Estimating global variation in the maximum 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 organisms and communities directly in the environment. We produced a database of over 700 maximum growth rate predictions from genomes, transcriptomes, and metagenome-assembled genomes. 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 there is no correlation in their genomic potentials 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.

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 oceans in particular, protists dominate, accounting for approximately 30% of total marine biomass and of primary producer biomass as well [5]. Marine systems account for about half of all global primary production [6], hence the abundance of protists in these systems indicates a central role for protists in regulating global biogeochemical cycles [7]. And yet, our tools for studying the ecology and evolution of eukaryotic microbes are still quite limited, at least in comparison to their prokaryotic relatives [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, and of entire mixed-species communities, directly from their genome or metagenome 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 global variation of eukaryotic growth potential across organisms and environments.

Results and Discussion

Predicting maximal growth rates of eukaryotic microbes

We compiled a database 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 rates than the marine

eukaryotic microbes found in MMETSP (S1 and S2 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. There are many forces acting on codon bias – including mutational biases and selection for mRNA stability [35] – but the codon usage patterns of genes are also thought to be optimized to the relative abundance of tRNAs within the cell to facilitate rapid translation. This optimization is particularly apparent among highly-expressed genes in fast-growing species [36–43]. The basic intuition is that the optimization of genes for rapid translation leads to increased CUB, and in turn facilitates rapid growth. We wanted to see whether such patterns can be leveraged to predict the growth rates of eukaryotic microbes [44]. 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, 40, 45]). 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 correlation with Box-Cox transformed doubling times, $\rho = -0.47$, $p = 5.64 \times 10^{-11}$). 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; S2 Fig). Optimal growth temperature has been seen to be an important parameter predicting the maximum growth rates of prokaryotes that is important to control for in genomic predictors [20, 22]. We found that our model could explain about a third of variation in doubling time among organisms (linear regression, $r^2 = 0.337$), and that both CUB ($p = 4.48 \times 10^{-8}$) and optimal growth temperature ($p = 3.18 \times 10^{-8}$) were significant predictors in the model. Interestingly, similar to what we have previously reported for prokaryotes [20], we saw that for eukaroytic microbes, the relationship between CUB and doubling time may saturate at a threshold doubling time, after which CUB no longer changes 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]. It appears that a similar dynamic may be at work among eukaryotes, although the small number of sequenced eukaryotic microbes with long minimal doubling times (greater than 30-40 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 (30-40 hours versus 5 hours respectively; [20]), likely due to constraints on eukaryotic growth related to cell size and complexity. It is possible that predicted maximum growth rates for very slow growing organisms will be overestimated using our tool, though we lack strong evidence that this will be the case. To determine pragmatic cutoffs for users, we redefined the doubling time cutoffs for reliable prokaryotic and eukaryotic prediction in terms of the CUB rather than the predicted doubling times to get temperature-independent cutoffs appropriate for organisms growing in colder environments (see Methods). A clear limitation of our approach is that we can never say for sure whether an organism's experimentally measured

maximum growth rate is its actual maximum growth rate, which may in part explain why organisms with slow empirical growth measurements sometimes have faster maximum growth rates inferred genomically.

We asked whether our eukaryotic model of growth improved predictions for eukaryotic organisms relative to the predictions made by previous tools developed for and trained exclusively on prokaryotes [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 magnitude, indicating 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]).

Finally, we considered the possibility that our model was overfit to the data, which would mean our predictor would perform poorly on new datasets. Overfitting is a particularly relevant concern when dealing with species data where shared evolutionary history can induce a hierarchical dependency structure between datapoints. In addition to traditional cross-validation, we applied a blocked cross-validation approach, which controls for phylogenetic structure by taking each phylum in our dataset as a fold to hold out for independent error estimation rather than holding out random subsets of our data as in traditional cross-validation. Even under these controls, eukaryotic gRodon showed a clear improvement over prokaryotic models as well as against prediction using the average growth rate of a group (S3 and S4 Figs). This suggests CUB-growth relationships extrapolate well across phyla. Given that the genomes taken from NCBI and transcriptomes from MMETSP are from very different sets of microbes, we wondered whether the choice of data source would impact our model. We trained models on each source dataset and saw good cross-prediction between training datasets (S5 Fig).

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 similar, though marginally distinguishable, doubling times to cultured organisms in MMETSP (means of 26.1 and 25.1 hours respectively; Kolmogorov-Smirnov test, $p = 0.0187$; 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 good job at capturing the distribution of growth rates among organisms, though they are slightly enriched for fast growing organisms that are relatively 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]. A caveat to these analyses is that while MAGs are likely to better approximate the natural distribution of organisms in the environment than culture collections, the binning process used to construct MAGs may also be biased, meaning that MAGs do not perfectly represent natural populations of microbes. That being said, we found only extremely weak (and likely spurious) associations between completeness and inferred growth rate across MAGs (and no or an opposite pattern was seen for relative abundance to that of growth rate; S6 Fig), suggesting no strong association between the ability to bin an organism's genome and its maximum growth rate.

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 (Mann-Whitney U Test, $p = 7.65 \times 10^{-6}$; 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 [46, 47]. Overall, the distribution of maximal growth rates varied across taxonomic groups, likely a product of both specialization for different niches and historical contingency.

Finally, we looked at the maximum growth rate and abundance of each MAG across the Tara Oceans dataset [15, 48] to get group-specific maximum growth rates across the global oceans. We found the average maximum growth rates of several major groups of microbial eukaryotes with many representative MAGs by taking an abundance-weighted maximum growth rate for each Tara surface sample (Fig 2e-h; < 100 meters). Groups varied in their global growth patterns, but in general the Indian Ocean was a hotspot for organisms with high maximum growth rates and the Southern Ocean was dominated by organisms with low maximum growth rates. Ochrophyta (Fig 2f) and Dinoflagellata (Fig 2e) communities varied widely in their growth rates, whereas Chlorophyta (Fig 2g) community maximum growth rates varied less across samples. Haptophyta (Fig 2h) communities generally had the lowest maximum growth rates seen across samples.

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 MAG representatives of every community member in a particular environment, it is possible to apply CUB-based predictors to a metagenome to estimate the average maximum growth rate of that community [22, 49]. 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], and we recently updated and benchmarked metagenome mode v2 for improved community-level prediction [49]. Our eukaryotic model can be similarly applied to calculate the median maximal growth rate of the eukaryotic community represented in a metagenomic sample (see Methods for details). We benchmarked this approach against synthetic mixtures of genomes meant to simulate mixed-species communities and saw good performance predicting community-wide averages, in line with our previous results for prokaryotes (see S8-S9 Figs and Methods).

To demonstrate this application, we obtained assemblies of 610 globally-distributed marine metagenomic samples from the BioGEOTRACES dataset [50]. This dataset is particularly useful for our purposes because samples were not size-fractionated, allowing both prokaryotic and eukaryotic communities to be assessed simultaneously. 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; S10 Fig). The correlation between the growth potentials of prokaryotic and eukaryotic communities was striking (Pearson correlation of samples from < 100 meters, $\rho = 0.457$, $p = 1.74 \times 10^{-27}$; Fig 3a-b), though this relationship varied with depth (Fig 3a-d). A linear model confirmed a significant interaction between depth and the relationship between eukaryotic and prokaryotic growth rates (linear regression of prokaryotic growth rates, $\beta_{\text{eukaryotes}} = 0.185$, $p_{\text{eukaryotes}} = 2.81 \times 10^{-11}$, $\beta_{\text{depth}} = 4.78 \times 10^{-4}$, $p_{\text{depth}} = 2.02 \times 10^{-3}$, $\beta_{\text{eukaryotes:depth}} = -1.53 \times 10^{-4}$, $p_{\text{eukaryotes:depth}} = 0.0120$). Notably, these correlations were not simply byproducts of temperature shifts, as the CUB of eukaryotic and prokaryotic communities co-varied across surface samples (though correlations among deeper samples could be explained by temperature; Fig 3c-d). Additionally, these patterns could noy be attributed to differences in coverage. While doubling time did decrease with the relative abundance of eukaryotic contigs in a sample, as would be expected, samples with a lower proportion of eukaryotes were not particularly skewed in their estimated growth rates (S11 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 CUB of eukaryotic and prokaryotic communities with increasing depth aligns with a shift towards increasingly heterotophic eukaryotic communities (S13). Prokaryotic communities at the surface have correlated CUB with deeper prokaryotic communities, but eukaryotic communities are not correlated across depths, nor with prokaryotic communities at other depths (S12 Fig). Altogether, this suggests that any physical processes linking the surface to deeper depths (e.g., sinking particles) may occur on a different timescale to processes shaping the community growth potential (e.g., community assembly).

Conclusions

We developed and validated a new tool to estimate the growth potential of eukaryotic microbes directly from genomic, transcriptomic, and metagenomic 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.

We emphasize that gRodon estimates the maximum growth rate of an organism or the average maximum growth rate of a community, not the instantaneous growth rate at any given time. In the wild, it is likely that microbes rarely reach these maximum rates. Nevertheless, this parameter gives us an idea about the ecological role of an organism, in particular whether it has undergone selection for the ability to replicate rapidly. Microbial ecologists often conceptualize organisms' growth strategies on a spectrum from a "boom-bust" strategy favoring rapid maximum growth rates versus a "slow-and-steady" strategy with slower maximum growth rates, and call organisms with these two strategy sets "copiotrophs" and "oligotrophs" respectively [20, 21, 51]. In our previous work on prokaryotes, we found that these distinct growth classes have different evolutionary patterns, gene content, and functional profiles [20]. When applied at the community level, the average community-wide maximum growth rate can be thought of as an "index of copiotrophy" which measures the relative frequency of these two distinct growth classes in the community [49].

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 [52] and ggpubr [53] R packages. For mapping, we used the maps package [54], as well as the automap package for spatial interpolation (universal blocked kriging with the autoKrige() function; [55]). Ridgeline plots were generated using R package ggridges [56].

See S1.1 Text for supplemental methods describing how we generated the datset used to train our model (and the datasets and programs used [11–13, 33, 57–64]).

See S1.2-1.3 Text and S14-S15 Figs for supplemental methods describing the processes for predicting maximum growth rates from MAGs and Metagenomes using eukaryotic gRodon (and the datasets and programs used [10–12, 14–16, 20, 49, 50, 65–71]).

Fitting the Model

For each MMETSP transcriptome in our training dataset we used the annotations provided [58] (generated using dammit [72]) to locate coding sequence corresponding to ribosomal proteins. For each GenBank genome in our training dataset we searched among translated coding sequences for ribosomal proteins using blastp v2.10.1 [65] 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; [66]). 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 (Measure Independent of Length and Composition) statistic of CUB [40] using the coRdon R package [45], 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 CUB across all genes coding for ribosomal proteins for each genome or transcriptome.

We then fit a linear model to Box-Cox transformed doubling times (with the optimal λ chosen using the `boxcox()` function from the MASS package [73]) using (1) optimal growth temperature, and (2) the normalized median CUB of genes coding for ribosomal proteins

$$\text{Normalized CUB} = \frac{\overline{\text{MILC}}_{\text{All Genes}} - \overline{\text{MILC}}_{\text{Highly Expressed Genes}}}{\overline{\text{MILC}}_{\text{All Genes}}}$$

(see [22, 49] for details of normalization procedure) 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 gRodon's eukaryotic metagenome mode the codon-usage bias was calculated separately on a per-gene basis prior to normalization, as done with prokaryotic metagenome mode v2 [49].

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`; [74]) 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.

Appropriate CUB cutoffs for prediction were determined by taking a maximum growth rate model trained only on mesophilic organisms (optimal growth temperature between 20C and 60C), but otherwise not accounting for temperature (otherwise as above). We then determined at what CUB value the model predicted a doubling time of 30 hours (approximately where the CUB vs growth relationship saturates; Fig 1b; normalized CUB of 0.012 for eukaryotes, a similar procedure with a 5 hour threshold identified a CUB cutoff of 0.59 for the original prokaryotic model [20]). These cutoffs can be used to determine whether a maximum growth estimate from gRodon is likely to be an overestimate, independent of the growth temperature of an organism (e.g., Fig 3e-f).

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.

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.

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. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, et al. Codon optimality is a major determinant of mRNA stability. Cell. 2015;160(6):1111–1124.

36. 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.

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

38. 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.

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

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

41. 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.

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

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

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

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

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

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

48. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348(6237):1261359.

49. Weissman J, Peras M, Barnum TP, Fuhrman JA. Benchmarking community-wide estimates of growth potential from metagenomes using codon usage statistics. accepted, mSystems. 2022;.

50. 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.

51. Giovannoni SJ, Thrash JC, Temperton B. Implications of streamlining theory for microbial ecology. The ISME journal. 2014;8(8):1553–1565.

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

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

54. code by Richard A Becker OS, version by Ray Brownrigg Enhancements by Thomas P Minka ARWR, Deckmyn A. maps: Draw Geographical Maps; 2021. Available from: `https://CRAN.R-project.org/package=maps`.

55. Hiemstra PH, Pebesma EJ, Twenhöfel CJW, Heuvelink GBM. Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network. Computers Geosciences. 2008;.

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

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

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

59. 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.

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

61. 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.

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

63. 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.

64. 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.

65. 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.

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

67. 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.

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

69. 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.

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

71. 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.

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

73. 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`.

74. 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.

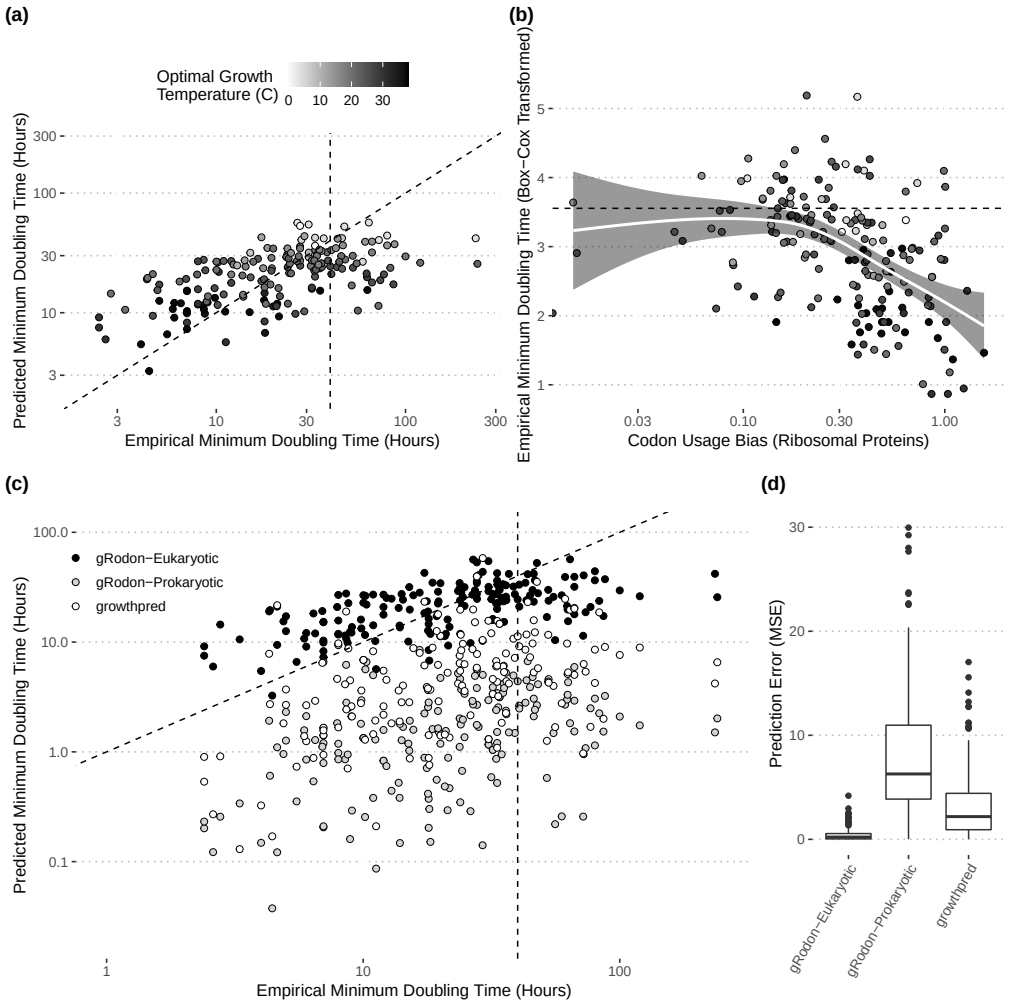

Fig 1. Codon usage bias (CUB) and temperature predict the maximum growth rates of microbial eukaryotes. (a) Predictions from a linear model of minimum doubling time with CUB and temperature as predictors on our training set generally reflect empirically observed doubling times ($r^2 = 0.328$). (b) The relationship between CUB and minimum doubling time is roughly linear and negative until approximately 30-40 hours (dashed line at 40 hours; Pearson correlation, $\rho = -0.48$, $p = 3.74 \times 10^{-9}$), after which the relationship levels off (Pearson correlation, $\rho = -0.099$, $p = 0.55$). Gray line is a smoothing line generated with ggplot2. (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 $x = y$.

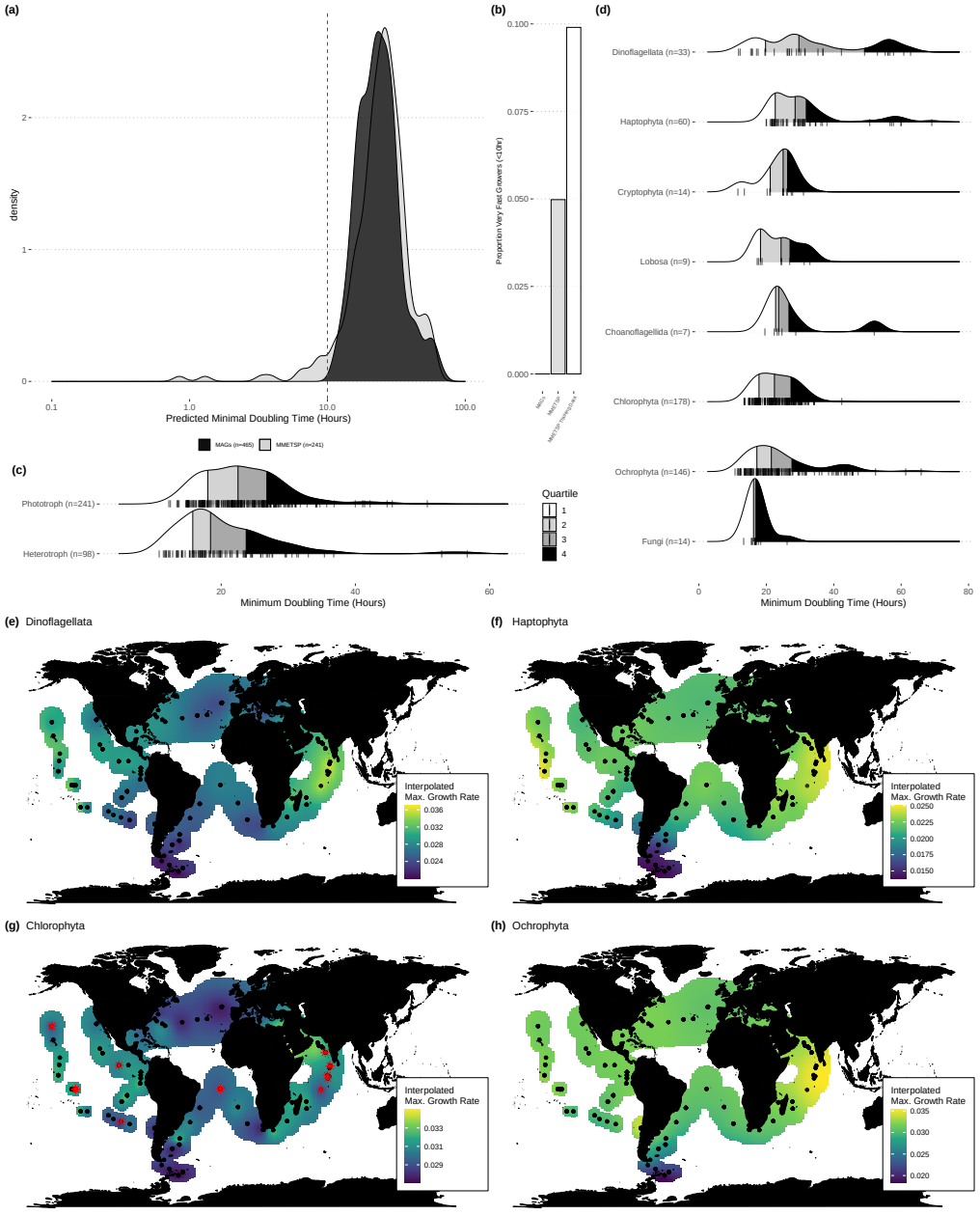

Fig 2. Environmentally derived genomes of eukaryotic microbes reveal differences in growth potential across sampling sources, lifestyles, and taxonomic groups. (a) The distribution of predicted minimum doubling times of organisms represented in the MMETSP is nearly indistinguishable from the distribution of predicted minimum doubling times among marine eukaryotic microbes represented by MAGs. (b) Even so, there are small number of very fast growing organisms (minimum doubling time < 10 hours) in MMETSP that form a long tail absent in the MAG data. (c) MAGs from organisms predicted to be heterotrophic were associated with faster maximum growth rates than those predicted to be phototrophic. (d) Different taxonomic groups have distinct distributions of predicted growth potentials among their members, as predicted from MAGs. (e-h) Using the abundance of MAGs from groups with many representatives mapped back to the Tara Oceans surface metagenomes (< 100 meters) we found group-specific patterns in average maximum growth rate (h^{-1}) across the world's oceans. Red asterisks denote possible overestimates due to saturation of the CUB-growth relationship. For panels (a-d) see S7 Fig for an analysis of possible overestimated rates, which cannot account for observed patterns.

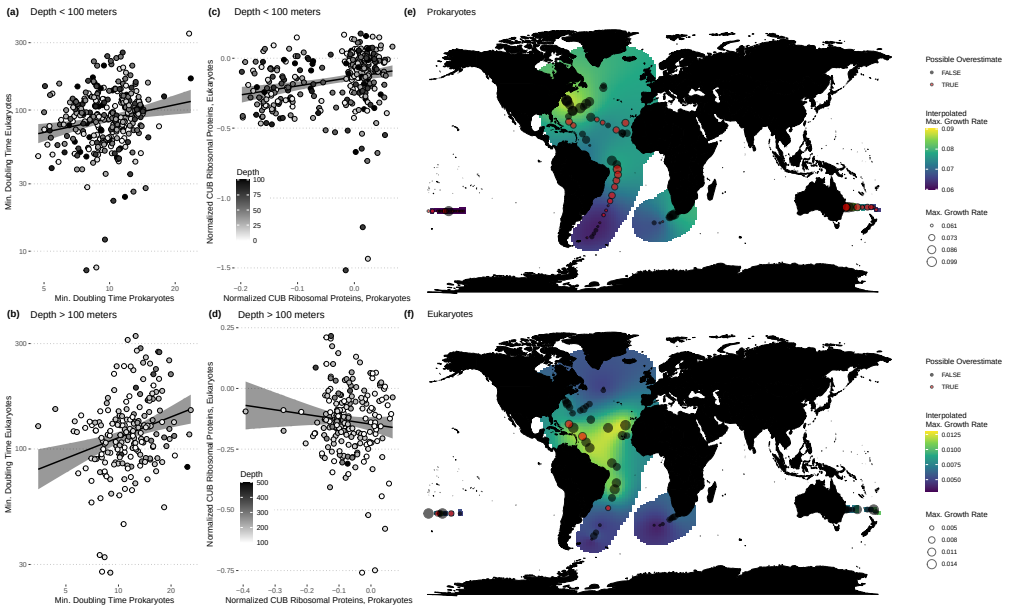

Fig 3. Bulk community-wide prediction of growth potential from metagenomes yields insights into global patterns of eukaryotic and prokaryotic growth in the oceans. (a-b) The average community-wide maximum growth rate of eukaryotic and prokaryotic communities are correlated near the ocean surface (< 100 meters; $\rho = 0.240$, $p = 2.48 \times 10^{-5}$), and at deeper depths (> 100 meters; $\rho = 0.457$, $p = 5.83 \times 10^{-12}$). (c-d) Yet this correlation is at least partly driven by temperature, particularly for deeper samples. While (c) the normalized codon usage bias of eukaryotic and prokaryotic communities are correlated at the surface (< 100 meters; $\rho = 0.223$, $p = 7.70 \times 10^{-5}$), (d) they are not at depth (> 100 meters; $\rho = -0.0884$, $p = 0.208$). (e-f) Average community-wide growth rates near the surface for eukaryotes and prokaryotes vary substantially across the global oceans (< 100 meters) and, while associated, have distinct overall patterns.