bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Estimating the maximal growth rates of eukaryotic microbes from cultures and metagenomes via codon usage patterns

JL Weissman[1*], Edward-Robert O. Dimbo[1], Arianna I. Krinos[2,3], Christopher Neely[4], Yuniba Yagües[5], Delaney Nolin[1], Shengwei Hou[1‡], Sarah Laperriere [1], David A. Caron[1], Benjamin Tully[6,7], Harriet Alexander[2], Jed A. Fuhrman[1],

1 Department of Biological Sciences–Marine and Environmental Biology, University of Southern California, Los Angeles, USA
2 Biology Department, Woods Hole Oceanographic Institution, Woods Hole, USA
3 MIT-WHOI Joint Program in Oceanography, Cambridge and Woods Hole, USA
4 Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, USA
5 Department of Chemical Engineering, University of California Berkeley, Berkeley, USA
6 University of Southern California, Wrigley Institute for Environmental Studies, Los Angeles, USA
7 University of Southern California, Center for Dark Energy Biosphere Investigations, Los Angeles, USA

‡Current Affiliation: Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen, China
* Corresponding author: jakeweis@usc.edu

Abstract

Microbial eukaryotes are ubiquitous in the environment and play important roles in key ecosystem processes, including accounting for a significant portion of global primary production. Yet, our tools for assessing the functional capabilities of eukaryotic microbes in the environment are quite limited because many microbes have yet to be grown in culture. Maximum growth rate is a fundamental parameter of microbial lifestyle that reveals important information about an organism's functional role in a community. We developed and validated a genomic estimator of maximum growth rate for eukaryotic microbes, enabling the assessment of growth potential for both cultivated and yet-to-be-cultivated organisms. We produced a database of over 700 growth predictions from genomes, transcriptomes, and metagenome-assembled genomes, and found that closely related and/or functionally similar organisms tended to have similar maximal growth rates. By comparing the maximal growth rates of existing culture collections with environmentally-derived genomes we found that, unlike for prokaryotes, culture collections of microbial eukaryotes are only minimally biased in terms of growth potential. We then extended our tool to make community-wide estimates of growth potential from over 500 marine metagenomes, mapping growth potential across the global oceans. We found that prokaryotic and eukaryotic communities have highly correlated growth potentials near the ocean surface, but that this relationship disappears deeper in the water column. This suggests that fast growing eukaryotes and prokaryotes thrive under similar conditions at the ocean surface, but that there is a decoupling of these communities as resources become scarce deeper in the water column.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Microbial eukaryotes are ubiquitous in the environment, and play diverse roles relevant to ecosystem (e.g., [1, 2]) and human (e.g., [3, 4]) health. In the ocean in particular, protists dominate, accounting for approximately 30% of total marine biomass [5]. Among marine primary producers alone, protists account for a third of total biomass [5]. Marine systems account for about half of all global primary production [6], so that the abundance of protists in these systems suggests an important overall role for protists in regulating global carbon cycles [7], among other biogeochemical cycles. And yet, our tools for studying the ecology and evolution of eukaryotic microbes are still quite limited, at least in comparison to their prokaryotic neighbors [8].

Several recent developments have greatly advanced our ability to survey the ecology of microbial eukaryotes directly from the environment using metagenomics. Large-scale efforts to augment the sizes of our existing genomic and transcriptomic databases, specifically the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP; [9]), have expanded our ability to use database-dependent approaches for metagenomic analysis for both taxonomic and functional classification (e.g., [10–13]). At the same time, novel approaches for binning and validation have been applied by multiple groups to reconstruct high-quality metagenome-assembled genomes (MAGs) from environmental datasets [14–16].

With these new environmentally-derived genomes come new challenges – specifically that of inferring features of an organism's physiology and ecology from its genome sequence, a persistent challenge in metagenomics [17–19]. One trait of particular interest is the maximal growth rate of an organism, a fundamental parameter of microbial lifestyle that can tell us a great deal about an organism's ecology [20–22]. Among microbial eukaryotes, minimal doubling times range over two orders of magnitude, from hours (e.g., [23, 24]) to days (e.g., [25]), and potentially even weeks (e.g., [26, 27]). Temperature sets a well-studied upper-bound on the maximal growth rates of microbial eukaryotes (see work on the Eppley Curve, e.g., [28–33]), but there is a great deal of variation among species below this threshold. For prokaryotes, genomic signals of translation optimization can be leveraged in order to predict the maximal growth rates of an organism [20, 22, 34]. Here, we show that the same signals, specifically the codon usage bias of highly expressed genes, can be used to estimate the growth potential of eukaryotic microbes directly from their genome sequences. We compiled a database of 178 species of eukaryotic microbes with recorded growth rates in culture and either publicly available genomes or transcriptomes. We then used this database to build a genomic predictor of growth potential for eukaryotic microbes. We applied this tool to a set of 465 MAGs and 517 metagenomes to derive ecological insights about the variation of eukaryotic growth potential across organisms and environments.

Results and Discussion

Predicting maximal growth rates of eukaryotic microbes

We compiled an initial dataset of maximal growth rates and optimal growth temperatures recorded in culture for 178 species with either genomic or transcriptomic information publicly available (S1 Fig, S1 Table). A sizeable portion of this dataset corresponded to marine eukaryotic microbes, with 101 entries corresponding to organisms in the MMETSP, though eukaryotic microbes from other environments were also represented, including important human pathogens (e.g., *Giardia intestinalis*, *Entamoeba histolytica*, *Leshmania spp.*, etc.). In general, eukaryotic microbes with genomes in GenBank, which tended to be human associated, had faster maximal growth

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

rates than the marine eukaryotic microbes found in MMETSP (S1 Fig). This pattern is similar to that found in prokaryotes, where human-associated bacteria and archaea typically had much faster growth rates than those found in marine systems [20].

One of the most reliable signals of optimization for rapid growth in prokaryotic genomes is high codon usage bias (CUB) in highly expressed genes [22]. The degeneracy of the genetic code means that multiple codons may code for the same amino acid, but not all organisms use alternative codons at equal frequencies. In fact, many organisms, both prokaryotic and eukaryotic, are biased in their usage of alternative codons. The codon usage patterns of genes are thought to be optimized to the relative abundance of tRNAs within the cell, and this optimization is particularly apparent among highly-expressed genes in fast-growing species [35–42]. The basic intuition here is that CUB is a result of optimization of genes for rapid translation, which in turn facilitates rapid growth. We wanted to see whether such patterns could be leveraged to predict the growth rates of eukaryotic microbes [43]. For each genome or transcriptome, we calculated the CUB of a set of highly expressed genes relative to the expected CUB calculated from all other coding sequence in that genome or transcriptome (details of these calculations can be found in Materials and Methods; [20, 39, 44]). Because ribosomal proteins are expected to have high expression across species and many physiological conditions [20, 22], we take these as our set of highly expressed genes for all analyses (see Methods and S2 Table). We found a significant negative relationship between CUB of highly expressed genes calculated in this way and the minimum doubling time of an organism (Pearson's correlation with log-transformed doubling times, $\rho = -0.400$, $p = 3.14 \times 10^{-8}$). Thus we confirmed that high CUB is a signal of growth rate optimization among microbial eukaryotes.

We then built a linear model relating CUB of highly expressed genes and optimal growth temperature to the doubling times of eukaryotic microbes (Fig 1a). We found that such a model could explain about a third of variation in doubling time among organisms (linear regression, $r^2 = 0.328$), and that both CUB ($p = 1.16 \times 10^{-7}$) and optimal growth temperature ($p = 1.89 \times 10^{-10}$) were significant predictors in the model. Interestingly, similar to what we have previously reported for prokaryotes [20], we found that for eukaroytic microbes the relationship between CUB and doubling time saturated at a threshold doubling time, after which CUB no longer changed with increasing doubling time (Fig 1b). In our earlier work on prokaryotes, we took the presence of this threshold as evidence that for slow-growing organisms who experienced little selection for optimized codon usage, drift would overwhelm the evolutionary process [20]. Thus prokaryotes could be divided into two distinct evolutionary regimes related to their growth strategies. It appears that a similar dynamic may be at work among eukaryotes, although the relatively small number of sequenced eukaryotic microbes with minimal doubling times greater than 40 hours makes this hard to assess. If such a threshold does exist for eukaryotes, it is at a much higher doubling time than the one seen in prokaryotes (40 hours versus 5 hours respectively; [20]), likely due to constraints on eukaryotic growth related to cell size and complexity. In any case, similar to the threshold effect seen for prokaryotes, we note that above this threshold, predicted maximum growth rates are likely to be overestimated. Thus, our model can only reliably predict minimum doubling times up to 40 hours, after which we can only infer that a microbe grows "very slowly".

Finally, we asked if our eukaryotic model of growth improved predictions for eukaryotic organisms relative to the predictions made by previous tools developed for prokaryotes. Consider that tools able to predict growth rate from CUB already exist, though they have been trained exclusively on prokaryotic organisms [20, 22]. We applied these prediction tools to our eukaryotic dataset and found that they systematically overestimated the growth rates of eukaryotic organisms, often by more than an order of

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

magnitude, leading to quite poor performance on eukaryotes (Fig 1c-d). Thus our eukaryote-specific model is an important development, as prokaryote-specific models cannot accurately predict eukaryotic growth rates. We have incorporated this eukaryote-specific model into the open-source and user-friendly growth prediction R package gRodon, which we previously developed to predict prokaryotic growth rates (`https://github.com/jlw-ecoevo/gRodon`; [20]).

Environmentally derived genomes reveal biases in culture collections and ecological patterns

We obtained a large set of 1669 eukaryotic MAGs assembled and binned from the Tara Oceans metagenomes by two separate groups [15, 16]. Of these, we were able to predict the growth rates of 465 MAGs in which we found at least 10 ribosomal proteins (see Materials and Methods for details). These MAGs were uniformly slow-growing, with an average minimum doubling time of approximately one day, and none with a minimum doubling time less than 10 hours long (Fig 2a). These MAGs provide a baseline expectation of the maximal growth rates of eukaryotic microbes in marine environments, and while the reconstruction of MAGs from the environment is not a wholly unbiased process, we expect these MAGs to be more representative of the distribution of organisms living in the environment than what we find within our culture collections [20]. In fact, we found that MAGs were estimated to have only slightly longer doubling times than cultured organisms in MMETSP (27.5 vs 24.5 hours respectively; t-test, $p = 9.25 \times 10^{-4}$; Fig 2a). The differences between these two datasets were most apparent when looking at the tails of the distributions of growth rates, where the MMETSP had a long tail of fast-growing organisms that was absent among the MAGs (Fig 2a,b). Altogether the data suggest that our culture databases of eukaryotes do a relatively good job at capturing an accurate distribution of growth rates among organisms, though they are slightly enriched for fast growing organisms that are rare in the environment. This result is in stark contrast to the pattern seen among marine prokaryotic organisms where culture collections were shown to be systematically biased towards fast-growers [20].

Within the set of MAGs several patterns were apparent. First, while organisms classified as phototrophic and heterotrophic had largely overlapping growth rate distributions (Fig 2c), heterotrophs tended to grow faster than phototrophs (t-test, $p = 1.58 \times 10^{-3}$; trophic classification on the basis of the presence of metabolic pathways in a MAG, taken from Alexander et al. [15]). This reflects previous findings that at higher temperatures heterotrophic marine eukaryotic microbes had faster growth rates than phototrophic ones, though phototrophs outgrew heterotrophs at lower temperatures because their growth rates decreased less dramatically with decreases in temperature [33].

Just as growth rates varied among functional groups, they also systematically varied among taxonomic groups (Fig 2d). Overall, marine fungi had the fastest average estimated growth rates. MAGs belonging to the Chlorophyta also seemed to be relatively fast growing, with a somewhat narrow range of growth rates clustered around a doubling time of about a day. By contrast Dinoflagellata, Haptophyta, and to some degree Ochrophyta all had a considerable number of very slow growing representatives (minimal doubling time > 40 hours), though these groups had very broad distributions of growth rates and included many faster growing members as well. The diversity of growth rates in these groups is perhaps not surprising, as the cell sizes of diatom and dinoflagellate species vary over two orders of magnitude, indicating a wide diversity of morphologies and environmental niches [45, 46]. Overall, the distribution of maximal growth rates varied across taxonomic groups, likely a product of both specialization for

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

different niches and historical contingency.

In accordance with the variation of growth potential across taxonomic groups, we found that closely related organisms had more similar maximal growth rates than distantly related organisms (S2 Fig). That is, the absolute difference in doubling times between two organisms increased as a function of the patristic distance (distance from tip-to-tip on a phylogeny) between the two organisms, though this relationship had little explanatory value (linear regression, $p = 2 \times 10^{-16}$, $\beta = 0.26$, $r^2 = 0.01$). In any case, below a threshold distance of 0.1 substitutions per site maximum growth rates could be extrapolated between relatives with an average absolute error under six hours (S2 Fig), which may be taken as an acceptable level of error for a general guess at overall lifestyle. This means that for 18S rRNA amplicon sequence variants (ASVs) where a genus-level relative has a genome or transcriptome available, rough estimates of growth potential may be inferred, similar to prokaryotes [20].

Predicting the growth potential of prokaryotic and eukaryotic communities from metagenomes

It is often difficult to reconstruct high-quality MAGs for many organisms, both prokaryotic and eukaryotic, from the environment. Even when we cannot easily obtain MAGs representative of the entire microbial community in a particular environment, it is possible to apply CUB-based predictors to a metagenome to estimate the average growth potential of a community [22]. The prokaryotic growth predictor previously implemented in the gRodon package allowed the user to predict the median community-wide maximal growth rate of the prokaryotic community [20]. Our eukaryotic model can be similarly applied to calculate the median maximal growth rate of the eukaryotic community represented in a metagenomic sample. To demonstrate this application, we acquired assemblies of 610 globally-distributed marine metagenomic samples from the BioGEOTRACES dataset [47]. This dataset is particularly useful for our purposes because samples were not size-fractionated, allowing both prokaryotic and eukaryotic communities to be assessed simultaneously. We sorted these metagenomes into prokaryotic and eukaryotic contigs and applied prokaryotic gRodon (using "metagenome" mode) to the prokaryotic sequences and eukaryotic gRodon (using "eukaryote" mode) to the eukaryotic sequences. Because our eukaryotic model only uses one measure of codon usage bias applied on a gene-by-gene basis, it is similar to "metagenome" mode from prokaryotic gRodon (v1.0.0) as well as growthpred, and can be applied as-is directly to mixed-species metagenomic data [20, 22]. See Materials and Methods for details of this analysis.

Overall, we were able to predict the average community-wide maximal growth rates of the prokaryotic and eukaryotic communities in 517 samples with at least 10 ribosomal proteins each that could be classified as eukaryotic or prokaryotic (Fig 3; S3 Fig). The correlation between the growth potentials of prokaryotic and eukaryotic communities at the ocean surface was striking (Pearson correlation of samples from < 100 meters, $\rho = 0.566$, $p = 1.08 \times 10^{-27}$; Fig 3a), though this relationship disappeared among deeper samples (Pearson correlation of samples from > 100 meters, $\rho = -0.101$, $p = 0.146$; Fig 3b). A linear model confirmed a significant interaction between depth and the relationship between eukaryotic and prokaryotic growth rates (linear regression of prokaryotic growth rates, $\beta_{\text{eukaryotes}} = 0.0980$, $p_{\text{eukaryotes}} = 3.05 \times 10^{-7}$, $\beta_{\text{depth}} = -0.0106$, $p_{\text{depth}} = 5.72 \times 10^{-9}$, $\beta_{\text{eukaryotes:depth}} = 1.43 \times 10^{-4}$, $p_{\text{eukaryotes:depth}} = 1.74 \times 10^{-5}$). Notably, this was not simply an effect of temperature in the model, as the CUB of eukaryotic and prokaryotic communities co-varied across samples (S4 Fig). Additionally, these patterns cannot be attributed to differences in coverage. While doubling time did decrease with the relative abundance of eukaryotic

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

contigs in a sample, as would be expected, samples with a lower proportion of eukaryotes were not particularly skewed in their estimated growth rates (S5 Fig).

It is perhaps not surprising that the growth potentials of eukaryotic and prokaryotic communities would be correlated, since conditions favorable to more copiotrophic lifestyles (e.g., high nutrients) should be similar across both prokaryotes and eukaryotes. The observed decoupling of the growth potential of eukaryotic and prokaryotic communities with depth is consistent with a model of high productivity at the surface linked through particle sinking to productivity at deeper depths. We found that eukaryotic growth potential at depth ($>$ 100 meters) was correlated with eukaryotic, but not prokaryotic, growth potential at the surface ($<$ 100 meters), suggesting that eukaryotic productivity at the surface is a primary driver of community composition at deeper depths (S6 Fig), likely in part due to the relationship between cell size and sinking rate [48, 49]. The decoupling of eukaryotic and prokaryotic growth potential with depth is also reflected in the increasing heterotrophy of the eukaryotic community as depth increases. Leveraging the MAGs discussed in the last section which had been previously mapped to a globally distributed set of marine metagenomes [15, 50], we found that MAGs that were predicted to be phototrophic dropped off in abundance relative to heterotrophic MAGs after 100 meters (S7 Fig).

Conclusions

We developed and validated a new tool to estimate the growth potential of eukaryotic microbes directly from genomic and transcriptomic sequences. Using this tool, we were able to predict the maximal growth rates of a large set of uncultured marine organisms directly from reconstructed MAGs. We found distinct patterns in growth potential across functional and taxonomic groups and assessed existing culture collections for functional bias. We then applied our tool to a large set of marine metagenomes to predict the community-wide growth potential of eukaryotes along large ocean transects. We found a clear positive relationship between eukaryotic and prokaryotic growth potential at the ocean surface, suggesting that fast growing organisms from multiple domains of life thrive under similar conditions, and the same for slow growing organisms. With an increasing number of environmental metagenomes published each year, for many environments it will now be possible to build high-resolution maps of microbial growth potential across domains, yielding insights into the drivers of microbial community structure and function.

Our tool demonstrates the clear utility of genomic and metagenomic trait estimators for eukaryotic microbes. Yet, when working with eukaryotic microbes there are relatively few bioinformatic resources both in terms of methods and databases. Moving forward, as the complexity and subtlety of our bioinformatic tool-set increases, eukaryotic microbes represent a new frontier for methods development and ecological investigations with molecular data (e.g., [11, 12, 14–16]).

Materials and Methods

The code to generate all figure and analysis in the paper can be found at `https://github.com/jlw-ecoevo/eeggo`. The new gRodon v2 R package with the eukaryotic growth rate model implemented can be found at `https://github.com/jlw-ecoevo/gRodon2`. All visualizations were made using the ggplot2 [51] and ggpubr [52] R packages.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Training Data

From a list of protist species with transcriptomes in MMETSP and/or genomes in GenBank, we searched for recorded growth rates in culture, alongside the temperatures at which these rates were recorded, from the scientific literature (S1 Fig, S1 Table). If multiple rates were reported in culture, we always took the fastest rate we were able to find. We converted between doubling time and specific growth rate using the equation: doubling time $= \frac{\ln(2)}{\text{growth rate}}$. When growth rate was reported in terms of divisions per day we instead used the conversion equation: doubling time $= \frac{1}{\text{growth rate}}$. A considerable number of growth rates recorded for marine organisms came from previous database compilation efforts by others [33, 53].

All MMETSP transcriptome assemblies with annotated coding sequence were obtained from `https://zenodo.org/record/3247846` [54]. Species that had a eukaryotic prey species listed under experimental conditions were excluded from further analysis to reduce the possibility of contamination. For each species in MMETSP, we selected the single largest transcriptome assembly. We then removed any potential cross-contamination between species by identifying possible contaminants using CroCo v1.1 [55] with a threshold of 97% identity and otherwise default parameters (`--suspect-id 97`), removing any transcripts listed as "suspect". We additionally classified transcripts using kraken2 v2.1.1 [13, 56] with the 'nt' database and default parameters, and removed any transcripts classified as viruses or prokaryotes in order to remove any potential contaminants.

For assemblies from GenBank with growth rates listed in our dataset we first ran EukMetaSanity v1.0.0 [11], which incorporates repeat prediction [57], reference protein selection [12, 58], and ab-initio gene predictions to determine putative gene loci. The output GeneMark-EP/ES predictions were used for analysis [59, 60]. We then classified coding sequences using kraken2 v2.1.1 [13, 56] with the 'nt' database and default parameters, and removed any transcripts classified as viruses or prokaryotes in order to remove any potential contaminants.

Fitting the Model

For each transcriptome in our dataset we used the annotations provided [54] (generated using dammit [61]) to locate coding sequence corresponding to ribosomal proteins. For each genome in our dataset we searched among translated coding sequences for ribosomal proteins using blastp v2.10.1 [62] against a custom blast database of ribosomal proteins of eukaryotic microbes drawn from the Ribosomal Protein Gene Database (all genes coding for ribosomal proteins available from *Dictyostellium discoideum, Giardia lamblia, Phaeodactylum tricornutum, Plasmodium falciparum, Thalassiosira pseudonana*, and *Toxoplasma gondii*; S2 Table; [63]). In all downstream analyses we omitted any genomes or transcriptomes with fewer than 10 ribosomal proteins detected [20, 22].

For each coding sequence corresponding to a ribosomal protein in each genome or transcriptome we calculated the MILC statistic of codon usage bias [39] using the coRdon R package [44], the same as done for prokaryotic gRodon [20]. This statistic is both GC-content and length corrected and should be insensitive to both factors. For these calculations the expected codon usage was taken as the genome-wide average (across all coding sequences in a genome or transcriptome; [20]). As recommended in the coRdon documentation, in order to get a reliable estimate of codon bias we removed all genes with fewer than 80 codons. We then calculated the median bias across all genes coding for ribosomal proteins for each genome or transcriptome.

We then fit a linear model (`lm()` function from the base R stats package [64]) to Box-Cox transformed doubling times (with the optimal λ chosen using the `boxcox()` function from the MASS package [65]) using (1) optimal growth temperature, and (2)

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the median codon usage bias of genes coding for ribosomal proteins (see above) as predictors. We then implemented this model into the existing gRodon package for prokaryotic growth rate prediction, expanding the package's predictive range to eukaryotic organisms (using the new `mode=''eukaryotes''` setting; `https://github.com/jlw-ecoevo/gRodon2`).

For comparison with prokaryotic models we ran genomes and transcriptomes through growthpred (obtained as a docker image at `https://hub.docker.com/r/shengwei/growthpred`; [66]) and gRodon v1.0.0 on metagenome mode (the prokaryotic setting most similar to both growthpred and our eukaryotic model; [20, 22]), including the recorded optimal temperatures for prediction.

Estimating Growth Rates from MAGs

We obtained a set of 1669 eukaryotic MAGs assembled from the Tara Oceans metagenomic surveys by two groups [15, 16]. Previously, EukMetaSanity had been run on these MAGs to call genes and find coding sequences [11]. We used these annotations, and (similar to above) searched for ribosomal proteins using blastp v2.10.1 [62] against a custom blast database of ribosomal proteins of eukaryotic microbes drawn from the Ribosomal Protein Gene Database [63]. We ran EUKulele v1.0.6 to classify these MAGs taxonomically and omitted any organisms identified as Metazoa from downstream analyses [10]. Division-level classifications were taken as the division assigned as most likely by eukulele. After removing any MAGs with less than 10 ribosomal proteins detected or that were classified as Metazoa, we were left with a total of 465 eukaryotic MAGs.

To infer the optimal growth temperatures of each MAG we used distributional data across the Tara Oceans metagenomes. For MAGs from Delmont et al. [16], optimal temperatures had already been predicted by the authors as part of a machine-learning pipeline implemented to discover each MAGs niche. For the Alexander et al. [15] MAGs we took a simpler approach. For each MAG we took the top 1% of samples in terms of MAG relative abundance and calculated the mean temperature recorded for those samples (S8 Figure). For closely related MAGs found in both Delmont et al. [16] and Alexander et al. [15], we found that the two methods for estimating optimal growth temperature agreed well (S9 Figure).

Finally, we calculated maximal growth rate using gRodon v2.0.0 in eukaryote mode. Ridgeline plots were generated using R package ggridges [67].

Estimating Growth Rates from Metagenomes

Assemblies of the bioGEOTRACES metagenomes were obtained from Biller et al [47]. We then ran EukRep v0.6.6 on these assemblies to classify contigs as eukaryotic or prokaryotic (using settings `-m strict --tie prok`; [68]).

In order to call and annotate genes from prokaryotic contigs we ran prokka v1.14.6 (using settings `--norrna --notrna --metagenome --cpus 80 --centre X --compliant`; [69]). Ribosomal proteins were identified from prokka annotations. We then predicted the average community-wide maximal growth rate of the prokaryotic community using gRodon v2.0.0 in metagenome mode using the temperature metadata provided with the samples [20].

In order to call genes from eukaryotic contigs we ran MetaEuk v4.a0f584d (using setting `easy-predict`; [12]). Similar to our procedure with individual genomes, we searched for ribosomal proteins among translated proteins output from MetaEuk using blastp v2.10.1 [62] against a custom blast database of ribosomal proteins of eukaryotic microbes drawn from the Ribosomal Protein Gene Database [63]. We then predicted the

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

average community-wide maximal growth rate of the eukaryotic community using gRodon v2.0.0 in eukaryote mode using the temperature metadata provided with the samples.

Phylogeny

We ran BUSCO v5.2.2 against the `eukaryota_odb10` database on translated proteins from the complete set of MAGs described above as well as the decontaminated MMETSP transcriptomes [14]. For the purposes of tree-building only, we removed any organisms without at least 50% of BUSCO gene families present (out of 255). We then identified any gene families present in at least 80% of remaining organisms, yielding 51 gene families. We aligned each of these families using MUSCLE v3.8.31 with default parameters [70] and trimmed our alignments with trimal v1.4.rev15 (using setting `-automated1`; [71]). Alignments were concatenated (using `https://github.com/nylander/catfasta2phyml`) and trimmed again using trimal v1.4.rev15 (using setting `-automated1`). We then used Fasttree v2.1.10 to infer a phylogeny (using default settings; [72]).

We visualized our phylogeny using R package ggtree [73] and calculated patristic distance using R package ape [74].

Acknowledgments

J.L.W. was supported by a postdoctoral fellowship in marine microbial ecology from the Simons Foundation (Award 653212). A.I.K. was supported by the Computational Science Graduate Fellowship (DOE; DE-SC0020347). H.A was supported by a National Science Foundation grant (OCE-1948025). We also acknowledge support from Simons Foundation Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems (CBIOMES) Grant 549943 (to J.A.F.) and US NSF Division of Ocean Sciences (OCE) Grant 1737409 (to J.A.F.).

References

1. Caron DA, Countway PD, Jones AC, Kim DY, Schnetzer A. Marine protistan diversity. Annual review of marine science. 2012;4:467–493.

2. Geisen S, Mitchell EA, Wilkinson DM, Adl S, Bonkowski M, Brown MW, et al. Soil protistology rebooted: 30 fundamental questions to start with. Soil Biology and Biochemistry. 2017;111:94–103.

3. WHO. World malaria report 2020: 20 years of global progress and challenges. In: World malaria report 2020: 20 years of global progress and challenges; 2020.

4. Radwanska M, Vereecke N, Deleeuw V, Pinto J, Magez S. Salivarian trypanosomosis: a review of parasites involved, their global distribution and their interaction with the innate and adaptive mammalian host immune system. Frontiers in immunology. 2018;9:2253.

5. Bar-On YM, Milo R. The biomass composition of the oceans: a blueprint of our blue planet. Cell. 2019;179(7):1451–1454.

6. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components. science. 1998;281(5374):237–240.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

7. Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. Science. 2015;347(6223).

8. Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. Protists are microbes too: a perspective. The ISME journal. 2009;3(1):4–12.

9. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS biology. 2014;12(6):e1001889.

10. Krinos AI, Hu SK, Cohen NR, Alexander H. EUKulele: Taxonomic annotation of the unsung eukaryotic microbes. Journal of Open Source Software. 2021;6(57):2817. doi:10.21105/joss.02817.

11. Neely CJ, Hu SK, Alexander H, Tully BJ. The high-throughput gene prediction of more than 1,700 eukaryote genomes using the software package EukMetaSanity. bioRxiv. 2021;.

12. Karin EL, Mirdita M, Söding J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 2020;8(1):1–15.

13. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome biology. 2019;20(1):1–13.

14. Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv preprint arXiv:210611799. 2021;.

15. Alexander H, Hu SK, Krinos AI, Pachiadaki M, Tully BJ, Neely CJ, et al. Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton. bioRxiv. 2021;.

16. Delmont TO, Gaia M, Hinsinger DD, Fremont P, Vanni C, Guerra AF, et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. bioRxiv. 2021; p. 2020–10.

17. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, et al. Unifying the known and unknown microbial coding sequence space. bioRxiv. 2021; p. 2020–06.

18. Gilbert JA, Dupont CL. Microbial metagenomics: beyond the genome. Annual review of marine science. 2011;3:347–371.

19. New FN, Brito IL. What Is Metagenomics Teaching Us, and What Is Missed? Annual Review of Microbiology. 2020;74:117–135.

20. Weissman JL, Hou S, Fuhrman JA. Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. Proceedings of the National Academy of Sciences. 2021;118(12).

21. Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, et al. The genomic basis of trophic strategy in marine bacteria. Proceedings of the National Academy of Sciences. 2009;106(37):15527–15533.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

22. Vieira-Silva S, Rocha EP. The systemic imprint of growth and its uses in ecological (meta) genomics. PLoS genetics. 2010;6(1).

23. Lim EL, Dennett MR, Caron DA. The ecology of Paraphysomonas imperforata based on studies employing oligonucleotide probe identification in coastal water samples and enrichment cultures. Limnology and oceanography. 1999;44(1):37–51.

24. Hohl HR, Raper KB. Nutrition of cellular slime molds I: growth on living and dead bacteria. Journal of bacteriology. 1963;85(1):191–198.

25. Valach M, Gonzalez Alcazar JA, Sarrasin M, Lang BF, Gray MW, Burger G. An unexpectedly complex mitoribosome in Andalucia godoyi, a protist with the most bacteria-like mitochondrial genome. Molecular biology and evolution. 2021;38(3):788–804.

26. Zheng S, Wang G, Lin S. Heat shock effects and population survival in the polar dinoflagellate Polarella glacialis. Journal of experimental marine biology and ecology. 2012;438:100–108.

27. Sato N, Yoshitomi T, Mori-Moriyama N. Characterization and biosynthesis of lipids in Paulinella micropora MYN1: evidence for efficient integration of chromatophores into cellular lipid metabolism. Plant and Cell Physiology. 2020;61(5):869–881.

28. Kremer CT, Thomas MK, Litchman E. Temperature-and size-scaling of phytoplankton population growth rates: Reconciling the Eppley curve and the metabolic theory of ecology. Limnology and oceanography. 2017;62(4):1658–1670.

29. Bissinger JE, Montagnes DJ, harples J, Atkinson D. Predicting marine phytoplankton maximum growth rates from temperature: Improving on the Eppley curve using quantile regression. Limnology and Oceanography. 2008;53(2):487–493.

30. Brush MJ, Brawley JW, Nixon SW, Kremer JN. Modeling phytoplankton production: problems with the Eppley curve and an empirical alternative. Marine Ecology Progress Series. 2002;238:31–45.

31. Goldman JC, Carpenter EJ. A kinetic approach to the effect of temperature on algal growth 1. Limnology and Oceanography. 1974;19(5):756–766.

32. Eppley RW. Temperature and phytoplankton growth in the sea. Fish bull. 1972;70(4):1063–1085.

33. Rose JM, Caron DA. Does low temperature constrain the growth rates of heterotrophic protists? Evidence and implications for algal blooms in cold waters. Limnology and Oceanography. 2007;52(2):886–895.

34. Hockenberry AJ, Stern AJ, Amaral LA, Jewett MC. Diversity of translation initiation mechanisms across bacterial species is driven by environmental conditions and growth demands. Molecular biology and evolution. 2018;35(3):582–592.

35. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. Journal of molecular biology. 1981;151(3):389–409.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

36. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. Nucleic acids research. 1982;10(22):7055–7074.

37. Dong H, Nilsson L, Kurland CG. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. Journal of molecular biology. 1996;260(5):649–663.

38. Hooper SD, Berg OG. Gradients in nucleotide and codon usage along Escherichia coli genes. Nucleic acids research. 2000;28(18):3517–3523.

39. Supek F, Vlahoviček K. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC bioinformatics. 2005;6(1):182.

40. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proceedings of the National Academy of Sciences. 2018;115(21):E4940–E4949.

41. Qian W, Yang JR, Pearson NM, Maclean C, Zhang J. Balanced codon usage optimizes eukaryotic translational efficiency. PLoS genetics. 2012;8(3):e1002603.

42. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics. 2011;12(1):32–42.

43. Subramanian S. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics. 2008;178(4):2429–2432.

44. Elek A, Kuzman M, Vlahovicek K. coRdon: Codon Usage Analysis and Prediction of Gene Expressivity; 2020. Available from: `https://github.com/BioinfoHR/coRdon`.

45. Snoeijs P, Busse S, Potapova M. THE IMPORTANCE OF DIATOM CELL SIZE IN COMMUNITY ANALYSIS1. Journal of Phycology. 2002;38(2):265–281.

46. Gaines G, Elbrächter M, Taylor F. The biology of dinoflagellates. Blackwell Scientific Publications, Oxford, UK; 1987.

47. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, et al. Marine microbial metagenomes sampled across space and time. Scientific data. 2018;5(1):1–7.

48. Waite AM, Thompson PA, Harrison PJ. Does energy control the sinking rates of marine diatoms? Limnology and Oceanography. 1992;37(3):468–477.

49. Waite A, Fisher A, Thompson PA, Harrison PJ. Sinking rate versus cell volume relationships illuminate sinking rate control mechanisms in marine diatoms. Marine Ecology Progress Series. 1997;157:97–108.

50. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, et al. A global ocean atlas of eukaryotic genes. Nature communications. 2018;9(1):1–13.

51. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. Available from: `https://ggplot2.tidyverse.org`.

52. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots; 2020. Available from: `https://CRAN.R-project.org/package=ggpubr`.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

53. Thomas MK, Kremer CT, Klausmeier CA, Litchman E. A global pattern of thermal adaptation in marine phytoplankton. Science. 2012;338(6110):1085–1088.

54. Johnson LK, Alexander H, Brown CT. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. GigaScience. 2019;8(4):giy158.

55. Simion P, Belkhir K, François C, Veyssier J, Rink JC, Manuel M, et al. A software tool 'CroCo'detects pervasive cross-species contamination in next generation sequencing data. BMC biology. 2018;16(1):1–9.

56. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology. 2014;15(3):1–12.

57. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences. 2020;117(17):9451–9457.

58. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology. 2017;35(11):1026–1028.

59. Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR genomics and bioinformatics. 2020;2(2):lqaa026.

60. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic acids research. 2005;33(20):6494–6506.

61. Scott C. dammit: an open and accessible de novo transcriptome annotator. in prep. 2016;.

62. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC bioinformatics. 2009;10(1):1–9.

63. Nakao A, Yoshihama M, Kenmochi N. RPG: the ribosomal protein gene database. Nucleic acids research. 2004;32(suppl_1):D168–D170.

64. R Core Team. R: A Language and Environment for Statistical Computing; 2020. Available from: `https://www.R-project.org/`.

65. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002. Available from: `http://www.stats.ox.ac.uk/pub/MASS4`.

66. Long AM, Hou S, Ignacio-Espinoza JC, Fuhrman J. Benchmarking metagenomic marine microbial growth prediction from codon usage bias and peak-to-trough ratios. bioRxiv. 2019; p. 786939.

67. Wilke CO. ggridges: Ridgeline Plots in 'ggplot2'; 2021. Available from: `https://CRAN.R-project.org/package=ggridges`.

68. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. Genome-reconstruction for eukaryotes from complex natural microbial communities. Genome research. 2018;28(4):569–580.

69. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–2069.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

70. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792–1797.

71. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–1973.

72. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one. 2010;5(3).

73. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017;8(1):28–36.

74. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35:526–528.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

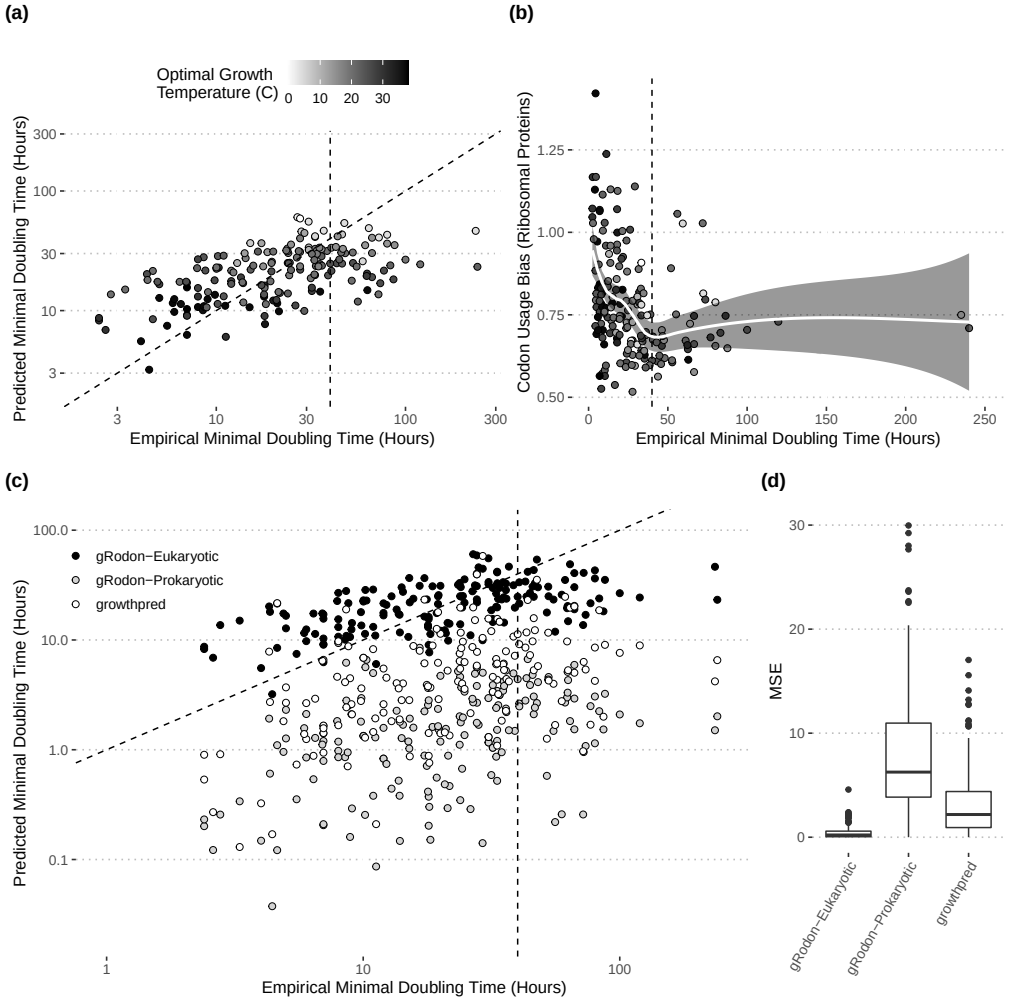

Fig 1. Codon usage bias (CUB) and temperature predict the maximum growth rates of microbial eukaryotes. (a) Predictions from a linear model of minimum doubling time with CUB and temperature as predictors on our training set generally reflect empirically observed doubling times ($r^2 = 0.328$). (b) The relationship between CUB and minimum doubling time is roughly linear and negative until approximately 40 hours, after which the relationship levels off ($\rho = -0.400$, $p = 3.14 \times 10^{-8}$). (c) Predictions of the maximum growth rates of microbial eukaryotes on the basis of CUB and temperature using models trained on prokaryotes are systematically biased towards faster growth predictions and (d) perform much worse than a model trained directly on eukaryotes in terms of mean squared error (MSE). Dashed vertical lines denotes 40 hours and dashed diagonal line denotes where $x = y$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

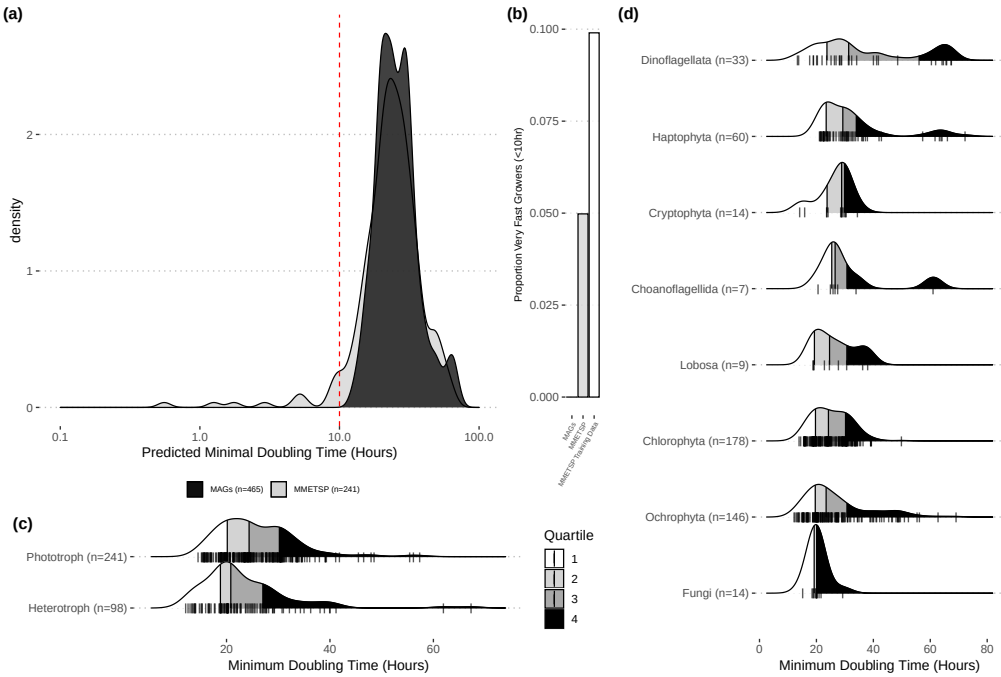

Fig 2. Environmentally derived genomes of eukaryotic microbes reveal differences in growth potential across sampling sources, lifestyles, and taxonomic groups. (a) The distribution of predicted minimum doubling times of organisms represented in the MMETSP ($n = 241$) is slightly shifted towards faster maximum growth rates as compared to the distribution of predicted minimum doubling times among marine eukaryotic microbes represented by MAGs (27.5 vs 24.5 hours respectively; t-test, $p = 9.25 \times 10^{-4}$; $n = 465$). (b) This difference in growth potential is primarily driven by a small number of very fast growing organisms (minimum doubling time < 10 hours) in MMETSP. (c) MAGs from organisms predicted to be heterotrophic were associated with faster maximum growth rates than those predicted to be phototrophic (t-test, $p = 1.58 \times 10^{-3}$). (d) Different taxonomic groups have distinct distributions of predicted growth potentials among their members, as predicted from MAGs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.15.464604; this version posted June 26, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

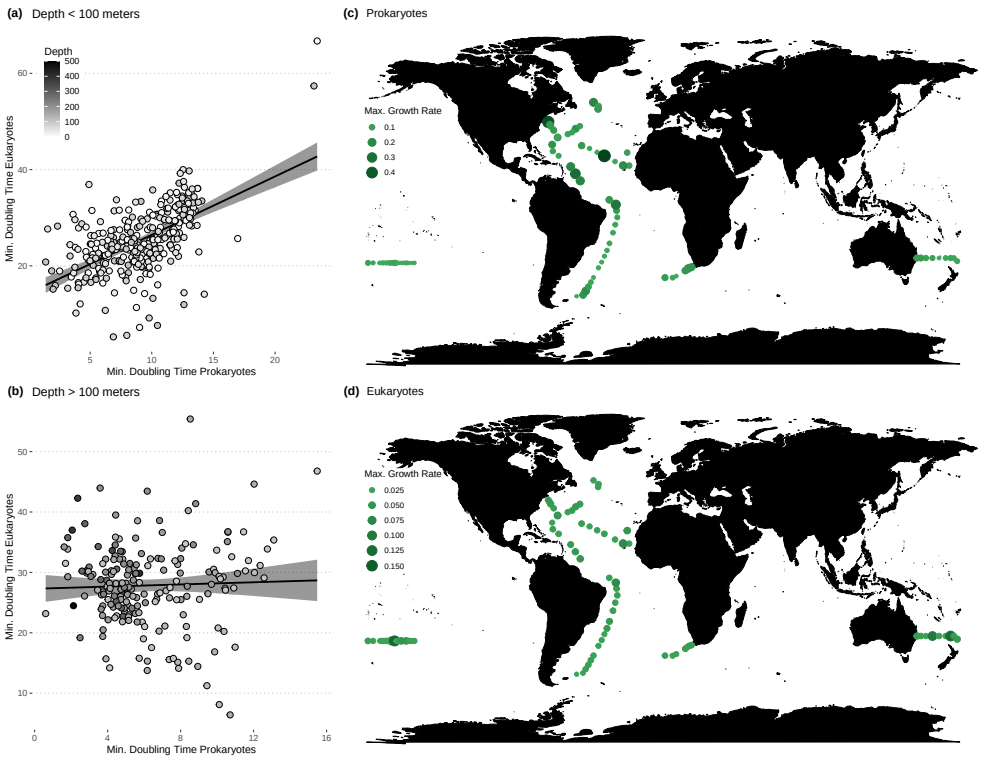

Fig 3. Bulk community-wide prediction of growth potential from metagenomes yields insights into global patterns of eukaryotic and prokaryotic growth in the oceans. (a) The average community-wide maximum growth rate of eukaryotic and prokaryotic communities are strongly correlated near the ocean surface (< 100 meters; $\rho = 0.566$, $p = 1.08 \times 10^{-27}$), but (b) not at deeper depths. (c,d) Average community-wide growth rates near the surface for eukaryotes and prokaryotes vary substantially across the global oceans (< 100 meters).