Genome-reconstruction for eukaryotes from complex natural microbial communities

Patrick T. West, Alexander J. Probst, Igor V. Grigoriev, et al.

Genome Res. 2018 28: 569-580 originally published online March 1, 2018
Access the most recent version at doi:10.1101/gr.228429.117

Method

Genome-reconstruction for eukaryotes from complex natural microbial communities

Patrick T. West,[1] Alexander J. Probst,[2,6] Igor V. Grigoriev,[1,3] Brian C. Thomas,[2] and Jillian F. Banfield[2,4,5]

[1]*Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA;* [2]*Department of Earth and Planetary Science, University of California, Berkeley, California 94709, USA;* [3]*US Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA;* [4]*Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720, USA;* [5]*Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

Microbial eukaryotes are integral components of natural microbial communities, and their inclusion is critical for many ecosystem studies, yet the majority of published metagenome analyses ignore eukaryotes. In order to include eukaryotes in environmental studies, we propose a method to recover eukaryotic genomes from complex metagenomic samples. A key step for genome recovery is separation of eukaryotic and prokaryotic fragments. We developed a *k*-mer-based strategy, EukRep, for eukaryotic sequence identification and applied it to environmental samples to show that it enables genome recovery, genome completeness evaluation, and prediction of metabolic potential. We used this approach to test the effect of addition of organic carbon on a geyser-associated microbial community and detected a substantial change of the community metabolism, with selection against almost all candidate phyla bacteria and archaea and for eukaryotes. Near complete genomes were reconstructed for three fungi placed within the Eurotiomycetes and an arthropod. While carbon fixation and sulfur oxidation were important functions in the geyser community prior to carbon addition, the organic carbon-impacted community showed enrichment for secreted proteases, secreted lipases, cellulose targeting CAZymes, and methanol oxidation. We demonstrate the broader utility of EukRep by reconstructing and evaluating relatively high-quality fungal, protist, and rotifer genomes from complex environmental samples. This approach opens the way for cultivation-independent analyses of whole microbial communities.

[Supplemental material is available for this article.]

Microbial eukaryotes are important contributors to ecosystem function. Gene surveys or DNA "barcoding" are frequently used to identify eukaryotes in microbial communities and have demonstrated the breadth of eukaryotic diversity (Pawlowski et al. 2012). However, these approaches can only detect species and are unable to provide information about metabolism or lifestyle in the absence of sequenced genomes. The majority of fully sequenced eukaryotic genomes are from cultured organisms. Lack of access to cultures for a wide diversity of protists and some fungi detected in gene surveys has resulted in major gaps in eukaryotic reference genome databases (Caron et al. 2008; Pawlowski et al. 2012). Single-cell genomics holds promise for sequencing uncultured eukaryotes and has generated partial genomes for some (Cuvelier et al. 2010; Yoon et al. 2011; Monier et al. 2012; Vaulot et al. 2012; Roy et al. 2014; Mangot et al. 2017). However, multiple displacement amplification limits the completeness of single-cell genomes (Woyke et al. 2010). Alternatively, metagenomic sequencing reads from environmental samples are mapped against reference genomes to detect organisms and constrain metabolisms, but this approach is restricted to study of organisms with sequenced relatives.

Many current studies of natural ecosystems and animal- or plant-associated microbiomes use an untargeted shotgun sequencing approach. When the DNA sequences are assembled, tens of thousands of genome fragments may be generated, some of which derive from eukaryotes. Exceedingly few metagenomic studies have systematically identified such fragments as eukaryotic, although some genomes for microbial eukaryotes have been reconstructed (Sharon et al. 2013; Kantor et al. 2015, 2017; Quandt et al. 2015; Mosier et al. 2016; Raveh-Sadka et al. 2016). In almost all cases, these genomes were recovered from relatively low-diversity communities where binning of genomes is typically less challenging than in complex environments. Here, we applied a new *k*-mer-based approach for identification of assembled eukaryotic sequences in data sets from diverse environmental samples. Identification of eukaryotic genome fragments enabled their assignment to draft genomes and improvement of the quality of gene predictions. Predicted genes on assembled metagenomic contigs provide critical inputs for further binning decisions that incorporate phylogenetic profiles as well as classification of the reconstructed genomes and assessment of their completeness. Our analyses focused on biologically diverse environmental samples, many of which came from groundwater. In addition, we investigated previously published metagenomes from infant fecal samples and a bioreactor community used to break down thiocyanate. Because the approach works regardless of a predetermined phylogenetic affiliation, it is now possible to reconstruct genomes for higher eukaryotes as well as fungi and protists from complex environmental samples.

[6]Present address: Group for Aquatic Microbial Ecology, Biofilm Center, Department of Chemistry, University of Duisburg-Essen, 45141 Essen, Germany
Corresponding author: Jbanfield@berkeley.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.228429.117.

Results

Crystal Geyser community structure

The deep subsurface microbial community at Crystal Geyser, Utah has been well characterized as being dominated by chemolithoautotrophic bacteria and archaea, including many organisms from candidate phyla (CP) (Probst et al. 2014, 2016; Emerson et al. 2015). It is our current understanding that a wide diversity of novel bacteria and archaea are brought to the surface by geyser eruptions (Probst et al. 2018). Such deep sedimentary environments are unlikely to have high organic carbon compound availability. Thus, we hypothesized that organic carbon addition to this system would profoundly shift the community composition by selecting against the novel geyser microorganisms and enriching for better known heterotrophs. To test this prediction, we analyzed a sample of wood that was added to the shallow geyser and had decayed in the groundwater conduit (hereafter referred to as CG_WC). This sample and a wood-free sample (CG_bulk) that was collected the day before CG_WC were subjected to metagenomic analysis. We identified 124 and 316 distinct strains in the CG_WC and CG_bulk samples, respectively. The CG_WC sample contained abundant eukaryotic sequences (Fig. 1A) that were not present in the surrounding geyser water (Fig. 1B). Twelve strains were present in both samples (Fig. 1C), including the archaeon *Candidatus* "Altiarchaeum hamiconexum" (Probst et al. 2014), which dominated the CG_bulk sample. A phylum-level comparison of the microbial communities is presented in Figure 1D. The presence of decaying wood strongly enriched for Actinobacteria and Proteobacteria, as well as eukaryotes such as Ascomycota, Basidiomycota, and an organism classified as part of the Arthropoda. A low abundance alga from the class Bacillariophyta was detected in both samples.

As predicted, the CG_WC sample contains very few CP bacteria and archaea, with the notable exception of three members of Saccharibacteria (TM7). Two Saccharibacteria genomes were >90% complete, and one 1.01 Mbp genome was circularized and curated to completion. To evaluate for the accuracy of the complete genome, we ruled out the presence of repeat sequences that could have confounded the assembly and carefully checked the consistency of paired reads mapped across the entire genome (Supplemental Data 1). The cumulative GC skew was used to identify the origin and terminus of replication (Brown et al. 2016). Although the skew has generally the expected form (consistent with genome accuracy), the origin defined based on GC skew was offset from the *dnaA* gene by ~46 kbp (Supplemental Fig. S1A). Short repeat sequences often associated with the origin were absent both from the predicted origin

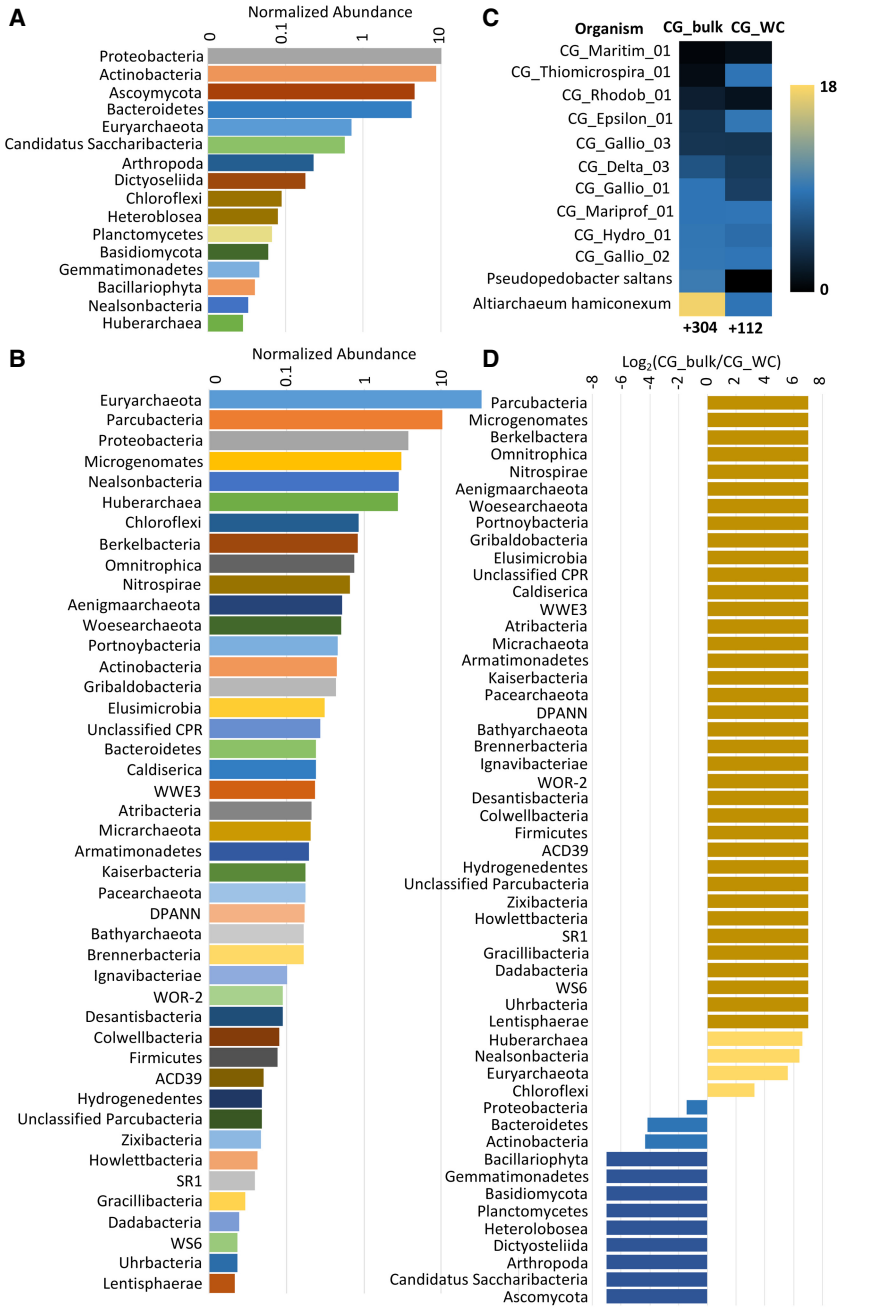

Figure 1. Comparison of CG_WC and CG_bulk community composition. The relative abundances of taxonomic groups in CG_WC (*A*) and CG_bulk (*B*) are depicted. Abundance was determined as the average coverage depth of the scaffolds containing annotated ribosomal protein S3 (*rpS3*) genes. Abundances were normalized for comparison across samples by multiplying the average coverage depth by the sample read count and read length. (*C*) Normalized coverage of *rpS3* containing scaffolds of strains common to both samples. The number of additional strains detected in each sample is listed *below* the respective sample heat map. (*D*) Log_2 ratio of normalized coverage of taxonomic groups from *A* and *B*. Taxonomic groups identified in only one sample are indicated by the darker yellow and blue bars.

and the region encoding *dnaA*, although they were identified close to the origin for another candidate phyla radiation bacterium (Anantharaman et al. 2016). We identified the origin region for a previously reported complete Saccharibacteria RAAC3_TM7 genome using cumulative GC skew and showed that repeats were not present in this genome either and that the predicted origin is 7.6 kb from the *dnaA* gene (Kantor et al. 2013).

EukRep tested on reference data sets

Typically, only prokaryotic gene prediction is performed on metagenomic samples, as these are the only algorithms specifically designed for this application (e.g., MetaProdigal) (Hyatt et al. 2012). For samples containing both prokaryotic and eukaryotic DNA, such as CG_WC, obtaining high-quality gene predictions for eukaryotes is complicated by the fact that distinct gene prediction tools are used for prokaryotic vs. eukaryotic sequences due to differences in gene structure. Specifically, eukaryote genomes have more complex promoter regions, regulatory signals, and genes spliced into introns and exons, variable between species. For this reason, it is not surprising that we found that prokaryotic gene predictors underperform when used on eukaryotic sequences. This can impact binning by affecting taxonomic profiling of scaffolds and bin quality metrics such as the presence or absence of single-copy genes (Supplemental Fig. S2). To address this issue and obtain high-quality eukaryotic gene predictions from metagenomes, we present EukRep, a classifier that utilizes *k*-mer composition of assembled sequences to identify eukaryotic genome fragments prior to gene prediction (Fig. 2A). When previously used to taxonomically classify metagenomic sequences, machine learning algorithms have shown promise, but their success was limited when samples contained many different species (Vervier et al. 2016). We hypothesized that a supervised classification method could be applied to accurately classify sequences at the domain level for gene prediction purposes, avoiding complications from having a large number of taxonomic categories.

The EukRep model was trained using a diverse reference set of bacterial, archaeal, opisthokonta, and protist genomes (3.40 Gbps of sequence) (Supplemental Table S1). The *k*-mer frequencies were calculated for each 5-kb interval, resulting in 581,376 individual instances that were used to train a linear-SVM (scikit-learn) (Pedregosa et al. 2011). We found that 5-mer frequencies represented the best compromise between speed and accuracy for classifying eukaryotic scaffolds and that sequences can be classified with high accuracy at lengths of 3 kb or greater (Fig. 2B; Supplemental Fig. S3). A validation set of 486 independent genomes (Supplemental Table S2) was assembled to test the prediction power of EukRep. An important goal of EukRep is to be able to classify novel as well as known eukaryotic sequences and to avoid overfitting for existing eukaryotic sequences. Thus, the

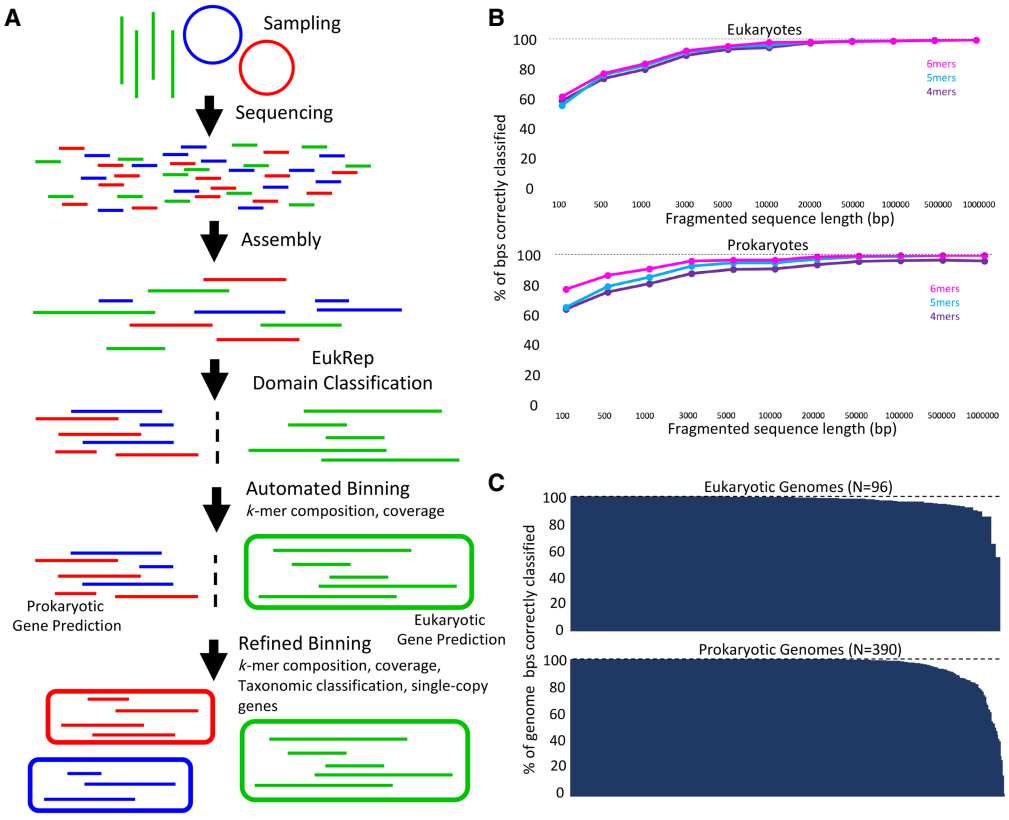

Figure 2. Identification of scaffolds for eukaryotic gene prediction with EukRep. (*A*) Schematic of the analysis pipeline used to identify and bin both eukaryotic and prokaryotic genomes within this paper. (*B*) A subset of genomes from Supplemental Table S2 was used to compare prediction accuracy of linear-SVM models trained on *k*-mer frequencies of *k*-mers ranging in length from 4 to 6 bp. For each sequence size category, sequences longer than the specified length were fragmented to the specified length and sequences shorter were excluded. (*C*) Accuracy of EukRep domain prediction on a per-genome level for both eukaryotes and prokaryotes. Percent of the genome correctly classified is defined as the percent of base pairs within a given genome predicted to belong to the genome's known domain. Each bar represents the percent of a single genome that was classified correctly. Genomes used for training and testing of EukRep along with their prediction results are listed in Supplemental Tables S1 and S2.

training and validation sets were chosen so as to taxonomically overlap at a maximum of genus level. Using the described validation set to test EukRep, we found that the classifier was able to accurately predict the domain of 97.5% of total tested eukaryotic sequence length and 98.0% of prokaryotic sequence length.

An important note is that EukRep is designed so as to miss as little eukaryotic sequence as possible. To ensure this, the program classifies every sequence in a sample, even sequences whose composition signals will be weak because the sequences are relatively short. Further, given the continuum between confident and less confident classification of eukaryote sequences, we chose settings that maximized classification outcomes (recall). The 2% of incorrect classifications of prokaryote as eukaryote sequences represent false positives that can be removed using standard binning methods (especially those that include phylogenetic signal).

We examined classifier accuracy on a per-genome basis to test whether the classifier performance varied for organisms of widely different types (Fig. 2C). This metric differs from that reported above because it refers to the accuracy of classifying individual artificially fragmented genomes rather than overall accuracy on all scaffolds tested from every genome. Ninety-four percent of tested eukaryotic genomes were classified with >90% accuracy, whereas 88% of tested prokaryotic genomes were classified with >90% accuracy. In a small number of prokaryotic genomes, more than half of the contigs were misclassified as eukaryotic. Notably, all of these were small genomes of organisms inferred to be parasites or symbionts. However, almost all of the sequences composing the eukaryotic genomes tested were correctly classified, indicating this method can successfully identify scaffolds whose analysis would benefit from a eukaryotic gene prediction algorithm.

In a complex metagenomic sample, obtaining sequences from novel lineages is a relatively common occurrence, and EukRep's ability to classify novel eukaryotic sequences is critical. We tested the ability of EukRep to do this by having it classify both eukaryotes ($n = 18$) and prokaryotes ($n = 46$) from phyla not represented in EukRep's training set (Supplemental Fig. S3). Although the genomes were fragmented into 3-kb pieces, EukRep maintained an overall accuracy of 90%. When tested on sequences fragmented to 20 kb, accuracy improved to 98%. Thus, we conclude that EukRep can be relied upon to correctly classify the majority of genomes from potentially entirely new phyla, even when fragmented.

Other taxonomic binning algorithms such as *taxator-tk* (Dröge et al. 2015) rely upon alignment to reference databases to make taxonomic classifications. Although these algorithms are typically designed for classifying reads at the lowest taxonomic level possible (e.g., species), they can potentially classify scaffolds at the domain level and perform the same function as EukRep. In order to test whether EukRep represents a significant improvement in this specific area, we compared EukRep to *taxator-tk* by classifying genomes from phyla unrepresented in EukRep's training set at the domain level. *taxator-tk* was selected for comparison because it includes eukaryotes in its prebuilt reference data set. *taxator-tk* was run twice. In the first test, many of the fragments to be classified were present as genomes in the reference data set (11/18), and it classified 47% of the total eukaryotic sequence tested as eukaryotic at 3 kb and 76% at 20 kb (Supplemental Fig. S3). In the second test, where the test genomes were removed from the reference set at the genus level so that the fragments represented genomes from novel genus level organisms at a minimum, the tool classified 24% at 3 kb and 44% at 20 kb of total eukaryotic sequence as eukaryotic (Supplemental Fig. S3). Due to the fact that EukRep does not rely upon alignment-based methods, it also does not require a reference database and can process metagenomes quickly, at a rate of up to two Gbp an hour on a single core. Thus, we conclude that EukRep represents an improvement over this approach for the purpose of identifying scaffolds for eukaryotic gene prediction.

Testing eukaryotic gene predictions on reference genomes

Eukaryotic gene prediction algorithms rely on a combination of transcriptomic evidence or protein similarity (AUGUSTUS [Stanke et al. 2006]; SNAP [Korf 2004]) and sequence signatures (GeneMark-ES [Ter-Hovhannisyan et al. 2008]) to make predictions. Given the frequent lack of sequenced close relatives to organisms identified in metagenomes and the lack of transcript data in many metagenomic studies, we tested how well eukaryotic gene predictors function in a diversity of eukaryotic genomes without transcriptomic evidence or homology evidence from close relatives. We applied the MAKER2 pipeline (Holt and Yandell 2011) with GeneMark-ES in self-training mode along with AUGUSTUS trained using BUSCO (Simão et al. 2015) to nine diverse eukaryotic genomes obtained from JGI's portal (Grigoriev et al. 2011) and NCBI's genome database (Fig. 3A; NCBI Resource

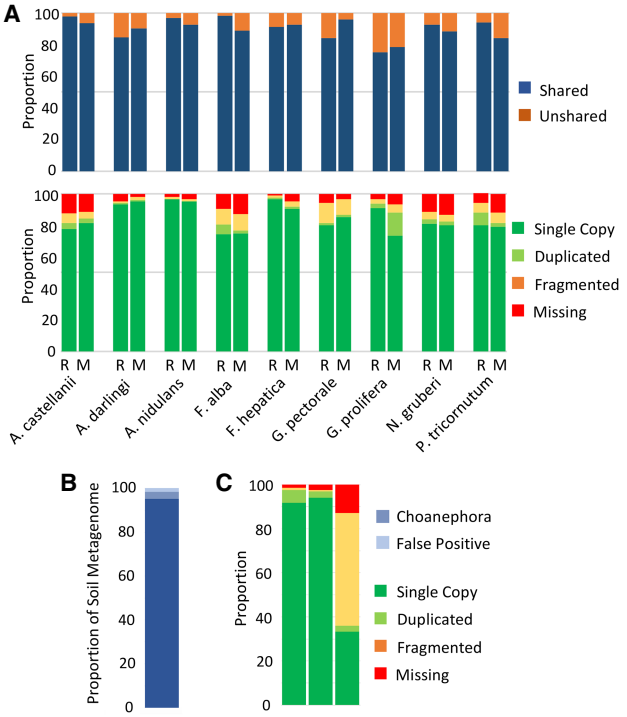

Figure 3. Eukaryotic gene prediction on metagenomic scaffolds. (*A*) Gene predictions for nine diverse eukaryotic organisms including fungi, a Metazoa, a Stramenopile, an Archaeplastida, and a Rhizaria. Columns labeled "R" refer to reference gene sets, whereas M columns refer to gene sets predicted without transcript or close homology evidence. The *top* panel displays the proportion of total genes either overlapping (shared) or not overlapping (unshared) a gene model from the other respective gene set for a given genome. The *bottom* panel is an analysis of presence or absence of single-copy genes in each gene set as determined by BUSCO using the eukaryota_odb9 lineage set. (*B*) Proportion of a soil metagenome spiked with the genome of *Choanephora cucurbitarum* predicted to be either noneukaryotic, eukaryotic and belonging to the *Choanephora*, or predicted to be eukaryotic but has homology to prokaryotic sequences. (*C*) BUSCO analysis of the binned *Choanephora cucurbitarum* genome with protein sets from (*left* to *right*) the reference protein set, trained MAKER2 output, and whole metagenome MetaProdigal output.

Coordinates 2017). The proteomes of *Chlamydomonas reinhardtii* (Merchant et al. 2007), *Neurospora crassa* (Galagan et al. 2003), and *Reticulomyxa filosa* (Glöckner et al. 2014) were also used as homology evidence. In each case, MAKER2-derived gene predictions were compared to reference gene predictions that incorporate transcriptomic evidence. The majority of the gene predictions identified without transcriptomic evidence were supported by reference gene predictions (78%–98%), and the majority of reference gene predictions overlapped a MAKER2-derived gene prediction (75%–98%). Estimated completeness of the predicted gene sets was measured by using BUSCO (Simão et al. 2015) to search for 303 eukaryotic single-copy orthologous genes within the predicted gene sets. The number of single-copy, duplicated, fragmented, and missing genes showed minimal differences with and without transcriptomic evidence (Fig. 3A). These results show the pipeline we assembled for eukaryotic gene prediction, even without transcriptomic evidence, is capable of detecting near complete gene sets similar to those from reference genomes, with the exception of untranslated regions and alternative splicing patterns.

To ensure that our proposed methodology can result in improved eukaryotic gene predictions in the context of a complex metagenomic sample, we spiked the genome of *Choanephora cucurbitarum* (Min et al. 2017) into a complex, 15-Gbp, soil shotgun metagenomic sample (Fig. 3B). The genome of *Choanephora cucurbitarum* was used because it is a fragmented draft (N50 = 24,238 bp) with scaffold lengths similar to what is often encountered in a metagenome and because it has gene models with many introns that would particularly benefit from eukaryotic gene prediction. EukRep was run on this mock data set and recovered 40.6 Mbp of sequence classified as eukaryotic. Of this, 26.6 Mbp were the *Choanephora* genome (91.6% of the entire genome, 99.6% of the genome longer than the 3-kb minimum sequence length cutoff). Next, 93.2% of the identified genome was placed into a single bin. Training and running eukaryotic gene predictors on this bin substantially improved gene predictions, increasing estimated completeness via single-copy genes from 36% to 97% (Fig. 3C). The gene models were substantially more similar to reference gene models in terms of total gene count and gene length than those predicted using MetaProdigal (Supplemental Fig. S4).

Analysis of newly reconstructed eukaryotic genomes

After benchmarking EukRep on reference data sets, the algorithm was applied to the CG_WC sample, and 214.8 Mbps of scaffold sequence was classified as eukaryotic. Because eukaryotic gene predictors are designed to be trained and run on a single genome at a time, CONCOCT (Alneberg et al. 2014), an automated binning algorithm, was applied to the identified eukaryotic scaffolds to generate two preliminary eukaryote genomes. In this way, GeneMark-ES and AUGUSTUS gene prediction could be performed, as described above, on each bin individually as if running on a single genome.

The availability of relatively confident gene predictions for eukaryotic contigs enabled re-evaluation of genome completeness based on the presence or absence of 303 eukaryotic single-copy genes as identified by BUSCO (Table 1; Fig. 4). An obvious finding was that one of the CONCOCT bins was a megabin. Using information about single-copy gene inventories, along with tetranucleotide frequencies, coverage, and GC content, we assigned the eukaryotic scaffolds into four genome bins. BLASTing gene predictions against UniProt identified three of the bins as likely fungi and a fourth as a likely metazoan. Gene prediction was redone on the new fungal bins with GeneMark-ES in self-training mode and AUGUSTUS trained with BUSCO. The bins ranged in size from 24.5 Mbps to 99.0 Mbps and encoded between 8947 and 18,440 genes. BUSCO single-copy orthologous gene analysis showed all four bins were relatively complete individual genomes based on gene content, with the lowest containing 243/303 (80%) and the highest containing 288/303 (95%) single-copy orthologous genes (Table 1; Fig. 4). Some genes expected to be in single copy were duplicated, as is often found with BUSCO analysis of complete genomes. The assembly quality of one bin, WC_Fungi_A, appeared to be quite high, with 50% of its sequences contained in scaffolds longer than 599 kb. We reduced potential contamination

Table 1. Summary of binned eukaryotic genome quality, contamination, and completeness

System	Sample	Name	Size (bp)	# genes	# scaffolds	N50	Completeness (%)
Crystal Geyser	CG_WC	CG_Fungi_A	24,984,438	8947	80	599,568	95
Crystal Geyser	CG_WC	CG_Fungi_B	26,644,854	10,026	1607	101,330	80
Crystal Geyser	CG_WC	CG_Fungi_C	24,500,285	9955	3654	16,121	80
Crystal Geyser	CG_WC	CG_Arthropod	99,046,889	18,440	8889	17,347	92
Crystal Geyser	CG_4_10_14_3.00	CG_Stremenopile	31,157,668	13,749	3424	11,782	77
Infant gut	182_001	Infant_A_Fungi_A	11,921,609	5160	265	66,296	94
Infant gut	182_002	Infant_A_Fungi_B	11,426,193	4999	334	45,092	92
Infant gut	N3_182_000G1	Infant_A_Fungi_C	11,895,925	5158	262	66,272	94
Infant gut	N1_023_000G1	Infant_B_Fungi_A	12,280,001	5293	1002	17,208	88
Infant gut	b023-d007	Infant_B_Fungi_B	12,603,413	5320	846	20,788	89
Infant gut	SP_CRL_000G1	Infant_C_Fungi_A	12,594,614	5309	885	22,402	87
Thiocyanite reactor	SCNpilot_cont_1000_bf	Rotifer_A	32,149,948	16,252	5055	6857	69
Thiocyanite reactor	SCNpilot_cont_1000_p	Rotifer_B	40,079,970	17,172	1961	25,209	83
Thiocyanite reactor	SCNpilot_cont_500_p	Rotifer_C	46,690,091	19,593	1702	39,424	87
Thiocyanite reactor	SCNpilot_cont_750_p	Rotifer_D	45,084,683	20,599	4908	11,284	87
Thiocyanite reactor	SCNpilot_expt_500_p	Rotifer_E	53,918,756	23,992	4626	15,939	89
Thiocyanite reactor	SCNpilot_expt_500_bf	Rotifer_F	52,830,794	23,942	5156	13,016	90
Thiocyanite reactor	SCN_pilot_cont_500_bf	Rotifer_G	59,551,575	24,973	3237	32,744	91
Thiocyanite reactor	SCNpilot_expt_500_p	SCNpilot_expt_500_p	22,299,472	10,815	3636	6549	69
Thiocyanite reactor	cn_treated	cn_treated	32,902,255	8342	559	77,947	85

Eukaryotic genomes identified from CG_WC, infant fecal-derived samples, and thiocyanate reactor samples are listed. Genome completeness is defined as the percent of BUSCO single-copy orthologous genes that were present either in a single copy or duplicated.

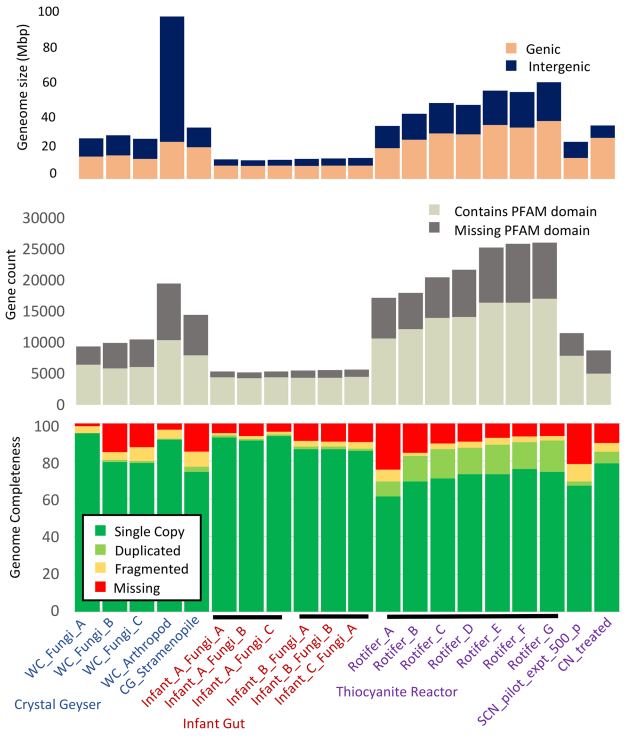

Figure 4. Overview of binned eukaryotic genomes. Genomes that share greater than 99% average nucleotide identity (ANI) are indicated by black bars. ANI comparisons are shown in more detail in Supplemental Figure S3. Genic regions refer to sequence located within predicted gene models whereas intergenic refers to all other sequence. Genes containing a PFAM domain were identified with PfamScan (Mistry et al. 2007). Genome completeness is measured as the percent of 303 eukaryotic single-copy orthologous genes found within a genome in a particular form with BUSCO.

of eukaryotic bins with prokaryotic sequence by BLASTing predicted proteins against UniProt and removing scaffolds with the majority of best hits belonging to prokaryotic genes.

A phylogenetic tree constructed from a set of 16 predicted, aligned, and concatenated ribosomal proteins (Hug et al. 2016) placed three of the bins within the fungal class Eurotiomycetes (Fig. 5). Each of these three bins ranged in size from 24.6 to 39.2 Mbps and in gene count from 8963 genes to 15,756 genes, within the range observed in previously sequenced Ascomycete fungi. The closest sequenced relative to all three bins was *Phaeomoniella chlamydospora,* a fungal plant pathogen known for causing Esca disease complex in grapevines (Morales-Cruz et al. 2015). The fourth bin, 99.7 Mbps in length and estimated to be 92% complete, was placed within the Arthropoda (Fig. 5). Its closest, although distant, sequenced relative is *Orchesella cincta* (Faddeeva-Vakhrusheva et al. 2016). *Orchesella cincta* is a member of the hexapod subclass Collembola (springtails), a diverse group basal to insects known primarily to be detritivorous inhabitants of soil. Although ribosomal protein S3 (*rpS3*) sequences belonging to Dictyosteliida, Heterolobosea, and Basidiomycota were detected, there were no genomes reconstructed for these organisms, likely due to low abundance or genome fragmentation.

Whole-community analysis, including eukaryotes

To test whether the presence of organic carbon within the CG_WC sample would enrich for heterotrophic metabolic pathways (and against members of chemolithoautotrophic communities typically associated with the Crystal Geyser community), we searched the CG_WC and CG_bulk samples using HMMs for CAZymes grouped by substrate (Cantarel et al. 2009), lipase HMMs from the Lipase Engineering Database (Fischer and Pleiss 2003), and a protease BLAST database from MEROPS (Rawlings et al. 2016). Predicted proteases and lipases were filtered to specifically identify putative excreted proteases and lipases by searching for proteins with secretion signals identified with SignalP (Petersen et al. 2011) and one or less transmembrane domains with TMHMM (Krogh et al. 2001).

Pathways previously described as dominant within the Crystal Geyser such as the Wood Ljungdahl carbon fixation pathway and Ni-Fe hydrogenases were depleted in CG_WC as compared to CG_bulk. Instead, CAZymes targeting cellulose, hemicellulose, pectin, starch, and other polysaccharides were enriched in CG_WC, indicating an increased capacity for degradation of complex carbohydrates (Fig. 6). A strong enrichment for excreted lipases and proteases was also detected, further indicative of an increase in the amount of heterotrophic metabolisms (Fig. 6). CG_WC also had a strong enrichment for methanol oxidation.

The four binned eukaryotic genomes contributed substantially to the putative heterotrophic categories (Supplemental Table S3). Fungi are known to exhibit different CAZyme profiles based upon their lifestyle (Ohm et al. 2012; Kim et al. 2016). An analysis of the CAZyme profiles of the three fungal bins focused on plant cell wall-targeting CAZymes supports the role of these fungi as possible plant pathogens or saprotrophs (Supplemental Table S4; Floudas et al. 2012; Ohm et al. 2012; Kim et al. 2016). A profile of CAZymes found within the Arthropoda bin revealed a large number of chitin-targeting CAZymes (Supplemental Table S3).

Testing EukRep in recovery of eukaryote genomes from other ecosystems

To test the broader application of EukRep, we applied the method to infant fecal samples and thiocyanate reactor samples in which eukaryotes had previously been identified (Sharon et al. 2013; Kantor et al. 2015, 2017; Raveh-Sadka et al. 2015, 2016). By using EukRep, we were able to quickly and systematically scan 226 samples for the presence of eukaryotic sequences. Six relatively complete fungal genomes were recovered from fecal samples from three infants (Fig. 4). Three are *Candida albicans* and were reconstructed from two different infants. The two genomes from the same infant are indistinguishable and very closely related to that from the third infant. All three are closely related to but distinguishable from the *C. albicans* reference strain WO-1 (Fig. 4; Supplemental Fig. S5A). The other three fungal genomes are strains of *Candida parapsilosis* that all occurred in a single infant. These are essentially indistinguishable from each other and from the C. *parapsilosis* strain CDC317 reference genome, with which they share >99.7% average nucleotide identity (ANI) (Fig. 4; Supplemental Fig. S5A,B; Sharon et al. 2013; Raveh-Sadka et al. 2015, 2016). *C. albicans* and *C. parapsilosis* are both clinically relevant human pathogens (Trofa et al. 2008; Kim and Sudbery 2011).

Within thiocyanate reactor samples, genomes of a rotifer, Rhizaria, and a relative of the slime mold *Fonticula alba* had previously been identified (Kantor et al. 2015, 2017). With EukRep, we were able to rapidly identify these eukaryotic genomes and evaluate their completeness. Genome completeness analysis benefited from improved gene predictions for single-copy orthologous genes and showed that the identified genomes ranged in completeness

Figure 5. Phylogenetic placement of binned eukaryotic genomes with maximum likelihood analysis of 16 concatenated ribosomal protein alignments. Genomes from Crystal Geyser, infant-derived fecal samples, and thiocyanate reactor samples are identified with blue, red, and purple circles, respectively. Branches with greater than 50% bootstrap support are labeled with their bootstrap support. Reference ribosomal proteins were obtained from Hug et al. (2016), JGI (Grigoriev et al. 2011), and NCBI (NCBI Resource Coordinators 2017).

from 69%–91%. (Fig. 4). As previously reported (Kantor et al. 2017), the rotifer was present in seven different samples (Rotifer_A-G) (Fig. 4), consistent with its persistence in the thiocyanate reactor community. All seven bins shared greater than 99% ANI (Supplemental Fig. S5B) indicating they are likely the same species.

Discussion

Using a newly acquired and two previously reported whole-community metagenomic data sets, we demonstrated that it is possible to rapidly recover high-quality eukaryotic genomes from metagenomes for phylogenetic and metabolic analyses. The key step implemented in this study was the presorting of eukaryotic genome fragments prior to gene prediction. By training and using eukaryotic gene predictors, we achieved much higher quality eukaryotic gene predictions than those obtained using a prokaryotic gene prediction algorithm on the entire data set (i.e., without separation based on phylogeny). This was critical for draft genome recovery and evaluation of genome completeness.

Classification of assembled genome fragments at the domain level was surprisingly accurate, with 98.0% (Fig. 2C) of eukaryotic sequences being correctly identified as eukaryotic, despite no close relative in the training set in many cases (Supplemental Table S2). The high accuracy of separation suggests some underlying pattern of *k*-mer frequencies that is different in eukaryotes compared to prokaryotes. In part, the signature may arise from different codon use patterns associated with the different genetic codes for bacteria and eukaryotes.

We anticipate that reexamination of environmental metagenomic data sets using the same approach as implemented here will yield high quality genomes for previously unknown eukaryotes. An important benefit from this and future sequencing efforts will be an expanded knowledge of the diversity, distribution, and functions of microbial eukaryotes, which are widely acknowledged as understudied (Pawlowski et al. 2012). Increasing the diversity of sequenced eukaryotic genomes would benefit evolutionary studies. Current eukaryotic multigene trees form a solid backbone of the eukaryotic tree of life (Parfrey et al. 2010) but suffer from sparse eukaryotic taxon sampling. Single-gene trees, which are possible to construct from gene surveys, lack the resolution of multigene trees (Rokas and Carroll 2005). Comprehensive sequencing of full genomes would help diminish the sparse taxon sampling problem in multigene trees and improve eukaryotic evolutionary reconstructions, with implications for understanding of eukaryotic protein function. For example, Ovchinnikov et al. (2017) demonstrated that it is possible to accurately predict protein structure by utilizing residue-residue contacts inferred from evolutionary data, but such analyses require large numbers of aligned sequences. More diverse eukaryotic sequences could expand the utility of this method for eukaryotic protein family analyses. Furthermore, a broader diversity of eukaryotic genomes would provide new insights regarding gene transfer patterns and whole-genome evolution.

EukRep, applied in the context of metagenomics, may prove useful for genome sequencing projects where isolation of the organism of interest may be difficult or not technically feasible. For example, it could be applied to study populations of bacteria

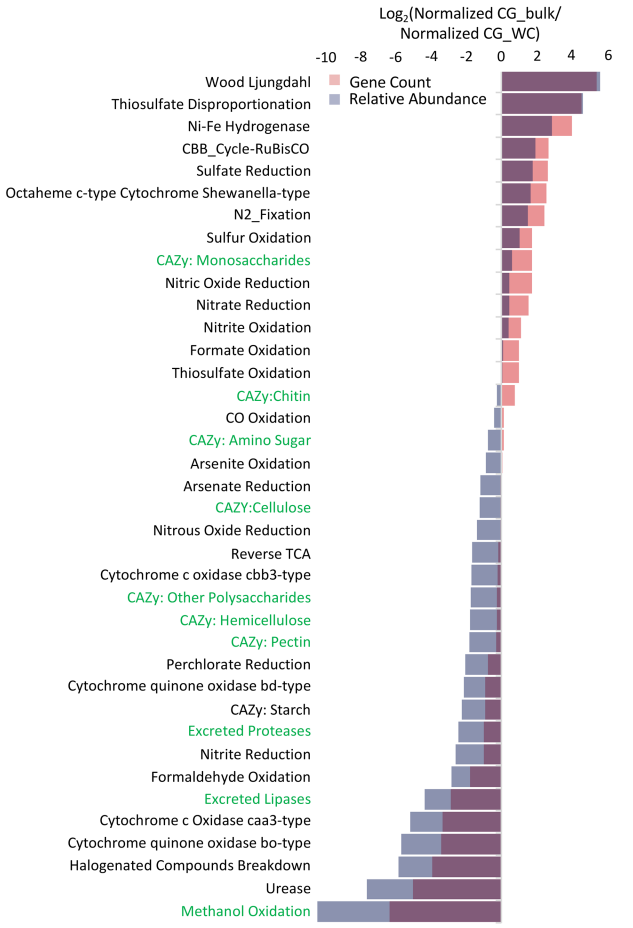

Figure 6. Comparison of CG_WC and CG_bulk metabolic capacity. Log_2 ratio of all annotated genes found within the CG_bulk sample against annotated genes found in the CG_WC sample. Annotated genes were grouped into categories based upon scores with a custom set of metabolic pathway marker HMMs (Anantharaman et al. 2016), CAZyme HMMs (Cantarel et al. 2009), and protease and lipase HMMs from MEROPs and the Lipase Engineering Database, respectively. Putative proteases and lipases were also filtered to only those containing a secretion signal and less than three transmembrane domains (see Methods). Gene count (red) is the ratio of total number of genes in each category for each sample normalized by the total number of genes found in the sample. Relative abundance (blue) is the ratio of average read coverage depth of the contig containing a given annotated gene in each category normalized by the sample read count multiplied by read length.

within the hyphae of arbuscular mycorrhizal fungi (Hoffman and Arnold 2010).

Eukaryotic cells frequently contain multiple sets of chromosomes (diploid or polyploid). These are often very similar but not identical and can result in the genome assembly alternating between collapsing and splitting contigs representing homologous genomic regions (Margarido et al. 2015). If reads are only allowed to map to one location when determining genome coverage, this could lead to variation of coverage values across different portions of a genome. As differential coverage of contigs is a parameter commonly used to help bin genomes, ploidy can complicate genome recovery. Another potential problem could relate to contamination of eukaryotic genome bins with some bacterial fragments. This will occur to some extent, given that some bacterial and archaeal contigs were wrongly classified as eukaryotic. Phylogenetic profiling of the predicted genes can be used to screen out most prokaryotic sequences.

During development, we noted that the frequency of correct identification of bacterial genomes was improved by increasing the number and diversity of eukaryote sequences used in classifier training. Further improvements are anticipated as the variety of reference sequences increases. However, there may be biological reasons underpinning incorrect profiles. The small number of cases where EukRep profiled bacteria as eukaryotes or vice versa may be interesting targets for further analysis. Notably, almost all are inferred or known symbionts or parasites, raising the question of whether their sequence composition has evolved to mirror that of their hosts.

We demonstrated the value of EukRep-enabled analyses through study of an ecosystem that had been perturbed by addition of a carbon source. The results clearly show a large shift in the community composition and selection for fungi. Of the binned genomes, the fungi have by far the most cellulose-, hemicellulose-, and pectin-degrading enzymes, consistent with their enrichment in response to high organic carbon availability from degrading wood. We also genomically characterized what appears to be a macroscopic hexapod that is related to springtails (Collembola), organisms known to feed on fungi (Chen et al. 1996). Given that the hexapod genome has a large number of chitin-degrading enzymes (Supplemental Table S3), we speculate that it may be part of the community supported by the fungi in the decaying wood. However, it is also possible that it was associated with the wood prior to its addition to the geyser conduit. Interestingly, the eukaryote-based community contains very few members of the candidate phyla radiation (CPR) and an archaeal radiation known as DPANN and other CP bacteria. These novel organisms are mostly predicted to be anaerobes and are highly abundant in groundwater samples that were likely sourced from deep aquifers under the Colorado Plateau (Probst et al. 2018). The results of the current study indicate that CPR and DPANN in the Crystal Geyser system are adapted to an environment relatively low in carbon availability, a finding that may guide future laboratory enrichment studies that target these organisms.

Overall, the results reported here demonstrate that comprehensive, cultivation-independent genomic studies of ecosystems containing a wide variety of organism types are now possible. Examples of future applications include analysis of the distribution and metabolic capacities and potential pathogenicity of fungi in the human microbiome, tracking of eukaryotes (including multicellular eukaryotes) in reactors used in biotechnologies, profiling of the built environment, and natural ecosystem research.

Methods

Crystal Geyser sample collection and DNA extraction

Details of filtration of groundwater for sample CG_bulk is given in Probst et al. (2016) (sample CG23_combo_of_CG06-09_8_20_14). Groundwater containing particulate wood was collected in a 50-mL Falcon tube. All samples were frozen on site on dry ice and stored at −80°C until further processing. The sample with the particulate wood was spun down, and DNA extraction was performed as described previously (Emerson et al. 2015).

Crystal Geyser DNA sequencing and assembly

Raw sequencing reads were processed with bbtools (http://jgi.doe.gov/data-and-tools/bbtools/) and quality-filtered with SICKLE

with default parameters (version 1.21; https://github.com/najoshi/sickle). IBDA_UD (Peng et al. 2012) was used to assemble and scaffold filtered reads. IDBA_UD was chosen as it is a widely used, publicly available program designed for metagenomic assemblies. Unlike almost all other such assemblers, it includes a scaffolding step. This is important because longer sequences can be more robustly binned. Scaffolding errors were corrected using MISS (I Sharon, BC Thomas, JF Banfield, unpubl.), a tool that searches and fixes gaps in the assembly based on mapped reads that exhibit inconsistencies between raw reads and assembly. The two Crystal Geyser samples used for binning and comparison in this study, CG_WC and CG_bulk, resulted in 874 and 529 Mbps of assembled scaffolds, respectively.

Prokaryotic genome binning and annotations

Protein-coding genes were predicted on entire metagenomic samples using MetaProdigal (Hyatt et al. 2012). Ribosomal RNA genes were predicted with Rfam (Nawrocki et al. 2015), and 16S rRNA genes were identified using SSU-ALIGN (Nawrocki 2009). Predicted proteins were functionally annotated by finding the best BLAST hit using USEARCH (UBLAST) (Edgar 2010) against UniProt (The UniProt Consortium 2017), UniRef90 (Suzek et al. 2007), and KEGG (Kanehisa et al. 2016). Prokaryotic draft genomes were binned through the use of emergent self-organizing map (ESOM)-based analyses of tetranucleotide frequencies. Bins were then refined through the use of ggKbase (ggkbase.berkeley.edu) to manually check the GC, coverage, and phylogenetic profiles of each bin.

EukRep training and testing

EukRep, along with trained linear SVM classifiers, are available at https://github.com/patrickwest/EukRep. A diverse reference set of 194 bacterial genomes, 218 archaeal genomes, 27 opisthokonta, and 43 protist genomes was obtained from NCBI and JGI (Supplemental Table S1). Hug et al. (2016), JGI Mycocosm database (jgi.doe.gov/fungi), and the NCBI taxonomy browser were used as references for selecting genomes from a broad taxonomic range. The contigs comprising these genomes were split into 5-kb chunks for which 5-mer frequencies were calculated (Anvar et al. 2014). Contigs shorter than 3 kb were excluded. The 5-mer frequencies were used to train a linear-SVM (scikit-learn, v. 0.18, default parameters with C = 100) to classify sequences as either of opisthokonta, protist, bacterial, or archaeal origin. The hyperparameter C was optimized using a grid-search with cross-validation and accuracy on a subset of test genomes used for scoring. To classify an unknown or test sequence, the sequence was split into 5-kb chunks, and 5-mer frequencies were determined for each chunk. Contigs shorter than 3 kb were excluded. The trained classifier was then used to predict whether the sequence is of opisthokonta, protist, bacterial, or archaeal origin. Once classified, the 5-kb chunks were stitched back together into their parent scaffold, and the parent scaffold's taxonomy was determined based upon majority rule of its 5-kb chunks. Accuracy for a given genome was considered to be the percent of total base pairs correctly identified as either eukaryotic or prokaryotic. To compare the effect of *k*-mer length on prediction accuracy, *k*-mer frequencies ranging in length from 4 to 6 bp from the same training set were used to train separate linear-SVM models. To determine the minimum sequence length cutoff, test genomes were fragmented into pieces of *n* length, and sequences shorter than *n* length were filtered out.

To test EukRep, a separate set of 97 eukaryotic and 393 prokaryotic genomes was obtained from NCBI and JGI (Supplemental Table S2). Genomes assembled into less than 10 contigs were fragmented into 100-kb pieces in order to better represent metagenomic data sets. EukRep was then run on each genome individually. Accuracy for a given genome was measured by dividing the total number of base pairs correctly classified by the total number of base pairs tested.

Eukaryotic genome binning and annotations

Scaffolds predicted to be eukaryotic scaffolds by EukRep were binned into putative genomes using CONCOCT (Alneberg et al. 2014). Eukaryotic genome bins smaller than 5 Mbp were not included in further analyses. Gene predictions were performed individually on each bin with the MAKER2 pipeline (v. 2.31.9) (Holt and Yandell 2011) with default parameters and using GeneMark-ES (v. 4.32) (Ter-Hovhannisyan et al. 2008), AUGUSTUS (v. 2.5.5) (Stanke et al. 2006) trained with BUSCO (v. 2.0) (Simão et al. 2015), and the proteomes of *C. reinhardtii* (Merchant et al. 2007), *N. crassa* (Galagan et al. 2003), and *R. filosa* (Glöckner et al. 2014) for homology evidence. These gene prediction strategies were employed due to their ability to be automatically trained for individual genomes. Completeness of the combined MAKER2 predicted gene set as well as the individual gene predictor gene sets were compared, and the most complete based upon BUSCO analysis was used in future analyses. Phylogenetic classification of the predicted genes along with presence or absence of single-copy orthologous genes was then used to refine each binned genome. CAZYmes were detected in both eukaryotic and prokaryotic bins through the use of HMMER3 (v. 3.1b2) (Eddy 1998) and a set of HMMs obtained from dbCAN (Yin et al. 2012). The presence or absence of various metabolic pathways was determined by using a custom set of metabolic pathway marker gene HMMs (Anantharaman et al. 2016) and HMMER3. Protease and lipases were predicted by using lipase HMMs from the Lipase Engineering Database (Fischer and Pleiss 2003) and BLASTing against a protease database obtained from MEROPS (Rawlings et al. 2016). Putative excreted proteases and lipases were identified by searching for predicted proteases and lipases with secretion signals identified with SignalP (Petersen et al. 2011) and no more than one transmembrane domain with TMHMM (Krogh et al. 2001). To find potentially contaminating prokaryotic scaffolds, predicted genes were BLASTed against UniProt. Scaffolds in which the majority of best hits belonged to prokaryotic genes were removed.

Read data sets for previously published metagenomes are available under Sequence Read Archive (SRA) accession numbers SRA052203 and SRP056932 at (SRA; http://www.ncbi.nlm.nih.gov/sra) and BioProjects PRJNA294605 and PRJNA279279.

Eukaryotic gene set comparisons

Nine gene sets were obtained from JGI's mycocosm database (Grigoriev et al. 2011) and NCBI. For each genome, genes were predicted without transcriptomic evidence by running assembled sequences through the MAKER2 pipeline with AUGUSTUS trained with BUSCO and GeneMark-ES in self-training mode. Gene sets predicted with transcriptomic evidence were obtained from the JGI portal and NCBI. For comparison against eukaroytic MetaProdigal predicted gene sets, MetaProdigal was run with the '-meta' flag.

Eukaryote genome completeness estimates

Genome completeness of predicted eukaryotic genomes was estimated based on the presence of conserved, low-copy-number genes. BUSCO (v. 2.0) (Simão et al. 2015) was run with default parameters using the "eukaryota_odb9" lineage set composed of 303 core eukaryotic genes. Completeness was considered to be

the percent of the total 303 core genes that were present in either single or duplicated copies. Additionally, the number of genes identified as duplicated was used as a way to estimate how much of a given binned genome appeared to be from a single organism.

Mock metagenome analysis

Bulk soil was collected from the Eel River Critical Zone Observatory (CZO) in Northern California. DNA extraction was performed as described previously (Emerson et al. 2015). Raw sequencing reads were processed with bbtools (http://jgi.doe.gov/data-and-tools/bbtools/) and quality-filtered with SICKLE with default parameters (version 1.21; https://github.com/najoshi/sickle). IBDA_UD (Peng et al. 2012) was used to assemble and scaffold filtered reads. The genome of *Choanephora cucurbitarum* was obtained from the NCBI genome database and spiked into the assembled soil metagenome. MetaProdigal was used to obtain gene predictions for the entire sample. EukRep was then used to classify scaffolds as eukaryotic. CONCOCT was used to bin predicted eukaryotic sequences, and gene predictions were reperformed on the *Choanephora* bin with the MAKER2 pipeline using GeneMark-ES and AUGUSTUS for gene prediction.

taxator–tk comparison

The microbial-full_20150430 database was obtained from the *taxator-tk* (Dröge et al. 2015) website and was used for mapping. Mapping of test genomes against the reference database was performed using BLASTN with default alignment parameters and output format described in the *taxator-tk* manual. In a second round of testing, scaffolds belonging to test genomes were removed from the test set to simulate genomes from novel organisms. Taxonomic assignment and binning were performed as described in the *taxator-tk* manual without filtering alignments.

Phylogenetic analyses

To determine ANI between genomes, dRep was used (Olm et al. 2017). To estimate taxonomic composition of Crystal Geyser samples, *rpS3* proteins were searched against KEGG (Kanehisa et al. 2016) with USEARCH (UBLAST) (Edgar 2010), and the taxonomy of the top hit was used to assign identified *rpS3*s to taxonomic groups. Abundance of identified *rpS3*s was determined by calculating the average coverage depth of the scaffolds containing annotated ribosomal protein S3 (*rpS3*) genes. Average coverage depth was calculated by dividing the number of reads mapped to the scaffold by the scaffold length. Abundances were normalized for comparison across samples by multiplying the average coverage depth by the sample read count times read length.

Four hundred sixty-one protein sets were obtained from binned eukaryotic genomes, publicly available genomes from the Joint Genome Institute's IMG-M database (img.jgi.doe.gov; Chen et al. 2016), NCBI, the Candida Genome Database (http://www.candidagenome.org/), and a previously developed data set (Hug et al. 2016). For each protein set, 16 ribosomal proteins (L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24, S3, S8, S10, S17, and S19) were identified by BLASTing a reference set of 16 ribosomal proteins obtained from a variety of protistan organisms against the protein sets. BLAST hits were filtered to a minimum e-value of 1.0×10^{-5} and minimum target coverage of 25%. The 16 ribosomal protein data sets were aligned with MUSCLE (v. 3.8.31) (Edgar 2004) and trimmed by removing columns containing 90% or greater gaps. The alignments were then concatenated. A maximum likelihood tree was constructed using RAxML (v. 8.2.10) (Stamatakis 2014), on the CIPRES web server (Miller et al. 2010), with the LG plus gamma model of evolution (PROTGAMMALG) and with the number of bootstraps automatically determined with the MRE-based bootstopping criterion.

Data access

Sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRS2648722. Assembled eukaryotic genomes from Crystal Geyser have been submitted to DDBJ/ENA/GenBank (https://www.ncbi.nlm.nih.gov/genbank/) under the accession numbers PCFH00000000, PCFI00000000, PCFJ00000000, and PCFG00000000. EukRep along with trained linear SVM classifiers are available at https://github.com/patrickwest/EukRep and as Supplemental Data 2.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1106400. This study was partially funded by the Sloan Foundation ("Deep Life," grant no. G-2016-20166041) and NSF Sustainable Chemistry grant (1349278). The work conducted at the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. A.J.P. was supported by the German Science Foundation under DFG PR 1603/1-1. We thank M.R. Olm, Dr. C.T. Brown, and Dr. A. Salamov for their contributions to this study.

References

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomics contigs by coverage and composition. *Nat Methods* **11:** 1144–1146.

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, et al. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Commun* **7:** 13219.

Anvar SY, Khachatryan L, Vermaat M, van Galen M, Pulyakhina I, Ariyurek Y, Kraaijeveld K, den Dunnen JT, de Knijff P, 't Hoen PAC, et al. 2014. Determining the quality and complexity of next-generation sequencing data without a reference genome. *Genome Biol* **15:** 555.

Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of replication rates in microbial communities. *Nat Biotechnol* **34:** 1256–1263.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37:** D233–D238.

Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. 2008. Protists are microbes too: a perspective. *ISME J* **3:** 4–12.

Chen B, Snider RJ, Snider RM. 1996. Food consumption by Collembola from northern Michigan deciduous forest. *Pedobiologia* **40:** 149–161.

Chen IA, Markowitz VM, Che K, Palaniappan K, Szeta E, Pillay M, Ratner A, Huang J, Anderson E, Huntemann M, et al. 2016. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res* **45:** D507–D516.

Cuvelier ML, Allen AE, Monier A, McCrow JP, Messié M, Tringe SG, Woyke T, Welsh RM, Ishoey T, Lee JH, et al. 2010. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc Natl Acad Sci* **107:** 14679–14684.

Dröge J, Gregor I, McHardy AC. 2015. *Taxator-tk*: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics* **31:** 817–824.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792–1797.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460–2461.

Emerson JB, Thomas BC, Alvarez W, Banfield JF. 2015. Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. *Environ Microbiol* **18:** 1686–1703.

Faddeeva-Vakhrusheva A, Derks MF, Anvar SY, Agamennone V, Suring W, Smit S, van Straalen NM, Roelofs D. 2016. Gene family evolution reflects adaptation to soil environmental stressors in the genome of the collembolan *Orchesella cincta*. *Genome Biol Evol*. **8:** 2106–2117.

Fischer M, Pleiss J. 2003. The Lipase Engineering Database: a navigation and analysis tool for protein families. *Nucleic Acids Res* **31:** 319–321.

Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martinez AT, Otillar R, Spatafora JW, Yadav JS, et al. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* **336:** 1715–1719.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422:** 859–868.

Glöckner G, Hülsmann N, Schleicher M, Noegel AA, Eichinger L, Gallinger C, Pawlowski J, Sierra R, Eureneuer U, Pillet L, et al. 2014. The genome of the foraminiferan *Reticulomyxa filosa*. *Curr Biol* **24:** 11–18.

Grigoriev IV, Cullen D, Goodwin SB, Hibbett D, Jeffries TW, Kubicek CP, Kuske C, Magnuson JK, Martin F, Spatafora JW, et al. 2011. Fueling the future with fungal genomics. *Mycology* **2:** 192–209.

Hoffman MT, Arnold AE. 2010. Diverse bacteria inhabit living hyphae of phylogenetically diverse fungal endophytes. *Appl Environ Microbiol* **76:** 4063–4075.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12:** 491.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life. *Nature Microbiol* **1:** 16048.

Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28:** 2223–2230.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44:** D457–D462.

Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, Thomas BC, Banfield JF. 2013. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *mBio* **4:** e00708-13.

Kantor RS, van Zyl AW, van Hille RP, Thomas BC, Harrison STL, Banfield JF. 2015. Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unraveled with genome-resolved metagenomics. *Environ Microbiol* **17:** 4929–4941.

Kantor RS, Huddy RJ, Iyer R, Thomas BC, Brown CT, Anantharaman K, Tringe S, Hettich RL, Harrison STL, Banfield JF. 2017. Genome-resolved meta-omics ties microbial dynamics to process performance in biotechnology for thiocyanate degradation. *Environ Sci Technol* **51:** 2944–2953.

Kim J, Sudbery P. 2011. *Candida albicans*, a major human fungal pathogen. *J Microbial* **49:** 171–177.

Kim KT, Jeon J, Choi J, Cheong K, Song H, Choi G, Kang S, Lee YH. 2016. Kingdom-wide analysis of fungal small secreted proteins (SSPs) reveals their potential role in host association. *Front Plant Sci* **7:** 186.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5:** 59.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305:** 567–580.

Mangot J, Logares R, Sánchez P, Latorre F, Seeleuthner Y, Mondy S, Sieracki ME, Jaillon O, Wincker P, Vargas C, et al. 2017. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci Rep* **7:** 41498.

Margarido GRA, Heckerman D. 2015. ConPADE: genome assembly ploidy estimation from next-generation sequencing data. *PLoS Comput Biol* **11:** e1004229.

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-laylin LK, Maréchal-Drouard L, et al. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318:** 245–250.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *Proceedings of the Gateway Computing Environments Workshop (GCE)*, pp. 1–8, New Orleans, LA.

Min B, Park JH, Park H, Shin HD, Choi IG. 2017. Genome analysis of a zygomycete fungus *Choanephora cucurbitarum* elucidates necrotrophic features including bacterial genes related to plant colonization. *Sci Rep* **7:** 40432.

Mistry J, Bateman A, Finn RD. 2007. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* **8:** 298.

Monier A, Welsh RM, Gentemann C, Weinstock G, Sodergren E, Armbrust EV, Eisen JA, Worden AZ. 2012. Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environ Microbiol* **14:** 162–176.

Morales-Cruz A, Amrine KC, Blanco-Ulate B, Lawrence DP, Travadon R, Rolshausen PE, Baumgartner K, Cantu D. 2015. Distinctive expansion of gene families associated with plant cell wall degradation, secondary metabolism, and nutrient uptake in the genomes of grapevine trunk pathogens. *BMC Genomics* **16:** 469.

Mosier AC, Miller CS, Frischkorn KR, Ohm RA, Li Z, LaButti K, Lapidus A, Lipzen A, Chen C, Johnson J, et al. 2016. Fungi contribute critical but spatially varying roles in nitrogen and carbon cycling in acid mine drainage. *Front Microbiol* **7:** 238.

Nawrocki EP. 2009. "Structural RNA homology search and alignment using covariance models." PhD dissertation, Washington University, St. Louis, MO.

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* **43**(Database issue)**:** D130–D137.

NCBI Resource Coordinators. 2017. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **45:** D12–D17.

Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, Condon BJ, Copeland AC, Dhillon B, Glaser F. 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen dothideomycetes fungi. *PLoS Pathogens* **8:** e1003037.

Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genome de-replication that enables tracking of microbial genotypes and improved genome recovery from metagenomes. *ISME J* **11:** 2864–2868.

Ovchinnikov S, Park H, Varghese N, Huang P, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides N, Baker D. 2017. Protein structure determination using metagenome sequence data. *Science* **355:** 294–298.

Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Marrison HG, Sogin ML, Patterson DJ, Katz LA. 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Syst Biol* **59:** 518–533.

Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle J, Dunthorn M, et al. 2012. CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal plant and fungal kingdoms. *PLoS Biol* **10:** e1001419.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in python. *JMLR* **12:** 2825–2830.

Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomics sequence data with highly uneven depth. *Bioinformatics* **28:** 1420–1428.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8:** 785–786.

Probst AJ, Weinmaier T, Raymann K, Perras A, Emerson JB, Rattea T, Wanner G, Klingl A, Berg IA, Yoshinaga M. 2014. Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat Commun* **5:** 5497.

Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, Hug LA, Burstein D, Emerson JB, Thomas BC, et al. 2016. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO_2 concentrations. *Environ Microbiol* **19:** 459–474.

Probst AJ, Ladd B, Jarett JK, Sieber CMK, Emerson JB, Thomas BC, Stieglemier M, Kling A, Woyke T, Ryan MC, et al. 2018. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nat Microbiol* **3:** 328–336.

Quandt CA, Kohler A, Hesse CN, Sharpton TJ, Martin F, Spatafora JW. 2015. Metagenome sequence of *Elaphomyces granulatus* from sporocarp tissue reveals Ascomycota ectomycorrhizal fingerprints of genome expansion and a *Proteobacteria*-rich microbiome. *Environ Microbiol* **17:** 2952–2968.

Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, Sharon I, Baker R, Good M, Morowitz MJ, et al. 2015. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *eLife* **4:** e05477.

Raveh-Sadka T, Firek B, Sharon I, Beker R, Brown CT, Thomas BC, Morowitz MJ, Banfield JF. 2016. Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants. *ISME J* **10:** 2817–2830.

Rawlings ND, Barrett AJ, Finn RD. 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* **44:** D343–D350.

Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol* **22:** 1337–1344.

Roy RS, Price DC, Schliep A, Cai G, Korobeynikov A, Yoon HS, Yang EC, Bhattacharya D. 2014. Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci Rep* **4:** 4780.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23:** 111–120.

Simão FA, Waterhouse RM, Loannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31:** 3210–3212.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7:** 62.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23:** 1282–1288.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* **18:** 1979–1990.

Trofa D, Gácser A, Nosanchuk JD. 2008. *Candida parapsilosis*, an emerging fungal pathogen. *Clin Microbiol Rev* **21:** 606–625.

The UniProt Consortium. 2017. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res* **45:** D158–D169.

Vaulot D, Lepère C, Toulza E, De la Iglesia R, Poulain J, Gaboyer F, Moreau H, Vandepoele K, Ulloa O, Gavory F, et al. 2012. Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS One* **7:** e39648.

Vervier K, Mahe P, Tournoud M, Veyrieras J, Vert J. 2016. Large-scale machine learning for metagenomics sequence classification. *Bioinformatics* **32:** 1023–1032.

Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, Lapidus A, Wu D, McCutcheon JP, McDonals BR,et al. 2010. One bacterial cell, one complete genome. *PLoS One* **5:** e10314.

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **40:** W445–W451.

Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S, Bhattacharya D. 2011. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332:** 714–717.

Received July 31, 2017; accepted in revised form February 27, 2018.