proteoba	Mollicutes	Mammalia	Insecta	ıctinopteryg	Bacilli	*Sn.* (%)	*Sp.* (%)
48.35	25.57	62.7	15.95	69.55	67	16.2	87.6	60.87	57.14	17.15	35.9	83.25	59.16	50.63
75.38	52.5	100	63.96	87.35	75.22	44.94	87.56	81.98	50.39	13.87	41.75	95.97	43.72	74.56
93.26	-	-	100	98.99	94.12	64.84	85.38	83.75	71.69	36.36	62.5	100	29.73	85.93
78.3	100	-	71.43	-	100	-	86.44	96.43	80.98	55.22	63.16	100	20.32	87.42
-	93.75	-	-	-	-	-	96.25	99.63	57.63	50	56.69	-	12.71	84.45
-	-	-	-	-	-	-	95.24	97.17	0	-	16.67	-	5.92	52.27
53.66	29.05	34.54	32.22	79.17	72.45	23.9	98.02	67.95	68.97	37.7	62.8	85.82	70.61	56.27
68.89	53.31	84.75	35.6	87.89	79.11	44.36	95.32	88.78	61.41	34.46	65.9	92.88	74.49	72.14
84.26	95.43	96.7	67.15	94.28	90.78	50	94.36	90.92	75.13	39.45	76.15	96	76.58	82.8
87.7	80.12	100	57.41	99.42	95.51	65.56	95.9	95.6	76.04	56.06	82.1	98.56	65.52	87.3
100	89.87	100	100	99.88	97.58	94.64	98.43	97.64	69.07	49.62	77.37	98.8	55.16	92.72
-	100	-	-	-	-	-	99.85	99.09	0	-	97.76	-	14.03	85.21
56.08	38.38	30.11	38.93	80.89	73.92	24.26	99.02	68.66	73.16	42.29	74.5	87.66	72.8	59.17
69.7	62.59	76.17	32.56	88.86	80.9	43.98	97.21	92	70.85	40.43	72.1	92.62	78.22	73.07
82.45	86.25	91.53	59.14	93.67	93.07	47.92	96.56	92.86	76.42	42.27	79.05	95.75	81.36	80.73
90.94	81.35	98.95	52.67	97.99	95.35	55.61	98.52	94.97	76.65	54.97	84.5	97.52	77.09	85.34
98.15	88.83	99.69	94.59	99.56	97.07	71.98	99.88	96.55	76.06	50.44	79.61	98.2	73.78	90.25
-	100	-	-	-	-	95.12	100	98.77	21.88	100	99.53	-	24.11	93.84
61.02	38.95	28.75	47.16	84.52	74.85	29	99.75	74.87	75.32	51.05	79.71	90.28	75.74	62.51
73.46	68.38	76.34	32.66	90.15	83.41	48.45	98.56	94.2	75.59	48.23	78.65	93.44	81.72	75.36
85.19	85.48	86.39	56.69	92.93	95.14	51.6	98.77	94.09	79.08	42.83	83.33	95.83	85.15	80.74
94.62	84.11	95.91	46.77	94.83	96.15	55	99.47	95.49	74.94	52.81	88.45	97.44	83.06	84.33
97.39	90.48	98.03	80.34	97.14	96.23	61.27	99.77	96.74	78.57	49.89	84.57	97.3	82.52	88.74
100	96.67	-	-	100	100	86.19	100	98.82	89.18	93.3	98.17	99.95	50.7	97.54
64.52	42.25	24.93	53.07	84.24	74.97	33.97	99.87	74.76	75	54.89	86.05	92.03	76.95	63.92
78.12	70.91	73.7	33.86	90.02	84.46	48.52	99.66	94.25	75.32	48.9	85.25	94.37	82.42	76.08
86.63	86.92	82.93	56.67	91.45	96.19	50.8	98.88	94.02	79.2	46.87	88.96	96.79	86.39	81.28
95.85	81.71	93.95	48.11	93.65	97.18	56.46	99.73	94.59	73.51	54.11	94.6	97.26	84.59	83.81
97.39	91.76	90.78	76.38	96.03	96.08	62.83	100	97.1	77.34	50.23	92.56	97.23	84.43	88.64
100	95.95	100	100	100	99.78	78.49	100	99.04	93.56	91.07	99.3	99.71	69.61	97.17
62.81	34.69	14.63	62.27	86.59	73.25	47.28	100	77.51	75	62.57	96.01	93.13	79.19	66.15
81.17	76.27	63.32	45.9	92.84	84.77	59.41	100	97.01	75.51	55.4	96.66	93.04	84.11	78.7
86.81	91.41	72.73	61.72	93.15	97.8	58.38	100	95.03	76.55	51.08	95.82	96.34	88.41	83.13
98.43	91.11	87.08	60.23	93.52	98.07	60.26	100	96.07	74.62	56.41	99.67	96.85	86.12	85.17
99.21	100	54.17	77.52	95.85	96.19	73.29	100	97.58	75.13	55.22	98.63	94.69	86.29	88.26
100	95.35	100	99.46	99.76	99.46	72.4	100	99.38	83.94	83.5	100	99.06	83.08	95.07

Table S7: Evaluation at order level of the relation of the fragment length used for model creation to the accuracy of assignments for genomic fragments from unknown organisms. For every clade, the clade-specific sensitivity (class *Sn.*), specificity (class *Sp.*) of assignments for genomic fragments of different lengths from the organisms of this clade among the 340 organisms is shown. The two rightmost columns display the overall sensitivity (*Sn.*) and specificity (*Sp.*) of de novo sequence classification (Supplementary Material, 'Evaluation procedures').

Eval With	Built	Other	Chroococc	Nostocales	Prochlorale	Xanthomor	Vibrionales	Pasteurella	Spirochaet	Bacillales	Bacteroida	Clostridiale	Lactobacill	Fusobacter	Actinomyc	Rhodospiri
		Class *Sn.* (%)														
1kbGF	1kbGF	33.39	38.33	82.67	51.67	71.33	81.67	77.42	8.5	64.96	74.67	42.17	62.04	92.67	73.19	13.33
1kbGF	3kbGF	85.22	29.67	53	6	38.33	11.67	51.92	10.17	12.71	46	21.5	36.67	75.67	34.52	0.67
1kbGF	5kbGF	95.56	14.33	4.33	0	9.5	0.33	8.42	0	0.04	6.33	2.83	36.78	27.33	11	0
1kbGF	10kbGF	98.64	0	0	0	1	0	1.08	0.83	0	0	0.17	6.78	34	0.05	0
1kbGF	15kbGF	98.9	0.33	0	0	0	0	2.92	2.67	0	0	0	10.59	0	0	0
1kbGF	50kbGF	99.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3kbGF	1kbGF	14.22	53.33	98	63.67	80	97.67	91.25	6.17	78.42	92	48.33	71.26	95.33	82.67	8
3kbGF	3kbGF	53.78	79.67	93.67	65	85.67	92.33	91.33	24.33	80.17	88.75	57.5	80.07	97.67	86.86	7
3kbGF	5kbGF	77.38	79.67	92.67	60	81.17	74	90.17	58.17	79	94.33	54.83	77.7	96	83.43	6.67
3kbGF	10kbGF	94.36	44.33	79	30.33	77.83	58	86	20.33	44.62	82.67	55.33	76.89	89.33	74.19	27
3kbGF	15kbGF	96.14	18	25	0	58.17	23	70.92	20.33	11.58	61.33	43	77.63	0	53.29	3.33
3kbGF	50kbGF	00 39.67	0	14.67	0	2.5	0	0	3.5	0	0	0	35.44	0	11.76	0
5kbGF	1kbGF	.46 88.36	59	100	61.33	82.5	99	94.67	9	81.17	94.33	51.5	76.93	97.67	84.86	6.67
5kbGF	3kbGF	81.37 1(	88	99	68.67	87.83	98.33	95.17	30.33	85.83	94	61.17	84	99	88.95	9.67
5kbGF	5kbGF	63.24	89.33	98	67.67	84.5	85	96.17	64.83	88.83	97.67	62.83	84	97.67	87.67	7.67
5kbGF	10kbGF	86.19	60.33	97.33	61.67	90.17	90.33	96.08	26	79.38	97.67	61.67	85.63	96	91.14	31.67
5kbGF	15kbGF	91.4	29	88	0	88.17	82.33	87.67	26.83	55.96	94.67	57.5	87.15	3.67	79.9	11.67
5kbGF	50kbGF	99.54	31	62.33	3.33	34.33	10.67	20.58	19.83	0	24.33	0	73.93	0	60.81	6.67
10kbGF	1kbGF	11.98	63	100	60	82.17	99.33	96.42	10.87	85.12	96.67	53.5	80.33	99	86.8	4.33
10kbGF	3kbGF	33.39	95.67	99.67	69.33	89.83	99	97.5	27.26	89.71	96.75	63.5	88.26	99.67	92.32	13.67
10kbGF	5kbGF	52.46	96	100	69.67	88	93.33	97.75	67.56	94.17	99.33	69.5	87.78	99.67	91.31	6.67
10kbGF	10kbGF	76.33	65.33	99.67	67.33	93.67	97.67	98.67	28.93	89.38	99.67	64.83	90.48	99	97.31	32.67
10kbGF	15kbGF	83.84	35	98	62.67	95.33	97	95.08	30.27	78	100	64	92.07	95.67	86.32	24.67
10kbGF	50kbGF	97.94	60.67	96	34.67	82.5	74.67	83.25	29.26	10.12	91.67	20	88.37	1.33	90.98	29.67
15kbGF	1kbGF	11.72	63.33	100	61.33	80.67	100	97.92	10	86.33	96.33	56.67	82.54	100	86.75	3.67
15kbGF	3kbGF	30.34	96.67	100	68.33	90.83	100	98.92	26.67	90.42	98.5	65.83	89.56	100	92.09	16.67
15kbGF	5kbGF	49.11	97.67	100	69	88	95.67	99.25	68.89	95.38	99.67	69.67	88.97	100	91.79	6.67
15kbGF	10kbGF	72.12	66	100	67.33	95	99.33	99.58	26.48	91.08	100	65.5	91.49	99.67	97.53	33.33
15kbGF	15kbGF	80.31	36	100	66	97.5	99.67	98.25	27.59	79.88	100	66.33	93.61	99.67	85.61	31.33
15kbGF	50kbGF	96.35	65.67	100	49.67	89.83	91	91.5	26.48	71.21	98.67	59.17	92.57	55.67	94.96	31.67
50kbGF	1kbGF	12	50.49	100	55.17	72.57	100	98.97	4.81	88.55	99.59	59.2	83.01	100	91.06	3.25
50kbGF	3kbGF	30.1	99.46	100	58.62	94.07	100	99.79	14.74	93.04	100	65.55	92.83	100	95.22	20.73
50kbGF	5kbGF	50.79	99.46	100	59.48	91.11	97.67	100	73.72	97.31	100	69.58	88.21	100	94.78	21.14
50kbGF	10kbGF	70.12	56.65	100	58.62	95.81	99.64	100	13.14	92.52	100	61.28	92.83	100	99.13	37.8
50kbGF	15kbGF	75.53	34.96	100	58.62	98.77	100	100	14.42	81.68	100	61.93	96.95	100	89.25	37.8
50kbGF	50kbGF	92.21	56.65	100	58.62	95.14	99.64	99.59	13.14	89.78	100	61.5	95.52	100	99.57	37.4

Supplementary Table S7

Rhodobact	Neisseriale	Mycoplasm	Campyloba	Thermococ	Thermopla	Rhizobiales	Chlamydial	Desulfurom	Diptera	Pseudomo	Rickettsiale	Burkholder	Enterobact	Methanosa	Other	Chroococc
															Class *Sp.* (%)	
33.33	77	67.17	88.78	91	66	56.89	84.78	62.33	65	48.56	84.11	73.07	66.75	43.67	49.75	32.95
29.67	75.67	46.33	87.44	71.67	51	3.22	1	34.67	70.67	16.56	25.44	57.4	15.08	25.33	27.95	71.77
24.67	68.83	39.83	79.22	5	32.33	0	0	27.67	37.33	2.22	0.44	16.8	0.75	2	24.7	72.88
0	0.5	8.17	62	2	1.33	0	0	1	0	0	0	0	0.17	0	22.9	-
0	0	0.5	48.78	0	0.33	0	0	0	0	0.89	0	0	0	0	22.76	100
0	16.67	0	28.22	0	0	0	0	0	0	0	0	0	0	0	22.68	-
46	78.17	74	90.22	98	89.33	73.78	89.33	77	78.33	62.22	91.22	82.4	81.39	56	66.97	43.13
54.33	80.67	60.5	96.22	97.67	90.33	71.44	87.22	66.67	77	63.78	89.67	85.27	88.11	62.33	62.38	63.56
59	80.83	62.83	97.11	96	87.33	81	84	65.67	71.33	63.67	84.33	85.87	82.44	42	57.54	62.89
50	76.83	50.5	80.33	83	64.67	56	31.89	48	44	43.33	45.78	67.8	69.36	22	42.82	82.1
38.33	74.5	47.33	87	24	60	0	5.89	32	39.67	43.67	4.56	39.27	41.14	24.67	32.46	50.94
0.67	65.67	2.83	73.56	0	0	0	0	0	0	0.11	0	0	1.69	0	24.26	-
49.33	79.33	77	90.22	99	92.67	75.44	89.11	81.33	75.67	65.33	92.67	85.87	84.06	62.33	73.32	45.62
58.33	82.33	65.33	96.78	99	95.67	74.11	88.67	72	81.67	70.33	91.22	90.47	90	67.67	69.44	64.39
68	82.17	66.83	97.89	97.67	92.67	94	88.67	73.33	72	73.22	92.56	91.33	88.94	53.67	67.83	63.81
60	80.5	57.33	80.56	93	80	90.78	84.67	65.67	58.67	62.56	84.44	85.8	89.75	42	61.36	71.26
53.67	79.5	59	92.78	83.33	78	5.89	69.44	57.67	56.33	60.44	64.89	73.73	85.59	56	47.82	47.8
35	74	23	78.11	1.33	17	0	0.22	2.67	7.67	19	0.56	0	37.67	0	28.94	100
55	82	79.77	88.78	99.33	92.33	77.11	89.22	91	85.33	70.22	95.56	88.73	87.35	80.33	76.33	51.92
63.67	83.33	72.78	98.78	99.33	97.33	76.89	88.78	78.33	85.67	73	93.67	93.33	92.19	82.33	76.16	65.23
74.33	83.17	73.16	99	99.33	97	97.22	88.89	77.67	76	75.11	96.67	93.8	91.98	73.33	76.85	66.21
63.33	82.67	65.78	82.22	97	92.67	98.22	88.67	74.67	74.67	71.56	91.33	92.6	94.46	67	73.13	64.9
61.33	83.17	68.62	97.78	95.33	94	95.89	88.56	70	69.67	70.44	88.67	87.4	94.64	80	65.42	47.51
58.67	79.5	47.45	81.11	68	52.33	5.44	59.44	46.33	39	61.78	61.78	12.6	87.79	0	41.19	94.79
53.67	82.5	81.23	90.11	99.67	97.33	77.44	86.27	92	82.33	72.78	96.03	90.27	89.54	86.67	78.01	53.98
62.33	83.33	76.35	99.22	100	99	77.44	86.27	81	85.33	74.78	93.38	93.53	93.4	88.33	78.02	65.61
75	83.33	76.35	99.67	100	99	96.78	86.27	79.67	75.33	76	98.28	93.73	93.07	84	80.07	69.93
64.33	83	70.44	82	99	97	98.78	86.13	79	74.67	72.89	92.45	93.93	95.43	78	76.73	62.66
60.67	83.33	72.24	99	98.67	98.67	98.78	86.27	72.67	71.33	74.33	89.14	90.6	95.68	86.67	68.42	45.96
59.33	80.33	49.61	83.33	88.67	66.33	68.22	79.2	60	52	70.78	83.44	66.27	93.28	33.33	57.49	87.95
53.76	69.65	83.62	88.64	100	97.98	90.13	79.22	92.69	77.67	84.45	98.23	92.97	92.9	96.09	83.76	52.26
63.91	69.65	77.59	100	100	100	89.44	79.22	85.77	82.67	84.45	94.69	94.14	94.72	96.44	81.58	73.2
80.45	69.65	78.45	100	100	100	98.96	79.65	86.92	65.67	84.93	99.56	93.72	94.62	94.82	85.27	76.25
65.04	69.65	71.55	80.72	100	100	99.86	78.79	84.23	78.67	83.35	96.9	94.97	97.13	94.4	82.58	67.69
64.66	69.65	74.14	100	100	100	99.87	79.22	79.62	75	85.05	96.02	92.83	97	95.73	72.63	47.02
63.53	69.65	56.9	83.33	99.07	90.91	99.86	78.79	68.46	67	84.33	88.94	94.48	97.07	77.59	73.61	81.48

Supplementary Table S7

Nostocales	Prochlorale	Xanthomor	Vibrionales	Pasteurella	Spirochaet	Bacillales	Bacteroidal	Clostridiale	Lactobacill	Fusobacter	Actinomyc	Rhodospiri	Rhodobact	Neisseriale	Mycoplasm	Campyloba
49.11	28.6	46.12	27.22	73.32	29.14	77.68	44.8	41.89	81.75	66.03	73.61	9.35	34.01	78.71	45.28	90.38
78.71	90	81.27	85.37	96.74	49.19	93.85	80.7	86.58	92.96	79.93	96.41	5.41	60.96	75.29	48.6	87.06
92.86	-	95	100	100	-	100	90.48	100	91.61	92.13	99.14	-	50.68	86.04	67.9	97.94
-	-	100	-	100	100	-	-	100	98.39	96.23	100	-	-	100	94.23	91.63
-	-	-	-	100	94.12	-	-	-	72.59	-	-	-	-	-	100	98.21
-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	-	98.83
47.65	25.3	58.11	22.35	82.45	36.27	80.22	48.17	45.1	88.01	69.76	69.94	5.87	53.91	93.99	50.86	98.9
78.06	54.78	66.58	43.9	87.26	54.89	86.59	78.02	47.85	90.8	79.19	82.65	6.54	59.49	86.27	55.76	97.09
85.8	85.31	78.8	58.89	93.84	97.21	92.35	86.02	70.45	94.42	86.23	88.71	20.83	74.06	92.03	59.37	97.65
89.43	100	84.91	89.23	98.66	81.88	98.17	93.58	90.22	95.27	97.81	96.05	67.5	92.02	99.57	85.11	97.18
100	-	96.41	95.83	98.84	82.99	99.29	99.46	88.36	91.37	-	99.29	14.71	98.29	99.33	84.27	98.37
97.78	-	100	-	-	100	-	-	-	99.17	-	99.6	-	100	99.75	100	100
50.76	25.24	63.79	22.57	85.67	50.47	82.51	46.24	46.75	90.82	70.43	70.8	5.13	59.68	97.14	51.74	99.75
75.57	48.58	71.22	37.15	85.86	65.7	85.44	74.31	43.85	90.72	82.73	78.59	7.69	63.87	88.37	55.21	98.42
83.05	64.44	76.01	48.66	91.95	87.61	89.73	80.72	56.44	94.82	82.77	83.42	11.98	77.57	94.44	56.96	98.77
78.49	97.88	81.97	67.75	97.38	79.19	96.36	83.24	87.89	95.07	97.3	93.59	42.04	88.67	99.59	77.48	98.37
99.25	-	94.63	90.48	96.87	88.95	98.82	97.26	79.31	93.11	100	96.66	20.23	99.38	98.55	73.75	99.29
97.91	100	98.56	100	100	100	-	100	-	98.37	-	98.38	100	100	100	97.18	100
52.08	27.03	66.98	23.24	86.99	59.09	85.27	48.17	55.25	95.22	77.34	71.56	3.64	61.34	99.19	53.62	99.5
74.01	46.95	77.11	33.79	84.17	65.73	86.4	73.85	41.78	91.48	86.17	74.81	9.47	68.46	93.46	57.81	99.55
78.12	58.38	77.42	41	89.95	86.88	88.63	80.98	49.7	94.99	87.43	79.26	7.35	76.37	97.08	58.11	99.44
67.49	93.09	83.76	52.14	96.42	79	95.04	72.05	87.02	96.75	100	89.93	32.13	83.33	99.4	72.96	99.73
95.15	100	89.38	74.05	95.24	89.6	97.75	91.19	73.7	95.32	99.31	92.92	22.63	99.46	99.6	67.47	99.77
88.62	100	92.7	94.12	100	97.77	100	97.17	100	98.64	100	96.39	95.7	100	100	95.8	100
53.1	28.84	69.34	23.75	85.77	58.7	86.8	48.98	62.5	96.78	82.87	70.33	3.18	60.3	99.6	52.84	99.88
74.07	48.69	81.46	33.19	83.71	66.98	86.28	74.62	43.41	93.23	92.31	72.62	11.04	63.82	94.52	56.46	100
77.52	54.33	81.99	38.84	89.08	85.52	88.55	84.7	49.76	96.11	86.96	76.82	6.97	74.01	98.23	57.12	99.89
65.08	90.99	86.89	46.27	96.53	77.3	94.39	66.96	85.43	97.16	100	87.76	30.49	83.91	99.4	69.9	99.86
95.54	100	89.86	65.14	92.04	89.76	97.86	91.46	75.52	96.89	99.01	90.96	25.2	98.91	100	64.01	99.89
85.96	100	91.82	87.22	99.91	97.28	100	93.08	100	98.38	100	95.05	82.61	98.89	100	94.61	100
58.59	24.52	78.28	31.01	83.68	55.56	88.06	51.36	73.68	97.85	85.03	74.64	2.96	51.62	100	55.11	100
75.19	35.79	92.93	39.49	78.01	75.41	88.59	75.55	52.16	88.69	96.15	76.23	12.38	72.96	87.2	57.69	100
76.92	37.1	89.35	42.96	87.28	92	86.46	92.88	58.14	96.87	89.93	80.74	19.33	79.26	94.78	57.96	100
60.85	73.12	93.74	46.19	93.12	85.42	93.27	67.87	89.94	98.57	100	91.25	29.71	82.38	100	68.03	100
91.74	76.4	84.21	60.74	87.28	100	96.27	91.46	81.79	97.91	99.21	92.35	27.68	100	99.54	62.77	100
79.16	100	95.99	74.4	99.39	95.35	98.44	86.46	98.93	95.26	100	96.28	53.8	98.83	100	86.84	100

Supplementary Table S7

Thermococ	Thermopla	Rhizobiales	Chlamydial	Desulfuron	Diptera	Pseudomo	Rickettsiale	Burkholder	Enterobact	Methanosarcinales	*Sn.* (%)	*Sp.* (%)
47.4	59.28	55.96	69.55	27.66	19.27	43.31	64.43	73.26	88.51	15.5	62.55	51.52
73.88	87.43	96.67	100	86.67	20.25	77.6	91.24	89.59	99.27	61.79	37.5	77.42
100	89.81	-	-	76.15	34.89	100	100	98.82	100	100	18.46	89.01
100	100	-	-	75	0	-	-	-	100	-	7.26	90.36
-	100	-	-	-	-	100	-	-	-	-	5.53	95.61
-	-	-	-	-	-	-	-	-	-	-	4.83	99.42
45.51	65.69	48.72	49.17	26.01	28.28	47.18	67.24	75.83	92.05	21.79	71.59	54.74
62.74	89.44	74.85	87.81	65.79	28.24	60.55	81.52	86.48	95.08	35.76	75.17	68.54
88.62	84.24	92.87	98.31	67.24	36.83	79.03	94.29	90.83	97.95	69.23	74.95	80.15
93.61	95.1	95.82	100	75.39	69.47	92.2	100	97.32	99.32	92.96	59.12	91.38
100	100	-	100	98.97	62.96	81.37	100	99.16	99.87	100	37.46	90
-	-	-	-	-	-	100	-	-	100	-	10.41	99.66
46.19	66.67	47.15	45.99	26.81	29.48	48.43	72.46	76.85	93.17	26.3	73.69	56.71
58.58	92.28	71.11	81.35	58.38	31.45	58.94	81.45	84.55	94.82	33.61	78.55	67.73
76.3	88.82	83.27	94.66	65.87	35.47	70.71	92.25	88.44	96.16	61.45	80.27	75.74
87.46	93.02	89.09	99.22	69.37	64.47	81.83	98.57	94.63	97.82	73.26	75.57	85.6
99.21	99.57	100	100	95.05	59.72	78.16	100	98.22	99.65	93.85	62.01	89.21
100	100	-	100	100	100	99.42	100	-	100	-	24.79	99.56
45.5	70.13	48.19	44.07	29.61	31.8	50.12	77.2	77.07	93.67	33.29	76.39	59.24
52.84	93.89	68.31	80.38	54.78	34.27	56.54	81.84	84.64	95.92	32.85	81.16	67.81
69.14	91.22	77.64	93.68	68.53	36.83	66.73	90.34	86.69	95.82	61.28	83.66	74.32
76.38	92.98	81.47	96.38	62.4	59.73	78.06	94.16	90.96	97	57.59	81.26	81.11
90.79	99.3	96.75	98.15	90.13	53.18	75.84	99.13	97.18	99.28	86.02	79.45	86.75
100	100	100	100	98.58	99.15	96.7	100	100	99.97	-	55.08	98.07
42.53	72.82	47.45	38.56	31.4	30.95	51.13	77.96	78.36	94.19	35.96	77.17	59.96
47.85	95.19	68	77.56	55.35	34.83	55.8	80.11	83.81	95.81	34.87	81.95	67.97
63.97	93.99	77.49	90.94	74.22	36.99	63.51	87.91	86.52	95.32	59.29	84.54	74.02
69.72	92.38	78.46	92.96	66.95	58.64	76.28	91.6	90.44	96.73	57.64	82.38	79.75
80.22	99	93.78	94.24	91.21	52.07	74.75	97.82	97.14	99.32	83.07	81.33	85.54
100	100	100	100	95.74	93.41	92.45	100	100	99.61	100	71.62	96.34
31.2	52.15	53.34	20.96	37.31	33.29	56	76.82	83.06	95.3	45.15	77.22	60.3
38.08	96.84	67.31	70.11	60.76	37.46	59.25	76.7	88.75	97.2	42.34	82.56	69.39
59.12	97.87	80.47	90.2	89.33	36.82	62.02	84.59	91.33	96.55	70.83	85.69	76.46
60.8	96.12	78.4	90.55	73.99	61.78	78.31	84.56	90.77	98.08	65.96	82.43	80.02
72.3	96.08	84.83	59.42	94.95	55.69	76.59	89.67	98.25	99.73	86.5	81.96	83.12
99.07	100	98.03	100	97.27	89.33	89.09	100	98.7	99.15	99.45	80.62	93.47

Table S8: Evaluation at genus level of the relation of the fragment length used for model creation to the accuracy of assignments for genomic fragments from unknown organisms. For every clade, the clade-specific sensitivity (class *Sn*.), specificity (class *Sp.*) of assignments for genomic fragments of different lengths from the organisms of this clade among the 340 organisms is shown. The two rightmost columns display the overall sensitivity (Sn.) and specificity (Sp.) of de novo sequence classification (Supplementary Material, 'Evaluation procedures').

Eval With	Built	Other	Prochlorococc	Staphylococci	Streptococcus	Bacillus	Clostridium	Lactobacillus	Listeria	Corynebacteri	Mycobacteriur	Streptomyces	Campylobacte	Helicobacter	Mycoplasma	Methanosarcii
		Class *Sn.* (%)														
1kbGF	1kbGF	32.6	52.33	88.5	87.87	82.22	50.33	32.67	86.33	58	90	96	93.67	85.67	82.33	50.67
1kbGF	3kbGF	91.26	7.67	26.5	33.07	11.11	41.17	23.33	23.33	1.33	20.67	48	81	66	53	30
1kbGF	5kbGF	97.09	0	1.33	23.13	0.56	10.33	32.33	0.67	0	4.67	4.67	24	52.67	62	0.33
1kbGF	10kbGF	99.73	0	0.17	1.73	0	4.5	19	0	0	0	0	0	55.67	10	2
1kbGF	15kbGF	99.05	0	0	0.4	0	0.17	10.33	0	0	0	0	0	15.67	10.67	0
1kbGF	50kbGF	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3kbGF	1kbGF	8.23	69	96.83	93	84.67	65.5	40.33	96.33	82.33	98	99.33	96.67	91.33	87	64.67
3kbGF	3kbGF	56.85	67	95.33	94.07	86.56	71.83	40.33	96.67	73	91.33	97	98.33	76.33	63.33	65.33
3kbGF	5kbGF	81.75	58.33	93.5	94	81.78	71.33	62.33	95.67	52	83.83	96	90	71	87.33	62.67
3kbGF	10kbGF	95.11	28	77.67	86.6	53.22	73.5	62.33	80.67	20.33	72.33	7	92	78.33	56	50.67
3kbGF	15kbGF	96.56	1	75.5	84.6	17.67	63.33	30.67	40.33	4	36.33	57.33	64.67	38.67	39	17.67
3kbGF	50kbGF	100	0	0	0.53	0	2.17	0	0	0	0	0	0	0	9.67	0
5kbGF	1kbGF	5.19	68	95.83	95.53	82	70.5	36.33	98.33	86.67	99.5	100	99	93.33	91	67.33
5kbGF	3kbGF	46.26	69	97.5	96.33	87.44	76.17	41	99	83.33	94.67	99.67	99.67	79.67	64.67	66.67
5kbGF	5kbGF	70.22	67	97.67	97.27	86.67	74.83	65	99	68.67	92.33	99	95.67	68.67	93.33	65.33
5kbGF	10kbGF	88.04	61.67	95	95.33	81.78	79.33	67	98.67	69.67	87.5	87.67	99.33	78	63.67	61.33
5kbGF	15kbGF	93.32	35.67	94	95.13	66.78	76	33.33	88.67	44.33	69.67	85.67	90	48	48	52.33
5kbGF	50kbGF	99.94	0	10.67	16.8	5	21	0	8.67	0	7.17	0	0	1.33	25	0
10kbGF	1kbGF	3.58	67.33	97.17	97.53	83.56	77.83	37.67	99	84.67	99.67	100	99.67	95.67	92.91	82.33
10kbGF	3kbGF	40.26	70.67	99.33	98	89.44	80.5	41.67	99.67	87.33	97.67	100	100	80	70.82	81
10kbGF	5kbGF	59.93	71	99.17	97.93	89.33	78.33	68	100	76.33	97.83	100	95.33	67.33	97.8	81.67
10kbGF	10kbGF	79.76	69	99.17	97.8	88.56	82	73.67	99	91.67	97	100	99.67	79	70.11	80.33
10kbGF	15kbGF	87.55	64	99	98.47	85.67	80.83	36.33	98.33	71.33	89.83	99.67	95	53.33	56.69	78.33
10kbGF	50kbGF	99.44	3.33	58.67	48.53	47.89	49.67	19	43	1	64.83	21.67	3	31	33.45	0
15kbGF	1kbGF	3.26	67	96.83	96.6	85	78.33	37.54	98.33	87.33	100	100	100	95.33	94.02	87.33
15kbGF	3kbGF	38.5	70.67	98.83	98.2	89.78	82	43.69	99.67	87.33	99	99	99.67	81.33	75.89	87
15kbGF	5kbGF	56.09	71.33	98.83	99	90	79.67	70.99	100	78.67	98.67	99.33	95.33	67.67	99.02	88.67
15kbGF	10kbGF	76.97	69.67	99.5	98.27	89.89	83.17	73	99.33	95.67	98.83	99.67	100	77.67	75	87
15kbGF	15kbGF	84.95	66.67	99.67	99.13	87.33	82.17	38.91	98.33	81.67	92.67	99	98.67	56	60.95	86
15kbGF	50kbGF	98.76	26.67	90.67	84.86	72.22	74.5	45	85	12	80	81.33	61	63.67	45.54	46.33
50kbGF	1kbGF	2.49	58.62	100	99.47	61.18	81.8	32.58	100	89.83	100	100	100	98.96	98.57	96.44
50kbGF	3kbGF	37.96	58.62	100	99.3	90.44	82.26	35.61	100	93.79	100	100	100	81.82	76.12	95.73
50kbGF	5kbGF	53.82	59.48	100	100	90.67	82.03	75.47	100	81.92	99.81	100	96.15	65	100	96.09
50kbGF	10kbGF	73.16	58.62	100	100	91.24	83.18	68.28	100	99.44	100	100	100	73	76.12	96.09
50kbGF	15kbGF	82.34	58.62	100	99.65	87.9	82.95	33.83	99.44	98.31	96.93	99.67	100	60.2	68	95.73
50kbGF	50kbGF	94.88	34.48	100	98.97	82.49	82.72	52.41	99.44	71.19	85.48	99.67	100	68	62.69	95.37

Supplementary Table S8

Pyrococcus	Brucella	Xylella	Pseudomonas	Burkholderia	Buchnera	Neisseria	Bordetella	Escherichia	Salmonella	Shigella	Yersinia	Vibrio	Haemophilus	Rickettsia	Chlamydophila	Fusobacterium	Other Class *Sp.* (%
91	92.67	82.67	75.67	85.67	93.67	93.67	83.67	37.83	76.67	24.33	79.83	78.67	91	94	93.17	92.67	86.71
79	66.67	36	13	75.67	24.33	92.33	46.67	3.17	31	9.33	0	61.33	54.33	16.67	0.17	72	63.59
42.67	23.33	9.67	0.5	3.83	0	82.67	18.33	0.33	17.67	0	0	17.67	3.67	0.67	0	31	58.33
13.67	0	0	0	1.17	0	63.67	2	0	14.5	0	0	0	0	0	0.33	0	56.68
1	0.33	1	0	0	0	17.33	0	0	0	0	0	0	0	1.33	0	0	55.29
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55.52
98.33	97.67	95.33	89.83	94	99.67	95.33	87	29	76.17	53	92.83	96	94	99.33	99.83	93.67	91.3
97	97.67	95	89	93.83	95	95.67	98.67	46.17	92.17	34.67	88.83	93.67	90	98	98.33	98	91.57
96.67	96.67	95.33	56.83	90.17	51.67	99.33	83.67	54.5	92	33.33	83.33	91.33	92	97	96.17	94.67	88.28
87.67	83	85.33	49.5	85.17	36.33	95.67	68.33	35.17	81.5	12.67	14	72	67	66.67	66.83	78	77.72
64	69	65.33	40.83	52.5	11	91.67	41	33.17	61.83	5.67	2.29	33	0	65.33	60.83	60.33	69.16
0	0	0	0	2.5	0	3.33	2	0	0.5	0	0	0	0	0	0	0	55.67
99	98.67	97.33	92.33	95.83	99.67	97.67	86	27	76.33	59.67	95.33	96	96.33	99	99.67	96.67	93.28
99	98.67	98.67	94.17	97	99.33	98	99.33	51.17	92.83	45.33	95.17	97.33	95	100	99.33	99.33	93.37
98.33	98.33	99.33	73	96	88.33	99	82.67	54	94.5	48.33	98.5	95.67	97.67	99.33	99.83	98	92.38
94.67	97.33	98.33	89.5	93.83	88	97.67	77	76.5	91.5	28	78.6	92.33	95.67	97.67	91.33	98	90.63
87	93.67	92.33	77	87.17	65.67	97	88	63.83	86	26.67	59.71	80.33	0.33	94.33	91.5	90.33	83.21
0	5.67	1	2	16.17	0.33	22.33	18	0	15.5	0.33	0	0.67	0	7.67	12.17	0	57.51
99.33	100	98.33	96.17	98.5	100	99	81.67	24.83	83.33	68.67	97.17	98.33	98.67	100	99.67	99	95.86
99.33	99.33	99.67	97.17	98.67	100	99.67	99	45	95.83	64.33	97.83	98.67	95.33	100	99.5	99.67	95.01
99.33	100	100	84	98.67	99.47	100	81	38.33	96.67	68	99.67	97.67	99.67	100	99.5	99.33	94.61
98	99	100	96.5	98.83	98.95	100	80.67	81.5	97.17	44	96.6	97.33	99.67	100	98.5	99.67	94.93
96.33	98.33	99.33	94.83	98.33	94.21	100	97.33	76.83	96.67	52	96.86	97	86.67	100	99.67	99.67	92.69
29.33	50.33	33	40.83	65	24.74	58.33	48.67	0	62.67	7.33	0	35.33	38.33	52.67	35.33	0.67	66.14
100	100	100	97.5	98.33	100	99.67	85.67	19.5	84.83	74.33	98.17	99.67	99.67	100	99.79	99.67	94.86
99.67	99.67	100	98.67	98.17	100	100	99.67	39	96	72	98.5	99	97.33	100	99.79	100	95.55
99.67	100	100	87.83	98	100	100	82.67	23.5	96.83	81.33	99.83	99	100	100	99.79	100	95.22
99	100	100	98.67	98.17	100	100	84.33	80.83	97.17	56.67	98.4	99	99.67	100	99.38	100	95.88
98.67	99.33	100	97.83	98.33	98.41	100	100	74.17	97.33	65.67	98.29	99	92	100	99.79	99.67	93.96
56.67	90.67	80	76	87.67	53.17	88	62	18.67	92.5	27.67	33.8	83.67	84.33	92.21	58.54	55	79.53
100	100	98.72	100	99.67	100	100	91.45	34.8	82.6	78.21	98.2	100	99.09	100	100	100	95.97
100	100	98.74	100	99.5	100	100	100	46.45	96.23	72.86	98.74	100	100	100	100	100	96.93
100	100	99.36	95.83	99.83	100	100	90.33	22.47	98.28	92.71	99.1	100	100	100	100	100	97.1
100	100	99.36	100	99.67	100	100	94.4	71.28	97.3	81.63	97.5	99.64	100	100	99.3	100	97.51
100	100	98.72	99.83	99.83	100	100	100	68.92	97.62	82.99	98.62	100	97.14	100	100	100	96.09
97.2	100	99.36	99.33	99	78.38	100	71.2	94.93	98.56	69.26	97.92	98.57	99.07	100	98.59	100	94

Supplementary Table S8

rochlorococcu	taphylococcu	Streptococcus	Bacillus	Clostridium	Lactobacillus	Listeria	corynebacteriu	Mycobacterium	Streptomyces	Campylobacte	Helicobacter	Mycoplasma	Methanosarcin	Pyrococcus	Brucella	Xylella	Pseudomonas
s)																	
21.02	69.32	73.14	63.68	47.63	17.72	38.6	30	54.27	30.13	87.81	70.8	37.31	8.45	41.3	43.44	39.24	30.72
95.83	95.78	93.41	96.15	76.95	20.83	89.74	100	91.85	81.82	97.98	80.16	36.98	50.56	63.71	89.29	87.1	82.98
-	100	95.07	100	93.94	23.77	100	-	90.32	100	100	93.49	44.82	100	95.52	97.22	90.62	100
-	100	100	-	100	67.86	-	-	-	-	-	84.77	96.77	100	100	-	-	-
-	-	100	-	100	15.58	-	-	-	-	-	97.92	65.31	-	100	100	100	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
20.58	79.05	74.4	59.76	59.1	30.1	28.73	22.68	44.11	24.57	97.64	76.11	49.15	8.9	41.09	28.61	48.89	28.72
42.95	87.6	81.33	70.63	63.95	33.8	63.74	66.57	71.63	43.43	82.63	83.88	35.38	19.56	53.99	60.04	61.82	44.76
83.73	93.66	87.09	84.31	81.99	44.52	78.85	80	81.13	59.63	95.07	88.38	50.68	70.15	77.33	79.02	73.33	61.89
100	98.73	92.92	96.18	86.64	66.55	97.19	100	94.76	100	97.18	83.04	77.78	87.36	90.69	99.6	86.78	93.4
100	98.48	95.34	100	68.72	41.07	100	100	93.97	100	100	58.59	59.39	100	97.96	100	90.74	94.96
-	-	100	-	100	-	-	-	-	-	-	-	100	-	-	-	-	-
18.77	86.34	75.14	60.29	65.08	28.68	28.67	21.87	43.32	23.6	99	77.13	53.85	9.18	40.91	26.4	54.58	29.69
32.6	89.72	80.91	66.64	64.01	33.24	58.01	55.31	66.51	37.19	85.67	86.59	35.02	15.67	54	53.82	62.58	39.73
53.17	92.72	85.12	75.58	77.82	47.79	68.12	70.55	75.89	47.14	93.79	91.56	51.66	47.8	70.07	67.05	69.79	52.52
95.36	96.77	89.77	90.86	81.93	65.9	84.09	96.31	85.78	85.67	96.44	82.98	68.95	76.67	83.53	94.19	75.64	80.39
100	98.43	93.88	96.31	63.78	42.37	98.88	96.38	87.08	94.83	99.63	59.75	59.75	96.32	91.58	99.29	88.5	85.08
-	100	99.6	100	100	-	100	-	100	-	-	100	89.29	-	-	100	100	100
20.85	91.67	76.12	60.79	69.6	32.38	26.42	20.57	42.47	22.76	100	77.15	60.44	10.94	40.43	25.64	62.5	30.21
28.49	94.01	82.4	65.77	64.75	34.15	59.33	52.3	62.74	34.36	84.75	82.19	40.61	15.69	52.65	52.01	70.02	36.67
41.76	93.7	84.86	69.73	74.37	50.12	60.36	57.25	74.3	40.11	91.37	90.18	54.05	36.62	61.57	62.89	77.32	44.56
82.47	96.59	90	85.79	80.79	60.71	75.38	88.14	79.51	59.88	95.22	81.72	69.37	63.59	74.06	91.67	72.12	66.25
99.48	98.18	93.24	91.03	60.4	44.31	92.19	91.45	82.8	82.6	100	61.07	59.5	92.89	80.28	98.33	88.69	74.38
100	100	100	100	99.67	74.03	98.47	100	99.49	100	100	100	81.74	-	100	100	100	100
22.31	93.56	77.35	61.74	73.44	34.7	26.13	20.79	43.35	22.37	99.67	76.47	63.77	11.39	40.32	25.77	71.26	30.74
28.38	95.34	82.6	64.95	66.49	37.76	62.68	52.4	63.12	32.6	83.75	78.96	43.26	16.19	49.34	50.94	77.12	35.6
40.23	94.58	86.08	68.07	73.88	52	60.36	53.64	74.09	36.21	87.46	87.12	54.28	33.84	56.1	60.73	82.42	41.83
77.12	97.07	90.48	83.49	79.84	59.03	73.76	83.67	78.03	52	92.02	76.9	73.36	57.62	69.72	90.91	74.26	63.05
96.62	99.34	93.75	88.51	58.97	49.14	92.19	90.07	80.7	76.35	99.66	62.45	59.2	88.36	75.51	97.07	92.02	72.02
100	100	100	98.48	99.11	75.84	98.46	100	96.39	100	100	100	85	100	100	100	100	97.64
16.27	98.24	64.84	62.99	72.75	34.13	16.86	16.44	49.15	28.38	100	70.37	66.99	15.22	30.31	21.24	71.63	35.63
16.83	99.1	76.64	66.53	69.19	35.07	68.9	43.12	66.11	40.21	84.48	71.68	41.46	22.29	45.53	46.56	88.7	37.78
21.23	97.1	84.38	72.74	77.56	48.78	58.03	45.03	77.47	41.38	81.3	81.25	53.04	39.02	46.93	50.39	87.57	43.4
39.77	98.21	89.64	83.46	80.4	63.46	66.79	68.75	78.05	50.34	85	67.59	76.12	63.53	59.44	91.16	78.39	59.52
73.12	99.4	94.28	87.9	68.44	49.45	80.82	88.32	79.91	74.38	99.04	58.42	58.62	83.8	68.59	97.03	93.9	69.25
100	99.7	100	91.09	94.97	73.79	91.19	100	83.48	90.06	100	100	84	100	99.05	100	100	88.04

Supplementary Table S8

Burkholderia	Buchnera	Neisseria	Bordetella	Escherichia	Salmonella	Shigella	Yersinia	Vibrio	Haemophilus	Rickettsia	Chlamydophila	Fusobacterium	*Sn.* (%)	*Sp.* (%)
53.15	42.97	64.01	30.91	35.64	81.42	12.59	52.07	34.4	32.23	53.41	57.45	62.61	76.13	45.72
75.42	54.89	71.39	72.92	76	97.38	30.43	-	58.41	50.62	98.04	50	84.71	38.72	75.04
100	-	96.12	91.67	66.67	100	-	-	73.61	100	100	-	89.42	17.68	89.69
100	-	96.95	100	-	96.67	-	-	-	-	-	100	-	9	95.62
-	-	100	-	-	-	-	-	-	-	100	-	-	4.92	87.88
-	-	-	-	-	-	-	-	-	-	-	-	-	3.12	
58.45	33.26	82.66	26.91	32.83	78.79	17.3	37.99	26.84	27.78	49.75	41	76.99	83.26	45.57
61	51.44	73.21	37.61	48.26	81.8	19.85	71.74	42.51	33.21	74.06	70.91	84.73	83.59	58.65
86.98	95.68	92.83	54.92	54.5	84.66	32.26	83.75	58.55	45.17	95.1	78.93	86.06	80.82	74.84
93.93	97.32	97.95	78.85	62.61	85.04	40	100	89.26	65.47	99.5	86.42	100	63.08	88.55
99.37	97.06	99.64	99.19	65.68	96.11	85	94.12	99	-	96.55	96.82	100	44.53	90.92
100	-	100	100	-	100	-	-	-	-	-	-	-	3.77	100
58.55	33.3	84.68	25.67	33.4	78.16	17.85	35.16	26.4	28.36	49.5	38.46	83.09	84.41	46.29
57.57	47.99	74.81	33.94	54.05	78.23	22.97	61.46	40.67	30.94	69.12	66.22	87.65	86.27	56.22
77.52	69.01	93.69	47.42	53.64	81.35	31.32	71.12	50.53	36.49	83.71	72.87	82.58	86.3	67.4
86.22	87.13	97.67	72.19	64.47	81.94	41.79	91.82	70.3	49.23	98.32	84.31	98	84.37	82.41
97.39	86.78	99.32	92.96	63.52	89.9	45.71	98.35	92.69	100	96.26	91.65	97.83	71.93	87.23
100	100	100	100	-	100	100	-	100	-	100	98.65	-	9.29	99.38
59.82	29.46	90.27	24.65	36.34	79.24	19.53	35.85	26.46	28.11	51.64	37.12	93.4	86.29	47.83
52.95	41.3	78.48	32.78	58.95	75.56	28.34	58.82	36.86	28.01	71.09	66.48	94.32	88.29	56.03
70.39	52.21	97.09	41.97	54.63	77.75	31.29	61.15	45.64	31.02	71.6	67.76	83.24	88.79	62.93
81.01	72.87	98.36	66.48	64.68	81.31	46.64	78.03	58.99	40.3	94.04	78.49	98.68	90.41	76.55
91.47	68.85	99.67	81.11	68.2	86.7	52	91.5	77.39	70.65	92.88	84.58	98.03	86.83	82.38
99.74	100	100	100	-	97.41	84.62	-	98.15	99.14	100	83.79	100	34.6	97.01
59.48	26.58	94.92	24.81	33.24	79.04	21.06	37.52	25.82	27.53	45.39	33.61	94.92	86.99	48.36
51.89	37.5	81.08	32.79	61.58	75.2	28.65	56.66	36.22	28.38	62.7	67.94	96.77	89	56.09
67.35	47.73	98.36	41.2	50.9	77.06	30.61	60.44	43.48	28.93	62.11	67.75	84.03	89.43	61.38
78.12	69.23	99.34	63.25	70.09	82.81	47.35	73.32	56.79	38.73	88.51	78.97	99.67	91.72	74.79
89.12	62.63	99.67	75.95	73.68	86.78	51.3	86.87	70.05	54.65	86.69	84.18	98.36	89.08	80.38
98.87	100	100	100	76.71	93.59	79.81	100	98.43	94.05	100	86.2	100	65.88	96.08
64.93	26.81	100	25.33	54.21	78.55	25.09	42.38	31.15	23.59	41.62	20.49	99.21	87.58	48.54
52.37	39.78	77.51	34.8	72.75	74.67	30.49	61.37	38.94	22.06	58.47	62.83	99.21	89.5	56.31
64.41	42.05	100	46.11	71.89	75.89	31.79	65.48	45.68	20.7	51.8	65.14	85.62	90.57	60.42
74.01	62.71	100	66.48	83.07	81.97	45.47	71.67	53.45	29.67	80.23	77.9	100	92.48	71.82
87.7	54.41	100	70.23	85.36	84.2	47.61	82.95	60.09	39.53	82.76	81.14	99.21	90.85	77.41
97.54	100	100	95.7	86.2	90.58	88.69	90.56	92	59.44	100	88.05	100	88.41	93.04

Table S9: Classification accuracy of Phylopythia for genomic fragments of unknown organisms (to the classifier) at the taxonomic levels domain, phylum, class, order and genus for genomic fragments of different lengths. For every clade, the clade-specific sensitivity (class *Sn.*) and specificity (class *Sp.*) is displayed. The two rightmost columns give the overall sensitivity (*Sn.*) and specificity (*Sp.*) of de novo sequence classification (Supplementary Material, 'Evaluation procedures').

Rank	kernel	Wn															
Domain	Gaussian	6,6,1	Bacteria	Archaea	Eukaryota	Bacteria	Archaea	Eukaryota								*Sn.* (%)	*Sp.* (%)
	Data		Class *Sn.* (%)			Class *Sp.* (%)											
	1kbGF		85.22	34.14	53.67	97.38	77.01	69.47								57.68	81.29
	3kbGF		92.39	48.86	74.88	98.86	89.92	85.13								72.04	91.3
	5kbGF		95.9	33.22	78.33	98.57	99.71	92.47								69.15	96.92
	10kbGF		96.9	71.84	90.49	99.24	84.65	92.33								86.41	92.07
	15kbGF		98.98	51.87	79.73	98.56	97.68	99.09								76.86	98.44
	50kbGF		98.45	75.52	93.6	99.54	94.91	99.32								89.19	97.92
Phylum	Gaussian	6,4,1	Other	Cyanobacteria	Proteobacteria	Firmicutes	Deinococcus-Thermus	Actinobacteria	Spirochaetes	Chlamydiae	Crenarchaeota	Euryarchaeota	Fusobacteria	Ascomycota	Arthropoda	Chordata	Bacteroidetes
	Data		Class *Sn.* (%)														
	1kbGF		76.27	8.71	64.77	12.83	34.33	50.04	44.67	0	40	20.11	79.33	2.83	43	89.89	41.5
	3kbGF		42.27	76.64	90.2	83.17	54.33	91.36	59.17	87.89	78	73.17	96	54.5	66.25	97.33	85
	5kbGF		34.51	85.64	93.2	89.4	60.33	93.52	60.83	88.56	79.33	76.78	98	71.5	66.75	98.78	92
	10kbGF		35.66	85.79	94.69	89.89	64.67	94.65	33.44	88.44	83.67	81.89	98.67	81.5	81.33	99.67	86.33
	15kbGF		47.23	86.43	95.06	89.45	65.67	93.39	30.74	85.85	79.67	81.22	98.33	79.5	79.33	99.89	80.33
	50kbGF		36.54	86.14	97.68	87.83	74.85	96.3	14.74	78.79	88.28	87.64	100	91.17	86.33	99.67	82.64
Class	Gaussian	5,5,1	Other	Gammaproteo	Sordariomyce	Actinobacteria	Thermoprotei	Thermoplasm	Thermococci	Clostridia		Deinococci	Bacteroides(c	Fusobacteria(	Spirochaetes(	Chlamydiae(cl	Methanomicrc	Alphaproteoba
	Data		Class *Sn.* (%)															
	1kbGF		92.55	26.68	0	16.88	38	25.75	10.67	3.78		26.67	48	58.67	3.17	0	2.5	7.04
	3kbGF		70.33	87.39	77	85.52	89.67	84.5	95	65.56		54.67	94.67	94.67	60.83	84.44	55.75	81.71
	5kbGF		57.81	93.21	88	90.52	91	90.75	98	73.67		62.33	98.33	98.67	65.33	89	64.75	91.25
	10kbGF		46.79	96.12	94.67	93.24	94.33	98.25	99	76		65.33	100	99.67	65.89	89	79.25	93.39
	15kbGF		43.87	97.13	97	93.35	96	99.25	99.33	76		66	100	99.33	60.74	86.27	85.5	93.67
	50kbGF		43.68	98.92	96.33	95.81	99.31	100	100	73.24		74.85	100	100	31.09	79.65	94.58	97.68
Order	Gaussian	5,5,1	Other	hroococcal	Nostocales	Prochlorale	nthomonad	Vibrionales	asteurellale	pirochaetal		Bacillales	acteroidale	Clostridiales	actobacillal	isobacteral	tinomyceta	hodospirillal
	Data		Class *Sn.* (%)															
	1kbGF		98.92	0.25	0	0	0	0	2.92	2.67		0	0	0	10.21	0	0	0
	3kbGF		96.18	13.25	28	0	51.29	17.25	70.92	20.5		10.69	61.33	36.86	76.46	0	53.62	3.33
	5kbGF		91.44	30.25	88.33	2.5	86	66.25	87.83	27.83		51.65	94.67	49.29	88.54	3.67	83.38	13.67
	10kbGF		83.79	48.5	98	67.75	95.86	92.75	96.17	31.27		73.23	100	60.57	94.43	95.67	94.77	31.67
	15kbGF		79.84	50.25	100	73.75	97.57	98.25	98.67	28.52		84.88	100	71	96.35	99.67	96.54	33.33
	50kbGF		74.66	47.33	100	67.79	97.43	100	99.79	14.42		90.57	100	67.6	97.31	100	99.75	37.8
Genus	Gaussian	5,5,1	Other	ochlorococc	aphylococc	treptococcu	Bacillus	Clostridium	actobacillu	Listeria		rynebacteri	ycobacteriu	treptomyce	ampylobact	elicobacte	lycoplasma	ethanosarci
	Data		Class *Sn.* (%)															
	1kbGF		99.05	0	0	0.35	0	0.17	7.75	0		0	0	0	0	11.75	6.4	0
	3kbGF		96.56	0.75	75.5	74.65	15.9	63.33	23	30.25		4	31.14	57.33	38.8	29	29.2	17.67
	5kbGF		93.32	26.75	94	83.94	60.9	76.17	25	73		44.33	64.29	85.67	54	37	43.8	52.33
	10kbGF		87.53	50.5	99	92.12	85.3	81.17	31.25	96.5		71.33	89.57	99.67	58.8	53.75	52.42	78.33
	15kbGF		84.9	70	99.67	97.53	88.2	82.5	44.27	98		81.67	94.14	99	85	68	60.47	86
	50kbGF		81.96	67.79	100	99.85	89.46	83.18	46.51	99.58		98.87	97.37	99.67	100	72.52	75.25	95.73

Supplementary Table S9

Other	Cyanobacteria	Proteobacteria	Firmicutes	Deinococcus-	Actinobacteria	Spirochaetes	Chlamydiae	Crenarchaeot	Euryarchaeota	Fusobacteria	Ascomycota	Arthropoda	Chordata	Bacteroidetes		*Sn.* (%)	*Sp.* (%)
Class *Sp.* (%)																	
6.47	93.13	92.62	97.26	91.15	93.57	42.68	-	93.75	94.27	82.07	73.91	29.01	78.09	71.55		40.55	79.47
18.06	94.37	94.94	95.21	90.06	91.03	57.35	97.65	81.82	80.01	86.23	73.98	43.02	88.57	73.12		75.68	81.95
24.64	93.09	94.98	95.03	87.02	91.04	64.49	96.96	84.4	79.98	88.55	69.98	42.58	92.51	72.73		79.28	82.38
21.09	97.25	95.94	97.24	88.99	92.67	60.61	98.03	95.8	83.14	98.34	74.54	50	95.94	67.8		80.02	85.45
22.39	97.27	96.92	98.04	94.26	95.6	71.24	96.12	97.95	88.02	100	77.81	52.31	97.51	64.61		79.47	87.69
23.17	98.13	97.1	99.18	98.43	98.77	75.41	82.35	93.43	86.85	100	78.59	55.46	99.78	65.56		80.57	87.79

Betaproteoba	Deltaproteoba	Epsilonproteo	Mollicutes	Mammalia	Insecta	Actinopterygii	Bacilli	Other	Gammaprotec	Sordariomyce	Actinobacteria	Thermoprotei	Thermoplasm	Thermococci	Clostridia	Deinococci	Bacteroides(c
Class *Sp.* (%)																	
33.38	7.38	83.8	72.33	72.5	32.5	43	0.54	14.82	94.63	-	98.37	73.08	92.79	100	97.14	86.96	80
85.08	48.5	94.7	92.67	92	67.75	80	80.11	44.58	92.13	74.28	90.59	85.67	92.6	86.1	86.38	83.25	80.91
89.62	54.75	96.6	94.67	96	71.75	84	91.39	58.34	92.16	65.51	88.16	89.8	91.21	76.56	75.95	84.62	74.12
93.08	59.75	98.2	94.71	99	77.25	92.33	95.2	67.98	92.51	63.82	84.93	93.09	95.39	72.62	74.03	83.4	70.26
94	60.62	98.9	96.15	100	79.25	95.67	96.46	71.98	92.95	62.85	83.43	95.36	93.41	69.95	74.84	84.62	68.49
96.48	63.54	99.71	97.75	99.5	84.25	96.33	97.55	82.67	93.62	65.68	87.12	94.12	92.81	69.03	85.71	91.24	72.12

odobacteria	Neisseriales	coplasmata	npylobacter	ermococca	moplasmat	Rhizobiales	hlamydiale	lfuromonac	Diptera	udomonada	Rickettsiale	urkholderial	terobacteria	hanosarcin	Other	hroococcal	Nostocales
															Class *Sp.* (%)		
0	16.67	0.5	58.5	0	0.25	0	0	0	0	0.8	0	0	0	0	22.26	100	-
38.33	75.5	47.5	88.3	24	45	0	5.89	32	39.67	39.3	4.56	39.27	41.14	18.5	31.15	50.96	98.82
54	79.5	60.33	93.5	83.33	61.75	5.3	69.44	57.67	56.33	57.9	64.89	73.73	86.08	42	45.62	56.28	98.15
63.33	83.17	71.83	98	95.33	90.5	86.3	88.56	70	69.67	73.5	88.78	87.47	95.16	60	64.1	65.54	88.29
62.67	83.33	75.58	99.1	98.67	96.25	93.8	86.27	72.67	72	77.5	89.14	91.73	96.4	65.75	70.65	64.01	85.96
65.04	69.65	75.86	100	100	100	99.87	79.22	79.62	76	86.57	96.46	95.31	97.36	84.34	77.38	68.21	78.95

Pyrococcus	Brucella	Xylella	seudomona	urkholderia	Buchnera	Neisseria	Bordetella	Escherichia	Salmonella	Shigella	Yersinia	Vibrio	laemophilu	Rickettsia	hlamydophi	sobacteriu	Other	
																	Class *Sp.*	(%
1	0.33	0.75	0	0	0	13	0	0	0	0	0	0	0	1	0	0	51.27	
64	69	49	35	45.14	11	68.75	30.75	28.43	46.38	4.25	2.29	24.75	0	49	60.83	60.33	62.91	
87	93.67	70	66	74.71	65.67	72.75	66	54.71	64.5	20	59.71	60.25	0.25	76.5	91.5	90.33	74.55	
96.33	98.33	94	85.57	94.86	94.21	86.5	73	65	86.25	40.75	96.86	80	80.5	96.5	99.67	99.33	85.47	
98.67	99.33	98.5	94.71	98.29	98.41	95.25	75	66.71	96.5	55.5	98.29	95.75	91	100	99.79	99.67	90.57	
100	100	99.04	99.86	99.86	100	100	79.43	93.79	98.74	76.84	98.77	99.47	97.87	100	100	100	95.31	

Supplementary Table S9

Fusobacteria(	Spirochaetes(	Chlamydiae(cl	Methanomicrc	Alphaproteob	Betaproteoba	Deltaproteoba	Epsilonproteo	Mollicutes	Mammalia	Insecta	Actinopterygii	Bacilli	*Sn.* (%)	*Sp.* (%)
78.92	95	-	83.33	98.99	94.12	64.84	81.36	84.55	54.92	39.51	62.32	100	30.69	83.04
88.2	89.24	96.82	69.04	95.61	92.65	60.62	93.39	92.36	66.67	45.93	76.68	96.05	79.24	83.42
90.24	84.67	91.44	63.48	95.05	94.59	58.4	97.09	93.73	70.46	48.07	82.89	95.53	83.97	81.99
92.86	88.14	85.12	56.71	92.4	96.67	58.79	98.69	95.07	75.57	47.69	84.97	95.54	86.98	81.74
92.55	87.7	77.08	55.88	91.6	97.37	57.6	99.4	95.04	76.48	50.8	92.58	95.85	87.59	81.63
91.24	89.81	45.32	65.01	92.75	98.19	64.46	100	96.67	78.66	55.98	98.63	94.28	87.84	82.84

Prochlorale	nthomonad	Vibrionales	asteurellale	pirochaetale	Bacillales	acteroidale	Clostridiales	actobacillal	usobacteral	Actinomyc	Rhodospiri	Rhodobact	Neisseriale	Mycoplasm	Campyloba	Thermoco	Thermopla
-	-	-	100	94.12	-	-	-	72.59	-	-	-	-	100	100	98.32	-	100
-	96.51	95.83	98.84	83.11	99.29	99.46	88.36	91.38	-	99.29	14.71	98.29	99.12	84.32	98.55	100	100
100	95.25	91.07	96.88	89.3	98.82	97.26	79.31	92.88	100	96.69	24.7	99.39	98.55	74.18	99.36	99.21	99.6
100	89.59	78.11	95.37	89.47	97.99	89.29	75.58	94.84	99.31	93.2	34.55	100	99.6	68.47	99.8	90.79	99.45
100	91.19	70.18	92.14	88.51	98.22	86.71	79.78	96.17	99.34	91.17	33.9	98.43	100	65.04	99.9	80.22	99.23
82.79	95.35	67.38	87.25	95.74	96	80.43	87.83	93.89	100	92.09	34.32	98.86	99.54	63.31	100	72.3	96.99

ochlorococc	aphylococc	treptococcu	Bacillus	Clostridium	actobacillu	Listeria	rynebacteri	ycobacteriu	treptomyce	Campyloba	Helicobact	Mycoplasm	Methanosa	Pyrococcus	Brucella	Xylella	Pseudomo
%)																	
-	-	100	-	100	15.58	-	-	-	-	-	97.92	65.31	-	100	100	100	-
100	98.48	95.34	100	68.72	41.07	100	100	93.97	100	100	58.59	64.6	100	97.96	100	90.74	94.96
100	98.43	93.88	96.36	63.83	42.37	98.98	96.38	87.89	94.83	99.63	60.41	69.3	96.32	91.58	99.29	88.61	85.08
99.51	98.18	93.6	91.82	60.42	47.71	93.92	91.45	84.84	82.6	100	67.82	70.83	92.89	80.28	98.33	90.82	75.35
97.56	99.34	94.36	89.63	59.07	59.59	94	90.07	83.21	76.35	100	73.12	74.28	88.36	75.51	97.07	93.81	74.41
80.16	99.4	94.92	88.91	66.48	63.49	84.53	88.38	82.36	72.05	100	69.34	76	83.8	68.59	97.03	95.39	72.36

Supplementary Table S9

Rhizobiales	Chlamydial	Desulfuron	Diptera	Pseudomo	Rickettsiale	Burkholder	Enterobact	Methanosarcinales	*Sn.* (%)	*Sp.* (%)
-	-	-	-	100	-	-	-	-	6.39	96.11
-	100	98.97	63.3	81.37	100	99.16	99.87	100	35.95	89.98
100	100	95.58	59.72	79.21	100	98.22	99.66	93.85	60.37	90.11
96.75	98.15	91.7	53.18	78.44	99.13	97.84	99.28	86.02	79.53	87.92
94.08	94.24	90.46	52.17	77.89	97.82	97.59	98.99	83.23	82.32	86.43
85.33	60.4	94.95	54.42	80.06	89.71	97.74	98.92	87.5	83.32	84.15

Burkholder	Buchnera	Neisseria	Bordetella	Escherichi	Salmonella	Shigella	Yersinia	Vibrio	Haemophil	Rickettsia	Chlamydor	Fusobacterium	*Sn.* (%)	*Sp.* (%)
-	-	100	-	-	-	-	-	-	-	100	-	-	4.42	87.88
99.37	97.06	99.64	99.19	65.68	96.11	85	94.12	99	-	96.55	96.82	100	38.62	91.1
97.39	86.78	99.32	92.96	63.52	89.9	45.71	98.35	92.69	100	96.53	91.65	97.83	63.38	87.61
92.35	68.85	99.71	81.11	68.32	87.79	52.58	91.5	79.01	74.88	94.38	84.23	98.35	81.09	83.66
90.53	62.63	99.74	75.95	73.89	87.13	55.78	86.87	75.1	61.28	89.49	83.45	99.01	87.52	82.6
89.27	54.41	100	69.67	86.53	85.92	73	82.54	67.02	46.62	86.24	80.23	100	92.23	80.8

Table S10: Comparison of Phylopythias' classification accuracy for 3kb genomic fragments from unknown organisms (3kb GF) to that of genomic fragments from unknown organisms which carry highly expressed ribosomal proteins (3kb RP GF). For each organism where this annotation was available (322 of the 340), 10 3kb fragments carrying one or more ribosomal protein encoding genes were classified. For the clades at the taxonomic levels domain, phylum, class, order and genus, the overall sensitivity (*Sn.*) and specificity (*Sp.*) of *de novo* assignment for these fragments is shown.

Rank	*Sn.* (%) 3kb RP GF	*Sp.* (%)	*Sn.* (%) 3kb GF	*Sp.* (%)
Domain	42.34	81.13	72.04	91.3
Phylum	76.08	83.26	75.68	81.95
Class	78.08	83.4	79.24	83.42
Order	27.17	91.39	35.95	89.98
Genus	26.22	91.42	38.62	91.1

Table S11: Phylogenetic classification accuracy of Phylopythia for genomic fragments of known organisms (other fragments of these organisms were included in the training set used for model creation) at the phylogenetic levels domain, phylum, class, order and genus for differently sized genomic fragments. For every phylogenetic level, the sensitivity (*Sn.*) and specificity (*Sp.*, averaged over all clades where 300+ fragments were available for classification) of *de novo* sequence classification is displayed. The specificity for 50k fragments at the genus level is undetermined (n.a.), due to too little available evaluation data. Δ S*n.* and Δ *Sp.* give the improvement of classification accuracy compared to that for unknown organisms.

Rank	GF	*Sn.* (%)	*Sp.* (%)	Δ *Sn.*	Δ S*p.*
Domain	1kb	57.71	88.68	0.03	7.39
	3kb	92.35	90.7	20.31	-0.6
	5kb	96.03	98.28	26.88	1.36
	10kb	98.15	98.03	11.74	5.96
	15kb	97.59	99.72	20.73	1.28
	50kb	99.8	99.95	10.61	2.03
Phylum	1kb	50.32	92.11	9.77	12.64
	3kb	94.75	94.84	19.07	12.89
	5kb	96.53	94.73	17.25	12.35
	10kb	97.93	97.22	17.91	11.77
	15kb	98.77	98.61	19.3	10.92
	50kb	97.99	98.82	17.42	11.03
Class	1kb	30.76	93.21	0.07	10.17
	3kb	94.64	94.29	15.4	10.87
	5kb	96.93	93.75	12.96	11.76
	10kb	97.65	96.19	10.67	14.45
	15kb	98.9	98.72	11.31	17.09
	50kb	99.53	98.45	11.69	15.61
Order	1kb	25.12	95.04	18.73	-1.07
	3kb	94.18	93.89	58.23	3.91
	5kb	96.93	92.75	36.56	2.64
	10kb	98.5	96.79	18.97	8.87
	15kb	98.66	96.7	16.34	10.27
	50kb	99.66	98.61	16.34	14.46
Genus	1kb	7.11	96.72	2.69	8.84
	3kb	69.16	92.5	30.54	1.4
	5kb	95.43	89.18	32.05	1.57
	10kb	98.07	83.68	16.98	0.02
	15kb	96.63	87.82	9.11	5.22
	50kb	99.52	n.a.	7.29	n.a.

Supporting Online Material

Alice C. McHardy, Héctor García Martín, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos

Materials and methods

Multi-class Support Vector Machine training

The compositional input vectors for the training of the SVM were created by mapping the input sequences s to the feature space π, that is defined by the pattern length w and number of literals l (in this context, a literal is a character from the DNA alphabet, see Figure S10*)*. The input vectors were normalized per row and scaled across columns in the range [0, 1]. Similar number of sequences where used for each class in model training. If it was not possible to include exactly the same number of sequences for each class, the miss-classification 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 was used as the kernel function with the SVM, which requires specification of the parameter γ. The optimal values for C and γ were 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 were used for the grid search and 1000 for the training of the classifier. For the domain level, 1000 sequences per clade were used for the grid search and 3000 for training of the classifier.

Combined metagenome classifier

The final classification framework of Phylopythia includes several classifiers that are built on fragments of different lengths for each level. Based on results of the extensive evaluation, we decided to include classifiers trained on 5, 10, 15 and 50kb fragments into

the framework for the level of the phylum and class, and classifiers trained on 50 and 15kb fragments for the lower levels, to maintain a high specificity level. Including further classifiers would yield an increased sensitivity, but we opted for a setting with higher specificity. Starting with the 50kb-trained classifier, a sequence is tested with these classifiers in descending order of training fragment lengths, until either an assignment is made, all classifiers have been applied, or a classifier is reached that has been trained with fragments shorter than the tested sequence. For the domain level, classifiers trained on 1, 3, 5, 10, 15, and 50kb fragments are used, each for fragments with a similar size to the respective training fragments. For the classification of fragments from known organisms, 10kb models are also included at the order and genus levels. For the metagenome sludge samples, additional 10, 15 and 50kb (15 and 10kb) 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.

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 generate training sequences and train the phylogenetic classifier. Thus, for nearly all organisms a model could be created which had not used any of the organisms sequence in training. To estimate accuracy for fragments of known organisms, for any given fragment length, a part of the genomic sequence was set aside for evaluation, while the other parts were 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 $Sn = \left(\sum_{i=1}^{N} \frac{tp_i}{t_i} + \frac{tp_{other}}{t_{other}} \right) \cdot \frac{1}{N+1}$, and 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, the sensitivity over (*N* + 1) clades, which includes the class 'Other'.

Supporting Figures and Tables

Fig. S1. Assignment accuracy for differently sized genomic fragments from unknown organisms at the level of the class (including 22 phylogenetic clades). 7 class-level classifiers were trained with tetranucleotide patterns (*w*4, *l*4) derived from genomic fragments of length 1, 3, 5, 10, 15 and 50kb as well as protein encoding sequences (CDS). For each classifier, the sensitivity (percentage of data classified correctly) (A) and specificity of assignments (B) with fragments of different lengths is shown for organisms that are unknown to the model (of which no sequences were used in training).

Fig. S2. *Wn* parameter search for the sequence composition space with the highest classification accuracy for genomic sequence fragments from unknown organisms (of which no sequence fragments were included in the model training data sets) at different phylogenetic levels. The legend gives in brackets the number of clades for each of the analyzed levels. The composition space is defined by the word length *w*, the number of literal characters *l*, and the step size *s*. Plot (A) shows the sensitivity and plot (B) the specificity, that is attainable in a given space with 15kb genomic fragments of organisms unknown to the classifier.

Fig. S3-7. Evaluation of the relation of genomic fragment length used for model creation and classification accuracy for genomic fragments of unknown organisms and different lengths. For the level domain (S3), phylum (S4), class (S5), order (S6) and genus (S7), the sensitivity (Sn.) and specificity (Sp.) of classification is displayed for genomic fragments of different lengths with classifiers trained on fragments of different lengths.

Fig. S8. Assignments at the domain level with Phylopythia for 50kb genomic fragments from unknown organisms (of which no sequence was used in training). Displayed is for every organism the percentage of fragments assigned to the archaea (green), bacteria (orange) and eukaryota (blue). Grey indicates an assignment to 'origin unknown'.

Fig. S9. Comparison of classification accuracy for 3kb fragments and 3kb fragments carrying ribosomal proteins with Phylopythia. For the clades at the domain, phylum, class, order, and genus level, the specificity of assignment for fragments from organisms unknown to the classifier is displayed, as well as the overall sensitivity and specificity of classification.

Fig. S10. Example of a (*w, l)*-pattern, where *w* is 6 and the number of literals *l* is 4. In this context, literals correspond to specified characters of the DNA alphabet, whereas the dot symbols (wildcards) match any of these characters.

Table S1. Species with (near-) complete genome sequences included in this study. The right column gives the number of organisms for every species.

Table S2. *Wn* parameter search for the sequence composition space with the highest classification accuracy for genomic fragments of 340 unknown organisms (to the classifier) at different phylogenetic levels. The composition space is defined by the word length *w*, the number of literal characters *l*, and the step size *s. Acc.* denotes the

percentage of correctly assignments for all tested fragments, *Sn.* and *Sp.* are the sensitivity and specificity of the classification (Supplementary material, 'Evaluation procedures').

Table S3. *Wn* parameter search for the sequence composition space with the highest classification accuracy for protein encoding genes of 340 unknown organisms at different phylogenetic levels. The composition space is defined by the word length *w*, the number of literal characters *l*, and the step size *s*. Results are shown for the multi-class classification without post-processing by the binary classifiers. *Acc.* denotes the percentage of correctly assignments for all tested fragments, *Sn.* and *Sp.* are the sensitivity and specificity of the classification (Supplementary material, 'Evaluation procedures').

Tables S4-8: Evaluation of the relation of the fragment length used for model creation to the accuracy of assignments for genomic fragments from unknown organisms. For every clade, the clade-specific sensitivity (class *Sn.*), specificity (class *Sp.*) of assignments for genomic fragments of different lengths from the organisms of this clade among the 340 organisms is shown. The two rightmost columns display the overall sensitivity (*Sn.*) and specificity (*Sp.*) of *de novo* sequence classification (Supplementary Material, 'Evaluation procedures'). The information in the tables describes the phylogenetic levels domain (S4), phylum (S5), class (S6), order (S7) and genus (S8).

Table S9: Classification accuracy of Phylopythia for genomic fragments of unknown organisms (to the classifier) at the taxonomic levels domain, phylum, class, order and genus for genomic fragments of different lengths. For every clade, the clade-specific sensitivity (class *Sn.*) and specificity (class *Sp.*) is displayed. The two rightmost columns give the overall sensitivity (*Sn.*) and specificity (*Sp.*) of *de novo* sequence classification (Supplementary Material, 'Evaluation procedures').

Table S10: Comparison of Phylopythias' classification accuracy for 3kb genomic fragments from unknown organisms (3kb GF) to that of genomic fragments from unknown organisms which carry highly expressed ribosomal proteins (3kb RP GF). For each organism where this annotation was available (322 of the 340), 10 3kb fragments carrying one or more ribosomal protein encoding genes were classified. For the clades at the taxonomic levels domain, phylum, class, order and genus, the overall sensitivity and specificity (*Sn., Sp*.) of *de novo* assignment for these fragments is shown.

Table S11: Phylogenetic classification accuracy of Phylopythia for genomic fragments of known organisms (other fragments of these organisms were included in the training set used for model creation) at the phylogenetic levels domain, phylum, class, order and genus for differently sized genomic fragments. For every phylogenetic level, the sensitivity (*Sn*.) and specificity (*Sp*., averaged over all clades where 300+ fragments were available for classification) of *de novo* sequence classification is displayed. The specificity for 50k fragments at the genus level is undetermined (n.a.), due to too little available evaluation data. Δ *Sn.* and Δ *Sp.* give the improvement of classification accuracy compared to that for unknown organisms.