

# CREST – Classification Resources for Environmental Sequence Tags

Anders Lanzén<sup>1,2\*</sup>, Steffen L. Jørgensen<sup>1</sup>, Daniel H. Huson<sup>3</sup>, Markus Gorfer<sup>4</sup>, Svenn Helge Grindhaug<sup>2</sup>, Inge Jonassen<sup>2,5</sup>, Lise Øvreås<sup>1</sup>, Tim Urich<sup>6\*</sup>

**1** Department of Biology and Centre for Geobiology, University of Bergen, Bergen, Norway, **2** Uni Computing, Uni Research AS, Bergen, Norway, **3** Centre for Bioinformatics, Tübingen University, Tübingen, Germany, **4** Fungal Genetics and Genomics Unit, AIT GmbH and University of Natural Resources and Life Sciences, Tulln, Austria, **5** Department of Informatics, University of Bergen, Bergen, Norway, **6** Department of Genetics in Ecology, University of Vienna, Vienna, Austria

## Abstract

Sequencing of taxonomic or phylogenetic markers is becoming a fast and efficient method for studying environmental microbial communities. This has resulted in a steadily growing collection of marker sequences, most notably of the small-subunit (SSU) ribosomal RNA gene, and an increased understanding of microbial phylogeny, diversity and community composition patterns. However, to utilize these large datasets together with new sequencing technologies, a reliable and flexible system for taxonomic classification is critical. We developed CREST (Classification Resources for Environmental Sequence Tags), a set of resources and tools for generating and utilizing custom taxonomies and reference datasets for classification of environmental sequences. CREST uses an alignment-based classification method with the lowest common ancestor algorithm. It also uses explicit rank similarity criteria to reduce false positives and identify novel taxa. We implemented this method in a web server, a command line tool and the graphical user interfaced program MEGAN. Further, we provide the SSU rRNA reference database and taxonomy SilvaMod, derived from the publicly available SILVA SSURef, for classification of sequences from bacteria, archaea and eukaryotes. Using cross-validation and environmental datasets, we compared the performance of CREST and SilvaMod to the RDP Classifier. We also utilized Greengenes as a reference database, both with CREST and the RDP Classifier. These analyses indicate that CREST performs better than alignment-free methods with higher recall rate (sensitivity) as well as precision, and with the ability to accurately identify most sequences from novel taxa. Classification using SilvaMod performed better than with Greengenes, particularly when applied to environmental sequences. CREST is freely available under a GNU General Public License (v3) from <http://apps.cbu.uib.no/crest> and <http://lcaclassifier.googlecode.com>.

**Citation:** Lanzén A, Jørgensen SL, Huson DH, Gorfer M, Grindhaug SH, et al. (2012) CREST – Classification Resources for Environmental Sequence Tags. PLoS ONE 7(11): e49334. doi:10.1371/journal.pone.0049334

**Editor:** Ludovic Orlando, Natural History Museum of Denmark, University of Copenhagen, Denmark

**Received:** June 26, 2012; **Accepted:** October 10, 2012; **Published:** November 8, 2012

**Copyright:** © 2012 Lanzén et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported by the Norwegian Research Council (<http://www.forskningsradet.no>; project number 179560), L. Meltzers høyskolefond (<http://meltzerfondet.no>; project number 805372) and the Austrian Federal Ministry of Science and Research (<http://www.bmwf.gv.at>; project “GEN-AU III InflammoBiota”). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: anders.lanzen@uni.no (AL); Tim.Urich@univie.ac.at (TU)

## Introduction

Marker gene sequencing (also known as “barcoding” or “meta-barcoding”) is an increasingly common technique for profiling the taxonomic composition and diversity of environmental samples. Facilitated by the rapid development of sequencing technologies, such studies are rapidly becoming routine and increasingly sophisticated. The technique has a clear potential for revolutionizing the field of microbial ecology as well as medical microbiology, by frequent and routine profiling of environmental as well as human microbiome samples. It has even been used in macroecology to monitor the distribution and dispersal of animal species [1]. For prokaryotes, the small-subunit ribosomal RNA (SSU rRNA) has become the *de facto* standard marker gene targeted by amplicon (or “tag”) sequencing [2]. However, the choice of marker, primers and marker region targeted varies among individual research laboratories and studies. In metagenomic or metatranscriptomic studies, sequences containing SSU rRNA or other markers can also be subjected to taxonomic profiling [3,4,5]. In either case, none of the existing “next-generation” sequencing

protocols available allow for full-length sequencing of the SSU rRNA gene.

Pyrosequencing as developed by 454 Life Sciences (Roche) was the first high-throughput sequencing technology to be applied for sequencing of SSU rRNA. The current generation of pyrosequencing instruments (GS FLX+) can generate shotgun sequencing reads up to 800 bp long, while amplicon sequencing is only supported using the previous generation chemistry at the time of writing with read lengths of approximately 450 bp [6]. Other sequencing platforms gaining popularity are Illumina Hi-Seq, yielding read lengths of 100–150 bp [7], or 200 bp if assembly of paired-end reads is used as described by [8] or [9], and IonTorrent, yielding read lengths over 200 bp. Regardless of technology used, accurate taxonomic classification is of paramount importance to the interpretation of the resulting sequencing data [2]. The quality of results depends on read length, choice of taxonomic marker, region (the latter particularly important for shorter reads) [10], and last but not least on the classification method and taxonomy applied. Indeed, the quality of the

taxonomy and reference database can have a more significant effect on results than the classification method [10].

In the last years, a large amount of SSU rRNA sequence data has been collected, organized and aligned in databases such as SILVA [11], the Ribosomal Database Project (RDP) [12] and Greengenes [13]. However, we are still only “scratching the surface” of global biodiversity with countless novel species and genera waiting to be discovered [14,15,16]. Many are also hidden among the hundreds of thousands of existing environmental sequences disguised under uninformative labels such as “uncultured bacteria” and thus remain without proper taxonomical descriptions. We can therefore expect that the Tree of Life representing our current understanding of the phylogeny of all living and extinct organisms, will receive many new branches and undergo many topological changes. Meanwhile, taxonomic classification will remain challenging. How to deal with the environmental sequences whose taxonomical affiliations remain unclear is a crucial consideration, as these unquestionably can be useful for classifying other similar environmental sequences. Including such environmental sequences in addition to cultured type strains may be crucial for phylogenetic or taxonomical work [17].

One challenging issue in taxonomical classification is polyphyletic taxa. Some of these, e.g. “Uncultured bacteria” are intentionally created as placeholders for sequences whose taxonomical affiliations are unclear. Others result from submission of sequences with incorrect taxonomical classification or incomplete knowledge of a phylogenetic group. Both categories can cause the classification sensitivity and resolution to drop [3]. Assignments to polyphyletic groups are also inherently less meaningful. Still, there are situations where taxa are well established in the literature, but known to be polyphyletic, e.g. *Clostridia* and *Bacilli* [18]. Removing them is not always desired until an alternative taxonomy has been established. Another challenging issue is identification of novel sequences. In order not to miss such potentially interesting information it is important to clearly be able to identify them, rather than assigning them incorrectly to an existing taxon [19].

To better deal with the above-mentioned issues, we present a set of resources for taxonomic classification that utilize environmental sequences together with reference strains. Branded as CREST (**C**lassification **R**esources for **E**nvironmental **S**equences **T**ags), we present a simple alignment- and lowest common ancestor (LCA) based taxonomic classification method, implemented as a web-server, command line tool and in a new version of the program MEGAN [20]. We also present a reference database and taxonomy for classification of environmental SSU rRNA sequences. This reference database, labeled *SilvaMod*, was derived using extensive manual curation from the taxonomically annotated SILVA Reference alignment (SSURef nr release 106) [11]. In addition, *SilvaMod* includes explicit rank information derived from the NCBI Taxonomy. A similar strategy was recently carried out by taxonomical annotation of Greengenes [21].

CREST is equally suitable for classification of sequencing data from SSU rRNA PCR amplicons as from shotgun metatranscriptome or metagenome sequences, not only from bacteria and archaea (prokaryotes), but also from eukaryotic taxa. We illustrate the performance of the databases and the assignment method, and compare this to the RDP Classifier [22] and SINA Aligner, both of which are commonly used methods for taxonomic classification of SSU rRNA sequences. We use two types of cross-validation; ten-fold, and removal of taxa, the latter to better simulate a situation where a novel taxon is discovered. We also apply the method to four environmental datasets generated using different sequencing

technologies and compared the number of identified taxa and the proportion of classified reads at different taxonomical ranks. While originally developed for classification of SSU rRNA sequence data, CREST has wide applicability since it provides a framework for generating and utilizing custom taxonomies and reference databases. This procedure only requires a taxonomically annotated custom alignment, created e.g. with the program ARB [23].

## Results

### Overview of CREST

Figure 1 presents the resources of CREST, along with the flow of information during the construction of a new reference database (top part) or classification (bottom part). CREST includes:

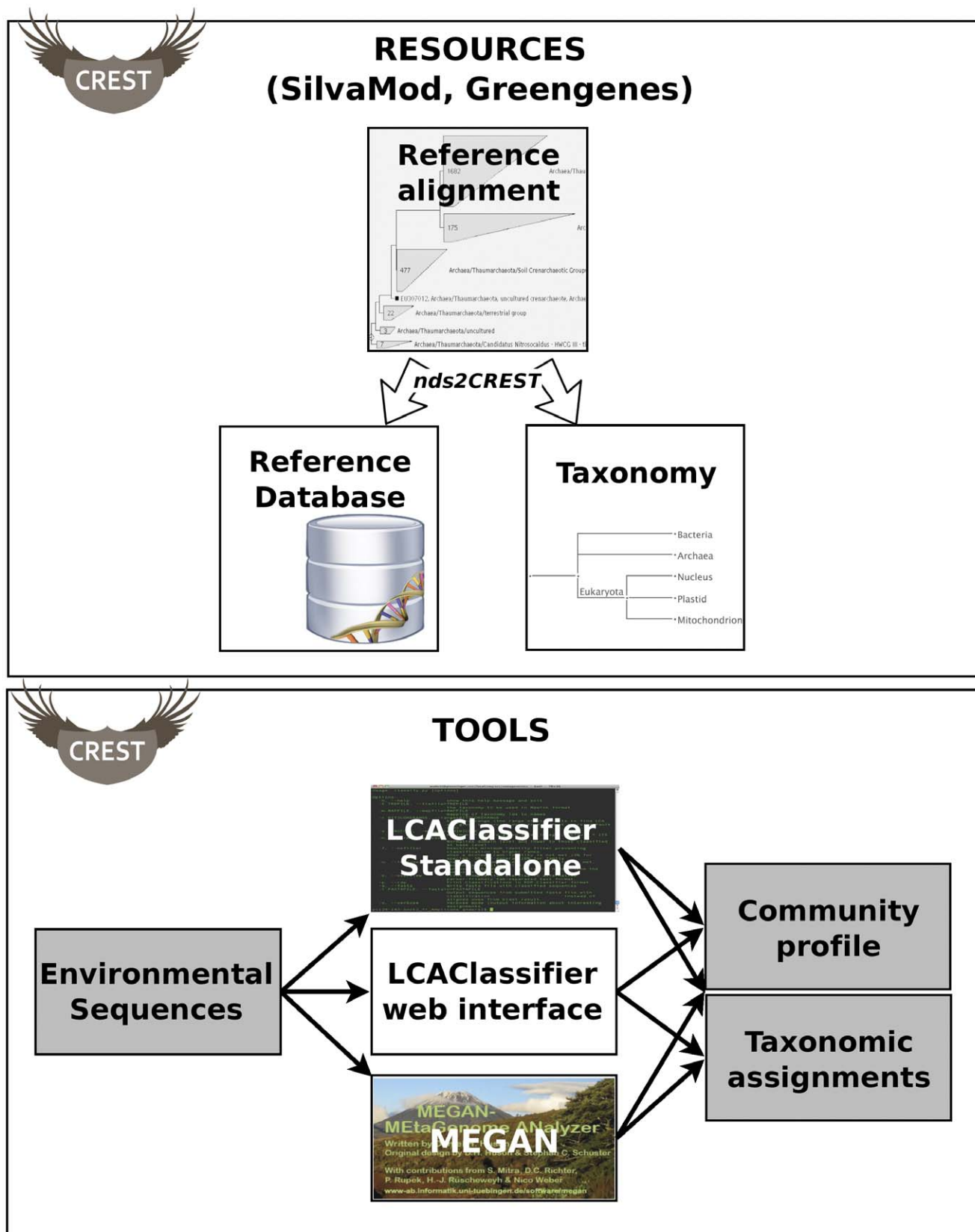
1. the manually curated SSU rRNA taxonomy and reference database *SilvaMod* based on a modification of the taxonomical annotation used in SILVA SSURef nr release 106;
2. supplementary files for using the Greengenes taxonomy and database as an alternative;
3. a simple classification method based on pairwise alignment and assignment to the lowest common ancestor (LCA) of the resulting highest-scoring alignments;
4. implementations of the classification method as webserver and command line tool (LCAClassifier), and;
5. a new version of the program MEGAN [20] offering CREST classification.

The LCAClassifier and MEGAN implement the same classification algorithm and can use any taxonomy and CREST-compatible reference database and taxonomy, in addition to *SilvaMod* and Greengenes. Starting with a reference alignment of a taxonomic marker sequences from a collection of taxa, such databases can be derived with the script *nds2CREST* (see Methods and online technical documentation at <http://code.google.com/p/lcaclassifier/wiki/Userguide> for details).

### *SilvaMod* and Greengenes Reference Databases

Release 106 of the SILVA non-redundant SSURef database includes manual taxonomic annotations of its aligned sequence clusters, with a resolution up to genus rank. According to the developers, annotations were based on Bergey’s Manual of Systematic Bacteriology (vol. 1 to 4) [24,25,26,27], the List of Prokaryotic names with Standing in Nomenclature [28], *Candidatus* taxa and names without standing in nomenclature (described in detail at <http://www.arb-silva.de/documentation/faqs/>). We curated these annotations and the taxonomical structure itself (the “SILVA Taxonomy”), in order to comply with recent phylogenetic work and to incorporate proposed environmental clades as suggested in a selection of phylogenetic studies (see Methods). Importantly, we only carried out such revisions if the resulting annotation agreed with the clustering in SSURef, in order to avoid inserting apparently polyphyletic taxa.

As the SILVA Taxonomy does not generally offer annotations for eukaryotic sequences at higher resolutions (order and above), we chose to instead use annotations to the NCBI Taxonomy for this domain. Annotations were manually verified and selected in order to avoid sequences whose taxonomic annotations were in conflict with the topology of the alignment-based tree (see Methods for details). To facilitate the identification and classification of plastid and mitochondrial 16S sequences as indicators of eukaryotic organisms, these were placed in the *SilvaMod* taxonomy as sub-domains together with



**Figure 1. Overview of the resources of CREST.** The flow of information during the construction of a new reference database (top part) or classification (bottom part) is represented by arrows. The classification tools MEGAN or LCAClassifier can utilize CREST taxonomy files and databases such as SilvaMod for classification of environmental sequences, aligned to the reference database with Megablast.  
doi:10.1371/journal.pone.0049334.g001

the nuclear 18S rRNA directly under the eukaryotic domain (see box “Taxonomy” in Fig. 1), even though this strictly does not agree with the phylogenetic origin of these genes.

The resulting SilvaMod database contains 254,671 sequences from the SILVA SSUref nr release 106, along with curated taxonomical annotations including explicit rank information, formatted for MEGAN and the CREST LCAClassifier. Out of these, 72% are annotated to genus rank and 92% to at least family rank or better and the taxonomy contains in total 99 phyla and 39 candidate divisions. At higher resolution, there are 1,237 orders, 3,933 families and 9,906 genera including candidate divisions and environmental clades.

We also make available taxonomic annotations for the 2011 release of Greengenes [21], in the format accepted by MEGAN and the CREST LCAClassifier. This database contains more sequences (408,315) from a similar number of taxa at the phylum rank. However, the Greengenes taxonomy contains a much smaller number of taxa at higher resolutions, e.g. only 230 orders, 394 families and 1061 genera. One reason for this difference is the many eukaryotic taxa present only in SilvaMod.

### Classification Tools: MEGAN and the LCAClassifier

Alignment-based classification using LCA and minimum similarity filters (see Methods) was implemented in the CREST LCAClassifier and by extending MEGAN [20] (v4.68+). MEGAN uses a graphical user interface and can also export assignments, community composition and taxon-specific sequences in text format. In addition, the composition of two or more communities can be compared [29].

The CREST LCAClassifier uses a command line interface and reports community composition in a simple tab-separated text format allowing for overview of taxon abundance and richness (for amplicon sequences) at each taxonomic rank. Several alignment files, constructed using Megablast, can be classified simultaneously, which facilitates easy comparison between classification results from several datasets (by adding output for taxa present in at least one dataset to all). In addition, assignment information can be exported along with sequence data in FASTA format, or without it as comma-separated text. The CREST LCAClassifier is also available through a web interface at <http://apps.cbu.uib.no/crest> including the Megablast alignment step. The user simply uploads one or several sequence files in FASTA-format. A maximum of 1,000 sequences is currently enforced by the webserver, but exceptions may be granted on request.

Default values of LCA parameters were chosen conservatively based on cross-validation testing (see below). However, the appropriate parameters depend on the community studied and can be adjusted in the classification tools. See “Alignment-based classification – LCAClassifier” in Methods for a discussion of parameter choice.

In addition to LCA classification, we added a minimum similarity filter in order to decrease the false positive rate for “novel” or noisy sequences, with low similarity to reference sequences (see Methods). Using the CREST LCAClassifier, such sequences are flagged as “Unknown” members of the taxon to which they were assigned after filtering and can be retrieved from the FASTA- or assignment output data. Using MEGAN, or verbose output of the CREST LCAClassifier, information about all such assignments is written to the output dialog.

When analyzing data from amplicon libraries, an important first step is quality filtering followed by noise removal [30] or clustering [31], as well as chimera removal, in order to compensate for artifacts resulting from sequencing or PCR [32]. For pyrosequencing or IonTorrent sequence reads, we recommend using

AmpliconNoise for pre-processing as this can remove more sequence noise than other available programs [30,33] and ensures compatible annotation. Regardless of the method used, the result is a set of unique sequences, each representing a variable number of reads. For AmpliconNoise, the reads of each unique sequence are determined as likely to originate from identical nucleotide sequences. Alternatively, sequences representing similar reads in a cluster (OTU) can be submitted if the sequence names in the FASTA-file containing filtered sequences or OTU representatives are annotated with read abundance (using “*weight = N*” or “*\_N*”), MEGAN or the CREST LCAClassifier will report both the weighted read abundance and number of unique sequences (i.e. richness) for each taxon. In addition, the CREST LCAClassifier calculates a Chao-estimate [34] of minimum richness for each taxon.

### Cross Validation Testing

To evaluate the performance of CREST with SSU rRNA sequences, we performed two types of cross-validation testing (a technique partitioning the reference dataset into subsets used for re-training and validation). Firstly, exhaustive ten-fold cross validation was used and secondly, removal of whole genera, families and phyla (see Methods for details). Tests were repeated with randomly cropped sub-sequences derived from these with lengths 450 and 100 bps; the approximate read lengths from the GS FLX Titanium (pyrosequencing) and Illumina Hi-Seq platforms, two of the most commonly used methods for high-throughput sequencing. Results from Megablast alignment followed by CREST LCAClassifier are hereafter referred to as “LCA”.

The same test regime was also carried out with the Greengenes database to compare effects of the two reference databases on classification results. To compare the performance with another popular method, the RDP Classifier [22], the same tests were carried out with this program using its default training dataset (v6/2.32). An alternative training dataset for the RDP Classifier using Greengenes was also evaluated, retraining the classifier via a QIIME script [35], designed intentionally to classify reads only to the family rank. This represents the recommended classification method for SSU rRNA QIIME.

Table 1 lists the resulting assignment accuracies (fraction of sequences classified correctly with default parameters) from ten-fold cross validation of the different classification strategies. SilvaMod with LCA performed best in five cases out of nine. Using the Greengenes database achieved slightly higher accuracy at family rank. Figure 2 shows the results from this test as precision-recall graphs, generated by varying the LCA range or confidence cut-off. This confirms that the CREST LCAClassifier was capable of classification with both a higher recall and precision, compared to the RDP Classifier. The RDP Classifier produced consistently higher false discovery rates (the fraction of all classifications made that were incorrect, or *1-precision*), up to 11% at the recommended bootstrap confidence cut-off at 0.8. For LCA it never reached above 3%. The minimum similarity filter contributes to reducing false assignments with about 30% at genus and family level for full-length sequences using the default LCA range.

Results from the second test, removal of whole taxa, are presented in Table 2. In this test, the RDP Classifier performed better for shorter sequences (100 bp) and for removal of whole phyla, whereas the CREST LCAClassifier performed better with longer sequences for removal of families or genera.

**Table 1.** Assignment accuracy from ten-fold cross validation.

Method	Training/ Reference set	Fragment length	Accuracy per rank <sup>a</sup>		
			Genus	Family	Phylum
LCA <sup>b</sup>	SilvaMod	F.L. <sup>d</sup>	<b>82%</b>	92%	<b>99.9%</b>
LCA <sup>b</sup>	SilvaMod	450 bp	62%	88%	<b>99.7%</b>
LCA <sup>b</sup>	SilvaMod	100 bp	<b>38%</b>	61%	<b>94%</b>
LCA <sup>b</sup>	Greengenes	F.L. <sup>d</sup>	69%	94%	99%
LCA <sup>b</sup>	Greengenes	450 bp	48%	87%	99%
LCA <sup>b</sup>	Greengenes	100 bp	33%	<b>65%</b>	<b>94%</b>
RDP <sup>c</sup>	Greengenes	F.L. <sup>d</sup>	–	<b>97%</b>	98%
RDP <sup>c</sup>	Greengenes	450 bp	–	<b>94%</b>	95%
RDP <sup>c</sup>	Greengenes	100 bp	–	49%	51%
RDP <sup>c</sup>	RDP v6	F.L. <sup>d</sup>	81%	95%	99%
RDP <sup>c</sup>	RDP v6	450 bp	<b>73%</b>	92%	98%
RDP <sup>c</sup>	RDP v6	100 bp	35%	56%	90%

<sup>a</sup>Assignment accuracy defined as number of correct assignments divided by the total number of sequences tested, given at three different ranks. The best values for each combination of rank and fragment length are indicated in bold.

<sup>b</sup>Classification using Megablast alignments and the CREST LCAClassifier within a 2% LCA range of the highest bitscore as well as percent similarity filters.

<sup>c</sup>Naïve Bayes classification using the RDP Classifier with a bootstrap of 0.8. With the Greengenes training set, RDP Classifier was run via the QIIME script `assign_taxonomy`, which does not classify sequences beyond the family level.

<sup>d</sup>Un-cropped full-length sequences from the reference or training dataset.  
doi:10.1371/journal.pone.0049334.t001

### Performance with Environmental Datasets

In order to evaluate the reference/training datasets and classification methods on “real life” environmental SSU rRNA datasets, the number of identified taxa and the proportion of classified reads were compared, using four different datasets (see Table 3). In addition to the approaches tested with cross validation, we also included the online SINA Aligner from SILVA in this test. Since submission to the online version of SINA was limited to a maximum of 500 sequences at the time of testing, the test could only be carried out with the one test dataset below this size. As opposed to the RDP Classifier, which utilizes nucleotide composition (“k-mers”), SINA is an alignment-based method like CREST. However, the sequences are compared directly to the reference alignment rather than using pairwise alignments.

SilvaMod with LCA performed best in terms of the share of reads classified for each environmental dataset and for each rank tested (phylum, family and genus), with only one exception (see Table 4). Compared to the RDP Classifier used with the default training, SilvaMod with LCA managed to assign on average about 50% more reads at family level. With this classification approach more taxa were also detected in each dataset and rank, in all cases but two (Table 4). Figure 3 shows the average share of classified reads in the four datasets, with SilvaMod and LCA showing the highest values across all ranks, followed with Greengenes, giving similar sensitivity with LCA and RDP Classifier assignment (except at genus level).

Classifications of the most abundant taxa agreed well, with a few exceptions. Sequences that were classified as *Sporichthyaceae* order *Frankinales* (21% abundance in the Lake Lanier dataset) using SilvaMod, were instead classified as *Actinomycetales* clade “ACK-M1” using Greengenes. Similarly *Oryzihumus* (22% in the Siberian soil dataset, according to SilvaMod) was classified as *Phycoccus* according to Greengenes.

### Resource Requirements and Execution Time

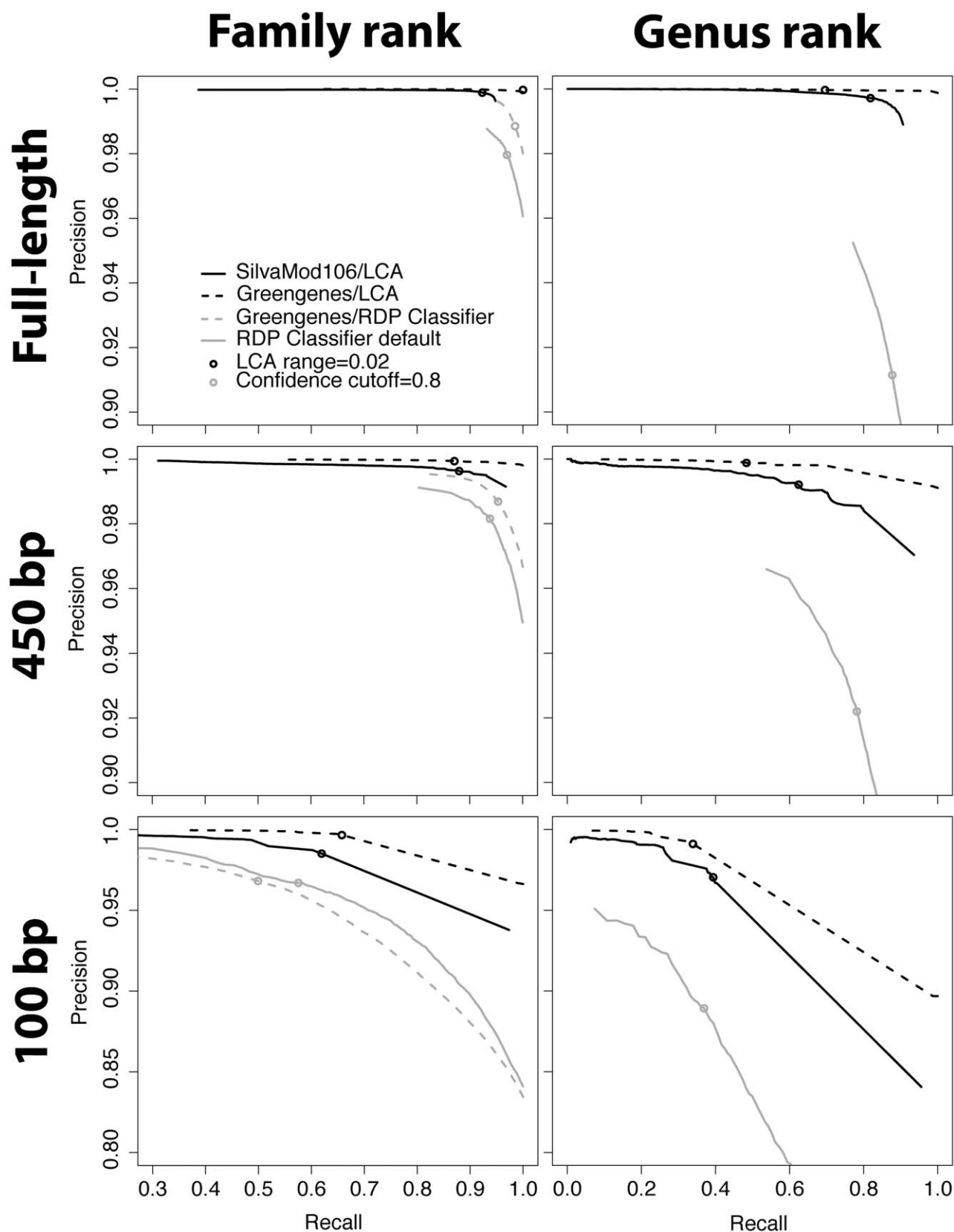
Running Megablast and LCA classification of 1,000 SSU rRNA sequences from the Forest Soil dataset took less than 12 minutes using a quad-core Apple MacBook Pro with a 2 GHz Intel i7. As a comparison, the RDP Classifier trained with Greengenes used 2 minutes on the same dataset and MacBook while the SINA Online Aligner used 8 minutes. Given that 1,000 sequences is a moderately high richness after AmpliconNoise analysis, this allows for classification in a very reasonable time of amplicon libraries sequenced with pyrosequencing or IonTorrent and subjected to AmpliconNoise prior to classification. Since typically only about 0.1% of shotgun metagenome sequences contain the SSU rRNA gene [4], this is also a very reasonable time for classification of such datasets. However, it requires a pre-screening to select the SSU rRNA containing sequences before classification, for which there are several alignment and compositional based tools available faster than Megablast, such as Metaxa [36] HoSeqI [37] and USEARCH [38].

### Discussion

Ten-fold cross validation tests indicate that CREST LCA Classification achieves better recall with higher precision compared to the RDP Classifier and SINA, regardless of reference database (SilvaMod or Greengenes). In general, CREST also gives a higher recall and precision when using the Greengenes database compared to SilvaMod at family rank. This could indicate that the Greengenes taxonomy is more robust, i.e. contain fewer polyphyletic or incorrectly affiliated taxa. On the other hand, the Greengenes reference database contains more sequences and fewer taxa. Ten-fold cross-validation is expected to underestimate the prediction accuracy for both small and more complex reference datasets, as the taxonomic redundancy will be lower, or in other words each taxon is represented by fewer sequences. Considering this bias, results indicate that both databases perform comparatively well using CREST LCA, whereas the RDP Classifier gives rise to more false positives, especially at genus rank.

Removal of whole taxa is less biased to differences in reference database size and showed contrasting results compared to ten-fold cross validation. As a consequence of lower sensitivity, the RDP Classifier always performed better when classifying shorter sequences. It also performed better for phylum removal, when using Greengenes as training dataset. This corresponds to a situation where a previously un-encountered clade at “phylum level” is discovered (SSU rRNA sequence similarity typically <85%). The alignment and LCA based method struggled with this, as partial alignments were often produced and thus an assignment to the closest related phylum was made. A majority of sequences with more realistic novelty are correctly identified as unknown by the minimum similarity filter of MEGAN or the CREST LCAClassifier. In addition, these can be extracted from the dataset for further phylogenetic study. We recommend caution when interpreting the existence of such taxa, however, particularly for amplicon sequences with low abundance (such as “singletons”) or for shotgun sequencing reads, as these can represent sequencing or PCR artifacts rather than true biological novelty.

Using actual environmental datasets, the CREST LCAClassifier with SilvaMod consistently demonstrated an ability to provide more detailed taxonomic classifications than the other approaches tested, in terms of both number of reads assigned and number of taxa detected. These results indicate a stronger sensitivity at all three taxonomic levels tested (phylum, family and genus), both quantitatively (number of reads classified) and qualitatively (number of taxa recognized). As demonstrated using cross-



**Figure 2. Precision-recall curves from ten-fold cross validation.** Shows the precision (number of correct assignments/number of assignments made) on the y-axis and measured recall (sensitivity or true positive rate) on the x-axis, when varying LCA range or confidence cutoff. Circles indicate the default cutoffs (cutoff for RDP=0.8, LCA range=0.02).  
 doi:10.1371/journal.pone.0049334.g002

**Table 2.** Assignment accuracy from removal-of-taxa cross validation.

Method	Training/Reference set	Fragment length	Accuracy <sup>a</sup> at removed rank level for removal of:		
			Genera	Families	Phyla
LCA <sup>b</sup>	SilvaMod	F.L. <sup>d</sup>	<b>98%</b>	90%	7%
LCA <sup>b</sup>	SilvaMod	450 bp	77%	64%	27%
LCA <sup>b</sup>	SilvaMod	100 bp	81%	66%	76%
LCA <sup>b</sup>	Greengenes	F.L. <sup>d</sup>	85%	<b>99.8%</b>	37%
LCA <sup>b</sup>	Greengenes	450 bp	<b>90%</b>	<b>85%</b>	24%
LCA <sup>b</sup>	Greengenes	100 bp	87%	72%	71%
RDP <sup>c</sup>	Greengenes	F.L. <sup>d</sup>	–	57%	<b>85%</b>
RDP <sup>c</sup>	Greengenes	450 bp	–	83%	<b>92%</b>
RDP <sup>c</sup>	Greengenes	100 bp	–	<b>99%</b>	<b>99%</b>
RDP <sup>c</sup>	RDP v6	F.L. <sup>d</sup>	62%	62%	21%
RDP <sup>c</sup>	RDP v6	450 bp	75%	78%	89%
RDP <sup>c</sup>	RDP v6	100 bp	<b>93%</b>	92%	96%

<sup>a</sup>Accuracy defined as number of correct assignments divided by the total number of sequences tested, given at three different ranks. The best values for each combination of rank and fragment length are indicated in bold.

<sup>b</sup>Classification using Megablast alignments and the CREST LCAClassifier within a 2% LCA range of the highest bitscore as well as percent similarity filters.

<sup>c</sup>Naïve Bayes classification using the RDP Classifier with a bootstrap confidence cutoff of 0.8. With the Greengenes training set, RDP Classifier was run via the QIIME script `assign_taxonomy`, which does not classify sequences beyond the family level.

<sup>d</sup>Un-cropped full-length sequences from the reference or training dataset.

doi:10.1371/journal.pone.0049334.t002

validation, incorrect assignments are not likely to have influenced this test. The fact that eukaryotic 18S rRNA sequences were only present in SilvaMod contributed to these results but cannot explain them alone; more bacterial and archaeal taxa were also consistently predicted. In general, SSU rRNA offers limited resolution as a taxonomic marker for eukaryotic sequences, particularly at species level, why internal transcribed spacers (ITS) or large subunit (LSU) rRNA is often utilized instead [39,40]. Similarly to our results using SSU rRNA, it has been shown that LCA classification as implemented in MEGAN yields better accuracy also for ITS [40] or LSU [41] rRNA sequences. CREST allows for the creation of such reference databases for the CREST LCAClassifier, from an alignment of ITS or LSU sequences.

The SINA Aligner performed significantly worse than all other methods on the one dataset it was tested on. However, the results of the comparison should be interpreted with some care; it may be that the default parameters used are more conservative than those used for the other methods. This is hard to estimate for SINA since it was not practically possible to include the method in the cross validation. It would have required the complete reference alignment to be re-built for each test for a fair comparison.

At the family rank, recall rates from the ten-fold cross-validation (Table 1) are roughly similar at family level to the fraction of sequences classified from the environmental amplicon datasets (Hydrothermal mat and Siberian soil) of corresponding lengths (450 bp and 100 bp, respectively). This indicates that the ten-fold cross-validation allows quite realistic testing of the methods at this resolution. At genus level, however, recall rates were consistently higher in the ten-fold cross-validation, probably because a large fraction of the reads in the environmental datasets belong to taxa that remain to be taxonomically described at this resolution.

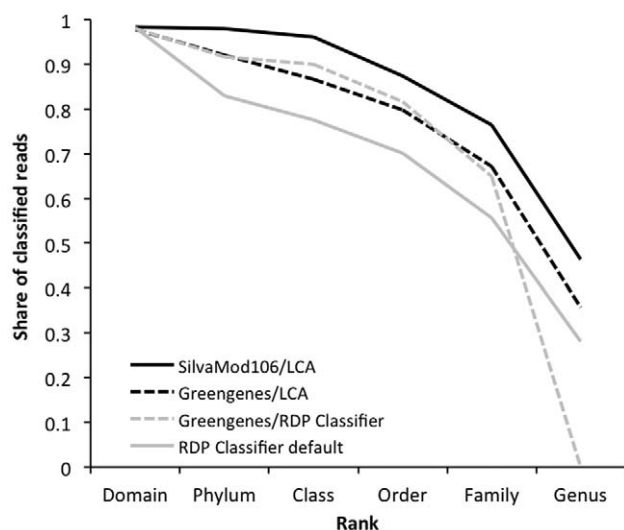
In conclusion, CREST provides for efficient and accurate taxonomic classification of environmental sequence tags, i.e. those containing a suitable taxonomic marker, such as SSU rRNA. We propose a classification scheme with Megablast used for alignment to the proposed SilvaMod reference database, and an extended algorithm for Lowest Common Ancestor classification, as implemented in MEGAN and the CREST LCAClassifier. This results in higher classification rates than with existing taxonomically annotated reference databases such as Greengenes. As shown using environmental datasets as well as cross-validation, it also outperforms the RDP Classifier, regardless of training dataset used, both in terms of recall and false positive rate.

**Table 3.** Datasets used for performance testing.

Dataset	Sequencing technology	Library type	Total SSU rRNA reads <sup>*</sup>
Lake Lanier	GS FLX Ti	Shotgun metagenome	558
Forest soil	GS FLX Ti	Shotgun metatranscriptome	51,202
Siberian soil	Illumina	16S rRNA amplicons	2,173
Hydrothermal mat	GS FLX Ti	16S rRNA amplicons	8,903

<sup>\*</sup>Reads with a BLASTN alignment bitscore >50 to a sequence in SilvaMod.

doi:10.1371/journal.pone.0049334.t003



**Figure 3. Average proportion of reads classified at different ranks in four environmental datasets.** The CREST LCAClassifier (analogous to MEGAN) was tested using the full SilvaMod and Greengenes [21] reference databases with their respective taxonomies, as well as the RDP Classifier [22] retrained with Greengenes (99%OTU dataset; executed via QIIME) and version 6 of the default RDP training dataset.

doi:10.1371/journal.pone.0049334.g003

In addition to classification with the SilvaMod reference database, CREST-compatible configuration files for the Greengenes database are available at the CREST website (<http://apps.cbu.uib.no/crest>). By using both databases for prokaryotic sequences, a best practice is ensured and differences can be identified and manually studied in more detail. Using the *nds2CREST* script distributed with the CREST LCAClassifier, new CREST reference databases can also be made from a taxonomically annotated alignment-based tree in ARB-format, from any taxonomic marker (e.g. *rpoB*, LSU rRNA or ITS).

## Methods

### Constructing the SilvaMod Taxonomy and Reference Database

The non-redundant SILVA SSURef release 106 was downloaded in ARB-format from the SILVA website at <http://www.arb-silva.de>. Using the ARB software package [23], we removed all sequences with a pintail score below 75, alignment quality score below 75 or length below 1,200 bp, in order to retain only high quality sequences. Further, we revised the taxonomy of several bacterial and archaeal taxa. The most significant improvements update the taxonomy of the Archaea to include the proposed phylum *Thaumarchaeota* [42,43], the *Actinobacteria* to comply with Bergey's Taxonomic Outline [44], the *Acidobacteria* to incorporate proposed subgroups [45] and the *Cyanobacteria* to comply with the CyanoDB [46] and in some cases specific studies (details given in Supplementary Table S1). Other added taxa include the *Zetaproteobacteria* [47], *Rubritaleaceae* [27] and *Armatimonadetes* [48].

**Table 4. Results from performance testing using environmental datasets.**

Method	Training/Reference set	Dataset	Share of reads assigned <sup>a</sup>			Unique taxa (B+A+E) <sup>b</sup>		
			Genus	Family	Phylum	Genera	Families	Phyla
LCA <sup>c</sup>	SilvaMod	Lake Lanier	<b>36.2%</b>	<b>73.7%</b>	<b>99.5%</b>	<b>31+0+1</b>	<b>45+0+2</b>	<b>11+0+2</b>
LCA <sup>c</sup>	SilvaMod	Forest soil	<b>30.4%</b>	<b>69.8%</b>	<b>99.1%</b>	<b>232+0+166</b>	<b>156+1+167</b>	<b>29+2+46</b>
LCA <sup>c</sup>	SilvaMod	Siberian soil	31.2%	<b>69.5%</b>	<b>93.6%</b>	<b>51+1+0</b>	<b>81+1+0</b>	<b>20+1+0</b>
LCA <sup>c</sup>	SilvaMod	Hydrothermal mat	<b>87.5%</b>	<b>93.0%</b>	<b>99.6%</b>	<b>36+2+1</b>	<b>42+8+1</b>	19+2+1
LCA <sup>c</sup>	Greengenes	Lake Lanier	11.5%	64.5%	98.9%	15+0+0	25+0+0	<b>13+0+0</b>
LCA <sup>c</sup>	Greengenes	Forest soil	14.7%	55.1%	84.1%	130+0+0	126+1+0	31+2+2
LCA <sup>c</sup>	Greengenes	Siberian soil	<b>39.0%</b>	60.1%	85.6%	38+1+0	53+1+0	18+1+0
LCA <sup>c</sup>	Greengenes	Hydrothermal mat	77.5%	89.0%	99.4%	15+1+0	23+6+0	<b>21+2+1</b>
RDP <sup>d</sup>	Greengenes	Lake Lanier	0	72.2%	91.8%	0	28+0+0	9+0+0
RDP <sup>d</sup>	Greengenes	Forest soil	0	52.2%	86.7%	0	111+0+0	16+2+1
RDP <sup>d</sup>	Greengenes	Siberian soil	0	53.4%	90.5%	0	53+1+0	10+1+0
RDP <sup>d</sup>	Greengenes	Hydrothermal mat	0	81.6%	97.8%	0	19+3+0	9+2+0
RDP <sup>d</sup>	RDP v6	Lake Lanier	9.3%	51.1%	87.1%	17+0+0	20+0+2	10+0+2
RDP <sup>d</sup>	RDP v6	Forest soil	11.9%	40.4%	80.9%	176+2+0	95+2+0	20+2+1
RDP <sup>d</sup>	RDP v6	Siberian soil	6.7%	39.7%	66.0%	36+1+0	39+1+0	10+1+0
RDP <sup>d</sup>	RDP v6	Hydrothermal mat	84.4%	91.7%	97.7%	21+2+0	17+2+0	8+2+0
SINA <sup>e</sup>	SSURef108	Hydrothermal mat	20.4%	27.7%	93.2%	32+1+0	25+5+0	9+2+0

<sup>a</sup>Proportion of the total reads in the dataset for which taxonomical assignment was achieved at the given taxonomical level.

<sup>b</sup>Number of unique taxa identified given separately for bacteria + archaea + eukaryotes. Where the highest total number of taxa was predicted from a test dataset, the number is indicated in bold.

<sup>c</sup>Classification using Megablast alignments and the CREST LCAClassifier within a 2% LCA range of the highest bitscore as well as percent similarity filters.

<sup>d</sup>Naïve Bayes classification using the RDP Classifier with a bootstrap confidence cutoff of 0.8. With the Greengenes training set, RDP Classifier was run via the QIIME script `assign_taxonomy`.

<sup>e</sup>LCA classification based on SINA Aligner, using default parameters at SILVA website.

doi:10.1371/journal.pone.0049334.t004



In addition, we identified a number of taxa whose taxonomic annotation disagreed strongly with the topology of the SSURef alignment-based tree and appeared poorly supported by phylogenetic studies. These were either re-assigned to existing parent taxa or novel ones labeled *incertae sedis*. Unique taxon names were always used and to this end we added the name of the only child taxon to several unlabeled or undetermined taxa, or removed them.

Annotations of the eukaryotic taxa using the NCBI Taxonomy were taken from the SSURef database and manually verified in order to remove all sequences where taxonomical affiliation was in clear conflict with the topology of the alignment-based tree. Selection of fungal reference sequences was done according to recent phylogenetic work [49,50].

All manual changes are listed in Supplementary Table S1, which can also be downloaded as a text file from <http://services.cbu.uib.no/supplementary/crest/and> is using an unambiguous format that can be parsed by the *nds2CREST* script (see below). In total, 82 new taxa were added, 123 were renamed and 17 deleted. All sequences remaining after curation were exported in FASTA format. During this procedure, sequences were cropped so as only the part corresponding to the SSU rRNA gene was saved. This was achieved by applying the *Escherichia coli* positional filter in ARB, selecting alignment column 1,000 and 43,183. A tab-separated text file listing the accession numbers and taxonomic placements of each sequence was exported (using “NDS export”).

We developed the python script *nds2CREST* distributed together with the CREST LCAClassifier in order to convert the exported sequence and taxonomic data from ARB into configuration files for MEGAN [20] and the CREST LCAClassifier. This script also reads a text version of the Manual Changes File (MCF; Supplementary Table S1). For each change specified in the MCF, it confirms that the change was properly carried out. In addition, the script removes all sequences without valid taxonomical annotation or specified to be removed in the MCF. After this procedure, it assigns taxonomic ranks for each taxon based primarily on the NCBI Taxonomy, where such information is available; secondarily on the name of the taxon using the suffixes “-ales” and “-aceae” to indicate family or order level, respectively; and lastly based on the parent rank. The output of *nds2CREST* is (1) a tree-file in Newick format describing the topology of the taxonomy, (2) a tab-separated “mapping file” specifying the name and rank for each taxon, and (3) a reference sequence database in FASTA-format. In addition to SilvaMod, we also prepared such files from the Greengenes Taxonomy [21] using the same procedure, however without manual curation or positional filtering.

### Alignment-based Classification - LCAClassifier

Taxonomical classification of environmental sequences starts with alignment to a reference sequence database (such as SilvaMod or Greengenes) using the NCBI *blastall* implementation of Megablast, with default settings except (optionally) restricting the output to 100 alignments to save disk space and calculation time, and deactivating the low complexity filter (the latter was not used during testing but has negligible impact on SSU rRNA alignments). BLASTN was also evaluated as an alternative, but as we did not notice an increased performance relative to the faster Megablast, we do not recommend it for SSU rRNA classification. The CREST LCAClassifier requires that Megablast output is saved in XML format, whereas MEGAN [20] can also parse the plain text output.

The classification is then carried out based on a subset of the best matching alignments using the Lowest Common Ancestor

(LCA) of this subset, as previously described in MEGAN [20]. Briefly, the subset includes sequences that score within  $x\%$  of the “bit-score” of the best alignment, providing the best score is above a minimum value. We selected a minimum bit-score of 155 and an LCA range ( $x$ ) of 2% as default parameters based on results from ten-fold cross-validation testing of SilvaMod, which resulted in relatively few false positives regardless of fragment length at the cost of slightly decreased recall. Lowering the LCA range increases the sensitivity at the cost of reduced precision, equivalent to moving to the right along the precision-recall curves of Figure 2, which can provide some guidance for selecting appropriate LCA range with different sequence lengths (note however, that the cross-validation testing only provides a rough indication of true precision and recall values). The appropriate LCA range also depends on the community studied. For example, the LCA range can be decreased to 1% if most sequences are similar to well-known type strains, or with longer read lengths (e.g. from Sanger sequencing or GS FLX+). Minimum bit-score has less effect on performance but we recommend increasing it when classifying amplicon sequences with longer read lengths, to e.g. 300 for FLX+.

The minimum similarity filter is based on a set of rank-specific requirements. Firstly, a sequence must be aligned with at least 99% nucleotide similarity to the best reference sequence in order to be classified to the species rank. For the genus, family, order, class and phylum ranks the respective cut-offs are 97%, 95%, 90%, 85% and 80%. These values were based on minimum similarities between closest neighbor SSU rRNA sequences inside the same taxa [51] then modified to further increase classification accuracy based on initial cross-validation testing. The filter ensures that classification is made to the taxon of the lowest allowed rank, effectively re-assigning sequences to parent taxa until allowed. Sequences with best-scoring alignments below the minimum bit-score are treated as unclassified and not analyzed by this filter.

The CREST LCAClassifier was implemented in Python (<http://www.python.org>) and can be executed on all major platforms.

### Performance Evaluation Using Cross-validation and Environmental Data

We performed exhaustive ten-fold cross-validation by randomly splitting the SilvaMod database into ten different sequence subsets of equal size. Each subset was then aligned to a concatenation of the other nine using Megablast and classified using the CREST LCAClassifier, in addition determining the LCA range at which the sequence could no longer be classified to each rank level. The default minimum bit-score (155) was also applied. Each test dataset was also randomly cropped into sub-sequences of 100 bp and 450 bp and the two resulting cropped subsets aligned. The same procedure was carried out for the Greengenes database and RDP Classifier default training dataset version 6 [22]. Instead of alignment, a re-training of the RDP classifier (v 2.3) was performed. Ten-fold cross-validation was also carried out on the Greengenes based training set *gg\_99\_otus\_4feb2011* using RDP Classifier v.2.2 through the QIIME (v.1.4.0; [35]) script *assign\_taxonomy.py*.

We also performed cross validation based on removal of whole genera, families and phyla. Testing was carried out using the four described reference (or training) datasets and was exhaustive, except for genera, where 100 genera were chosen randomly. The sequences of each taxon were aligned to a reference dataset with the sequences from that taxon missing. For the RDP Classifier tests, it was retrained in an analogous manner.

Assignments from cross validation tests were summarized using a custom python script and R (<http://r-project.org>). For each test, classification strategy, rank, and confidence cut-off (or LCA range), we calculated the number of true positives (*TP*; sequences correctly classified to a taxon), true negatives (*TN*; correctly unclassified), false positives (*FP*; incorrectly classified) and false negatives (*FN*; incorrectly unclassified). Assignment accuracy, precision and recall (sensitivity) was calculated using:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

In addition to cross-validation tests, the following four environmental datasets (summarized in Table 3) were used:

1. SSU rRNA gene-containing pyrosequencing reads from the metagenome of Lake Lanier [5],
2. SSU rRNA-containing pyrosequencing reads from the meta-transcriptome of an Austrian forest soil (unpublished),
3. Illumina-sequenced amplicon reads of the V4 region of 16S rRNA from Siberian tundra soil (unpublished), and
4. De-noised unique amplicon sequences of the V5–V6 region of 16S from a deep-sea hydrothermally associated microbial mat [4]

In addition to the mentioned classification strategies, dataset #4 was submitted to LCA-based classification with the SINA Aligner Online at the SILVA website ([www.arb-silva.de/aligner](http://www.arb-silva.de/aligner)) using default parameters and SSURef release 108 as reference alignment. Initially the Hidden Markov Model-based program SSuMMo [52] was also included in the comparison, using its

## References

1. Andersen K, Bird KL, Rasmussen M, Haile J, Breuning-Madsen H, et al. (2012) Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Mol Ecol* 21: 1966–1979.
2. Tringe SG, Hugenholtz P (2008) A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* 11: 442–446.
3. Urich T, Lanzén A, Qi J, Huson DH, Schleper C, et al. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* 3: e2527.
4. Lanzén A, Jørgensen SL, Bengtsson MM, Jonassen I, Øvreås L, et al. (2011) Exploring the composition and diversity of microbial communities at the Jan Mayen hydrothermal vent field using RNA and DNA. *FEMS Microbiol Ecol* 77: 577–589.
5. Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, et al. (2011) Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* 77: 6000–6011.
6. Glenn TC (2011) Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 11: 759–769.
7. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, et al. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 108 Suppl 1: 4516–4522.
8. Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD (2011) Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Appl Environ Microbiol* 77: 3846–3852.
9. Rodrigue Sb, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, et al. (2010) Unlocking short read sequencing for metagenomics. *PLoS ONE* 5: e11840.
10. Liu Z, DeSantis TZ, Andersen GL, Knight R (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* 36: e120.
11. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35: 7188–7196.
12. Cole JR, Chai B, Farris RJ, Wang Q, Syed-Mohideen ASK, et al. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* 35: D169–D172.
13. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72: 5069–5072.
14. Unterscher M, Jumpponen A, Opik M, Tedersoo L, Moora M, et al. (2011) Species abundance distributions and richness estimations in fungal metagenomics—lessons learned from community ecology. *Mol Ecol* 20: 275–285.
15. Pedrós-Alió C (2007) Ecology. Dipping into the rare biosphere. *Science* 315: 192–193.
16. Curtis TP, Head IM, Lunn M, Woodcock S, Schloss PD, et al. (2006) What is the extent of prokaryotic diversity? *Philos Trans R Soc Lond B Biol Sci* 361: 2023–2037.
17. Nilsson RH, Ryberg M, Sjökvist E, Abarenkov K (2011) Rethinking taxon sampling in the light of environmental sequencing. *Cladistics* 27: 197–203.
18. Wolf M, Müller T, Dandekar T, Pollack JD (2004) Phylogeny of Firmicutes with special reference to *Mycoplasmata* (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. *Int J Syst Evol Microbiol* 54: 871–875.
19. Rosen GL, Polikar R, Caseiro DA, Essinger SD, Sokhansanj BA (2011) Discovering the unknown: improving detection of novel species and genera from short reads. *J Biomed Biotechnol* 2011: 495849.
20. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17: 377–386.
21. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, et al. (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6: 610–618.
22. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73: 5261–5267.
23. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* 32: 1363–1371.

## Data Access

All resources including databases, taxonomy files and the source code for the CREST LCAClassifier and its web server are available under a GNU General Public Licence (v3) from <http://apps.cbu.uib.no/crest> and <http://lcaclassifier.googlecode.com>. Technical documentation describing how to install and use the program is available on the same websites. MEGAN can be downloaded from <http://ab.inf.uni-tuebingen.de/software/megan/along> with detailed documentation. All datasets used for testing can be downloaded from <http://services.cbu.uib.no/supplementary/crest/>.

## Supporting Information

**Table S1 Manual changes done during curation of the Silva Taxonomy (release 106) to SilvaMod.** (XLSX)

## Acknowledgments

We would like to thank Antje Gittel, Jack Gilbert and Natasha I. de Leon for sharing and preparation of environmental datasets for testing.

## Author Contributions

Conceived and designed the experiments: AL TU SLJ IJ DHH. Performed the experiments: AL SLJ TU DHH SHG MG. Analyzed the data: AL TU SLJ IJ LØ. Wrote the paper: AL TU LØ IJ SHJ.

24. Garrity GM, Boone DR, Castenholz RW, editors (2001) *Bergey's Manual of Systematic Bacteriology, Volume One: The Archaea and the deeply branching and phototrophic Bacteria*. 2 ed. New York: Springer - Verlag.
25. Garrity GM, Brenner DJ, Krieg NR, Staley JR, editors (2005) *Bergey's Manual of Systematic Bacteriology, Volume Two: The Proteobacteria*. 2 ed. New York: Springer - Verlag.
26. Vos PD, Garrity GM, Jones D, Krieg NR, Ludwig W, et al (2009) *Bergey's Manual of Systematic Bacteriology, Volume Three: The Firmicutes*. 2 ed. New York: Springer - Verlag.
27. Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, et al (2010) *Bergey's Manual of Systematic Bacteriology, Volume Four: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*. 2 ed. New York: Springer - Verlag.
28. Euzéby JP (1997) List of Prokaryotic names with Standing in Nomenclature 1997. Available: <http://www.bacterio.cict.fr/>
29. Mitra S, Klar B, Huson DH (2009) Visual and statistical comparison of metagenomes. *Bioinformatics* 25: 1849–1855.
30. Quince C, Lanzén A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12: 38.
31. Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* 12: 1889–1898.
32. Schloss PD, Gevers D, Westcott SL (2011) Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS ONE* 6: e27310.
33. Lee CK, Herbold CW, Polson SW, Wommack KE, Williamson SJ, et al. (2012) Groundtruthing Next-Gen Sequencing for Microbial Ecology - Biases and Errors in Community Structure Estimates from PCR Amplicon Pyrosequencing. *PLoS ONE* 7: e44224.
34. Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43: 783–791.
35. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7: 335–336.
36. Bengtsson J, Eriksson KM, Hartmann M, Wang Z, Shenoy BD, et al. (2011) Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie Van Leeuwenhoek* 100: 471–475.
37. Arigon AM, Perrière G, Gouy M (2008) Automatic identification of large collections of protein-coding or rRNA sequences. *Biochimie* 90: 609–614.
38. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.
39. Santamaria M, Fosso B, Consiglio A, Caro GD, Grillo G, et al. (2012) Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform*: doi: 10.1093/bib/bbs1036.
40. Porter TM, Golding GB (2011) Are similarity- or phylogeny-based methods more appropriate for classifying internal transcribed spacer (ITS) metagenomic amplicons? *New Phytol* 192: 775–782.
41. Porter TM, Golding GB (2012) Factors that affect large subunit ribosomal DNA amplicon sequencing studies of fungal communities: classification method, primer choice, and error. *PLoS ONE* 7: e35749.
42. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* 6: 245–252.
43. Pester M, Schleper C, Wagner M (2011) The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Curr Opin Microbiol* 14: 300–306.
44. Ludwig W, Euzéby JP, Schumann P, Busse H-J, Trujillo ME, et al. (2011) *Bergey's Taxonomic Outlines Volume Five: Road map of the Actinobacteria: Bergey's Manual Trust*.
45. Zimmermann J, Gonzalez JM, Saiz-Jimenez C, Ludwig W (2005) Detection and Phylogenetic Relationships of Highly Diverse Uncultured Acidobacterial Communities in Altamira Cave Using 23S rRNA Sequence Analyses. *Geomicrobiol J* 22: 379–388.
46. Komárek J, Hauer T (2011) *CyanoDB.cz - On-line database of cyanobacterial genera*. - Word-wide electronic publication, Univ. of South Bohemia & Inst. of Botany AS CR.
47. Emerson D, Rentz JA, Lilburn TG, Davis RE, Aldrich H, et al. (2007) A novel lineage of proteobacteria involved in formation of marine Fe-oxidizing microbial mat communities. *PLoS ONE* 2: e667.
48. Lee KCY, Dunfield PF, Morgan XC, Crowe MA, Houghton KM, et al. (2011) *Chthonomonas calidirosea* gen. nov., sp. nov., an aerobic, pigmented, thermophilic micro-organism of a novel bacterial class, Chthonomonadetes classis nov., of the newly described phylum Armatimonadetes originally designated candidate division OP10. *Int J Syst Evol Microbiol* 61: 2482–2490.
49. Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, et al. (2007) A higher-level phylogenetic classification of the Fungi. *Mycol Res* 111: 509–547.
50. Matheny PB, Curtis JM, Hofstetter Vr, Aime MC, Moncalvo JM, et al. (2006) Major clades of Agaricales: a multilocus phylogenetic overview. *Mycologia* 98: 982–995.
51. Cole JR, Konstantinidis K, Farris RJ, Tiedje JM (2010) Microbial Diversity and Phylogeny: Extending from rRNAs to Genomes. In: Liu WT, Jansson JK, editors. *Environmental Molecular Microbiology*. Wymondham: Caister Academic Press. 1–19.
52. Leach ALB, Chong JPJ, Redeker KR (2012) SSuMMo: rapid analysis, comparison and visualization of microbial communities. *Bioinformatics* 28: 679–686.