Benchmarking community-wide estimates of growth potential from metagenomes using codon usage statistics

JL Weissman[1*], Marie Peras[2], Tyler P. Barnum[2] & Jed A. Fuhrman[1],

1 Department of Biological Sciences–Marine and Environmental Biology, University of Southern California, Los Angeles, CA, USA **2** Trace Genomics, Inc., Redwood City, CA, USA
* Corresponding author: jakeweis@usc.edu

Abstract

Trait inference from mixed-species assemblages is a central problem in microbial ecology. Frequently, sequencing information from an environment is available, but phenotypic measurements from individual community members are not. With the increasing availability of molecular data for microbial communities, bioinformatic approaches that map metagenome to (meta)phenotype are needed. Recently, we developed a tool, gRodon, that enables the prediction of the maximum growth rate of an organism from genomic data on the basis of codon usage patterns. Our work and that of other groups suggests that such predictors can be applied to mixed-species communities in order to derive estimates of the average community-wide maximum growth rate. Here, we present an improved maximum growth rate predictor designed for metagenomes that corrects a persistent GC bias in the original gRodon model for metagenomic prediction. We benchmark this predictor with simulated metagenomic datasets to show that it has superior performance on mixed-species communities relative to earlier models. We go on to provide guidance on data pre-processing, and show that calling genes from assembled contigs rather than directly from reads dramatically improves performance. Finally, we apply our predictor to large-scale metagenomic datasets from marine and human microbiomes to illustrate how community-wide growth prediction can be a powerful approach for hypothesis-generation. Altogether, we provide an updated tool with clear guidelines for users about the uses and pitfalls of metagenomic prediction of the average community-wide maximal growth rate.

Introduction

A fundamental goal of microbial ecology is to characterize the variation of complex microbial traits across time and space and to understand the drivers of that variation [1–4]. Many of our best tools for understanding natural microbial communities are molecular, and rely on our ability to leverage environmental DNA or RNA sequences to produce insights about community composition (e.g., [5, 6]). Thus, a major bioinformatic challenge for microbial ecologists is that of inferring the traits of an organism or organisms on the basis of nucleotide sequences alone (e.g., [7–12]).

One trait that has proven to be amenable to such genomic inference is the maximum growth rate of an organism, which can be predicted from various signatures of optimization for rapid translation [13–21]. Specifically, organisms capable of fast growth have clear differences in their codon usage patterns relative to organisms which can only grow slowly [13, 21, 22]. In particular, because the production of many ribosomes is a

universal requirement for rapid growth [23], one can look for codon optimization for rapid translation in the genes encoding for ribosomal proteins on a genome as a particularly reliable indicator of maximum growth rate across phyla, and even domains of life [13, 24] (and across major metabolic divides, e.g., S1 Fig). Tools have been developed that can predict the maximum growth rate of an organism on the basis of its genome sequence [13, 21], and limited benchmarking on natural and experimentally-generated datasets suggest that these methods can also be applied at the community level to metagenomes describing a mixture of organisms to predict a community-wide average maximum growth rate [13, 21, 25]. Prediction of the average community-wide maximum growth rate from codon usage statistics is often performed as part of bioinformatics analysis pipelines, to assess how the growth potential of communities vary across environmental gradients (e.g., [9]). Yet, the performance of these methods on mixed communities has not been well characterized and it is unclear when we should expect such approaches to work well and/or fail.

Here we benchmark the performance of our recently developed predictor of maximal growth rate, gRodon [21], on mixed communities of microbes in order to better understand the limitations of this tool. We implement additional bias-corrections to the method that greatly improve performance. We go on to provide guidance on data pre-processing, and show that calling genes from assembled contigs rather than directly from reads dramatically improves performance. Altogether, we provide an updated tool with clear guidelines for users on when metagenomic prediction of the average community-wide maximal growth rate should be possible. Finally, we apply gRodon to two large-scale metagenomic datasets from habitats of particular interest: (1) the global oceans [26], and (2) the human body [8, 27]. In doing so, we show that community-level maximum growth rate predictions can yield real ecological insights by (1) demonstrating a dramatic decrease in the maximum growth rate of marine communities with depth after 100 meters, and (2) demonstrating clear differences in growth potential of the human microbiome across body sites.

Results and Discussion

Implementing a Correction for GC content and Consistency Into Metagenome Mode

The original gRodon paper [21] showed that gRodon's default "full" mode for predicting growth rates from individual genomes was not affected by genomic GC content. In contrast, we found that when applied to pairs of concatenated genomes (S2 Fig; simulating two-species communities with even abundances), gRodon v1's "metagenome mode" (MMv1), which applies a simpler codon-usage model than "full" mode [21], had a persistent GC-bias in which pairs with intermediate GC ranges ($\sim 50\%$) were predicted to be associated with shorter doubling times than pairs with more extreme genomic GC contents (S2 Fig).

At the same time, we found that when predicting the average minimum doubling time of a mixed community of microbes, MMv1 performed best when the organisms in the community shared a common set of preferred codons (S3 Fig). This pattern stems from a key assumption of gRodon that microbes in the same community have more similar codon preferences to one another than microbes from different communities. This assumption is based on studies showing that this is indeed the case for many systems [28], where microbes in the same environment tend to have biases towards similar sets of codons in their genomes. It is important that this similarity assumption is met for MMv1 to perform well, because our codon-based growth predictor compares the bias in codons used by some pre-specified set of highly expressed genes (generally

taken to be the ribosomal proteins) relative to a background estimate of codon usage bias in the rest of the genes found in a metagenomic sample. If the organisms in a sample do not share a common set of codon biases, then the background estimate of bias will be underestimated, leading us to overestimate the codon usage bias of our set of highly expressed genes and underestimate the average doubling time of the community (S3 Fig). We measure the similarity of codon usage across highly expressed genes in a genome or metagenome using "consistency" as a metric [21], which describes how far on average these genes are from each other in terms of codon usage. Because consistency is defined in a somewhat counter-intuitive way (higher consistency means more dissimilar codon usage), but to also avoid changing definitions between papers, we represent consistency here as Ψ, where larger values of Ψ mean that the ribosomal proteins in a sample have more dissimilar codon usages from one another.

Based in part on developments from an earlier software, growthpred [13], we made three changes to the model used by MMv1: (1) we calculated codon-usage bias on a per-gene level, which required an alternative randomization-based approach to calculate the expected codon frequency (see Methods), (2) we applied a normalization of codon usage from growthpred that the authors found was helpful in correcting for any GC-bias in their predictor (see Methods), and (3) we explicitly included GC content as a term in our growth model. Together, these developments effectively corrected the GC and Ψ biases apparent in MMv1 (S2c and S3c Figs). In fact, these two biases are closely related, since samples with very skewed GC content also have similar codon usage patterns among their constituents (low Ψ), potentially explaining the source of the GC-bias in MMv1 (Fig 1).

On further examination, we found that under certain conditions MMv1 actually outperformed our new bias-corrected metagenome mode (MMBC) on mixed communities (S4 Fig). This occurred when organisms in a sample had very similar codon preferences ($\Psi < 0.6$). Which model performed best appeared to be dependent on a trade-off inherent to the per-gene calculation of CUB in MMBC. When calculating CUB on a single gene there is a very limited amount of sequence information available with which to calculate the expected background codon frequencies, leading to a high variance in our per-gene estimates of CUB. On the other hand, MMv1 uses the entire metagenome to calculate background codon frequencies, leading to predictions with much lower variance but that were strongly biased for communities where organisms did not share codon preferences ($\Psi > 0.6$). In cases where organisms did share codon preferences ($\Psi < 0.6$), then MMv1 had both low bias and lower variance than MMBC. Since it is straightforward to calculate Ψ for any metagenomic sample, this means that one can decide whether to use MMv1 or MMBC in order to get the best result. We implemented a new mode in the open-source gRodon tool, metagenome mode v2 (MMv2), that automatically switches between MMv1 and MMBC based on a Ψ threshold of 0.6 (S4 Fig).

Finally, we considered the possibility that our MMv2 model was overfit to the training data, which would limit our ability to apply the predictor to new datasets. These models have relatively few parameters overall (S1 Table). Yet, overfitting is a particularly salient concern when training models on species data, because species do not represent independent data points [29, 30]. In addition to random cross-validation, we implemented a blocked cross validation approach to control for phylogeny in our error estimates [21, 29]. Together, random and blocked cross validation indicated MMv2 had an equal or better ability to extrapolate across folds compared to MMv1, and did not appear to be overfit (S5 Fig). Importantly, these single-genome error estimates do not necessarily tell the whole story of a model's ability to predict the average maximum growth rate of a mixed-species community (see next section), but rather provide a lower-bound for prediction error when applied to such communities.

gRodon's Metagenome Mode v2 Outperforms Other Models on Synthetic Communities

In order to benchmark gRodon's ability to predict the average maximal growth rates of mixed-species communities we generated several community-level datasets (Table 1). To start, we generated three sets of 10,000 "genome mixtures" drawn from three sources: (1) a set of RefSeq [31, 32] genomes de-replicated at the genus level (one genome chosen per genus), (2) a reference catalog of genomes from isolates from the human gut that covers about 70% of these communities [33], (3) a set of single-cell amplified genomes (SAGs) collected from the ocean surface in an unbiased manner [34]. We distinguish "genome mixtures" from true synthetic metagenomes in that our mixtures are simply the concatenated set of coding sequences from the constituent genomes, whereas a "true" synthetic metagenome is produced via the generation of synthetic reads from genomic references (we perform such an analysis later on). In this way, we focus the current analysis on issues of prediction on mixed communities only, and do not consider possible complications arising from the sequencing or assembly process. Additionally, our concatenation approach allows us to generate very large benchmarking datasets very efficiently (30,000 total generated mixtures). We presently consider mixtures where all organisms are present in equal proportions, though see the following section for an analysis of mixtures with varying species abundances.

In order to benchmark MMv2 against the earlier MMv1 and growthpred predictors, we compared the prediction of each method run on the entire mixture to the average prediction of gRodon's "full mode" on the individual genomes in the mixture. This can be taken as an estimate of the actual error of prediction (since gRodon's full mode is, to our knowledge, the currently best-performing single-genome predictor of maximum growth rate available [21]). Unfortunately, benchmarking against empirically measured growth rates is not currently feasible as the maximum growth rates associated with the genomes used here are either not known or not compiled in any easily accessible way. Any available sets of genome-growth rate pairs have been used to train gRodon and would be inappropriate to use as a benchmark (though see S6 Fig for the results of such benchmarking, which also indicate good performance of MMv2).

In general, MMv2 shows improved performance over the earlier MMv1 and growthpred predictors in accurately capturing the average maximal growth rate of a mixture of genomes (Fig 2a-c). In particular, MMv2 greatly outperforms MMv1 when $\Psi > 0.6$ (S7 Fig; high Ψ indicates dissimilar codon usage patterns among organisms in a sample). With one exception, MMv2 predictions are not affected by GC content, while MMv1 predictions show a strong dependency on GC across all datasets (Fig 2d-f). The one notable exception is that in the case of simulated marine communities, MMv2 predicts slower maximal growth for low-GC genome mixtures. This appears to be the product of a real biological pattern, as it is a well-studied phenomenon that slow-growing marine oligotrophs have streamlined genomes with low-GC content (possibly as an adaptation to nutrient limitation; [35]). MMv2 had a slight performance boost in samples where $\Psi < 0.6$ due to the bias-variance trade-off between MMv1 and MMBC discussed in the last section (S8 Fig).

All growth prediction models (all gRodon and growthpred modes) similarly implement optimal growth temperature as an optional linear predictor which may have important impacts on maximum growth rate estimates [13, 21]. We also included temperature as an optional predictor for MMv2, and simulations that incorporate temperature variation suggest that including temperature as a variable is important for datasets taken from a wide range of temperature conditions (S9-S11 Figs).

Finally, for samples where $\Psi > 0.6$, MMv2 had longer run times than MMv1, though both methods were able to predict growth rates in a matter of minutes or less (S12 Fig). On a single core on a 2.0GHz AMD EPYC 7702P processor, on a representative set of

Table 1. Simulated community level data used to benchmark gRodon MMv2. Synthetic metagenomes (reads generated from source genomes following a realistic error model) were analyzed both at the read level and after assembly at the contig level as two separate sets of community data.

Genome Source	Simulation Type	Temperature Simulated	Rel. Abund. Simulated	Number of Sumulations	Figure
RefSeq	Genome Mixture	N	N	1.00E+04	1, S2, S3, S4, S5, S8, S9, S17, S18
RefSeq	Genome Mixture	Y	N	1.00E+04	S10
RefSeq	Genome Mixture	N	Y	1.00E+04	2.00E+00
Marine Surface	Genome Mixture	N	N	1.00E+04	1, S8, S9, S17, S18
Marine Surface	Genome Mixture	Y	N	1.00E+04	S12
Marine Surface	Genome Mixture	N	Y	1.00E+04	2.00E+00
Human Gut	Genome Mixture	N	N	1.00E+04	1, S8, S9, S17, S18
Human Gut	Genome Mixture	Y	N	1.00E+04	S11
Human Gut	Genome Mixture	N	Y	1.00E+04	2.00E+00
Human Gut	Synthetic Metagenome - Reads	N	N	1.00E+02	3, S16, S20
Human Gut	Synthetic Metagenome - Assemblies	N	N	1.00E+02	3, S16, S20
gRodon Training Data	Genome Mixture	N	N	1.00E+03	S7
Mycobacterium (GTDB207)	Genome Mixture	N	N	1.00E+03	S19
Vibrio (GTDB207)	Genome Mixture	N	N	1.00E+03	S19

metagenomes from the human microbiome [8, 27], MMv2 could predict the average community-wide maximum growth rate in 1.9 minutes on average, with a maximum run time of 5.9 minutes for a single metagenome. By contrast, MMv1 could predict the average community-wide maximum growth rate in 14 seconds on average, with a maximum run time of 52 seconds for a single metagenome. This difference in run times is a consequence of how MMv2 calculates the background codon usage for each gene individually (see Methods). Still, even for datasets with thousands of metagenomes running MMv2 should be feasible in reasonable time-frames.

gRodon's Abundance-Correction Improves Performance

One of the key innovations of MMv1 was that gRodon is able to take coverage information into account when calculating the average maximal growth rate of a community in order to effectively weight community members by their relative abundances. In order to determine whether this abundance-correction actually improved predictor performance we again simulated three sets of 10,000 genome mixtures each as above (simulated from assembled genes of each genome), this time also drawing genome abundances from a log-normal distribution for each mixture which were then assigned to each gene accordingly. We used these simulated abundances in order to calculate the "actual" average maximal growth rate used as a benchmark, and then compared our predictions when we either provided gRodon with the relative abundances for each genome or did not. We found that MMv2 had improved predictions when relative abundance information was included (Fig 3). It appears that improvements were seen primarily in mixtures in which organisms have similar codon usage patterns ($\Psi < 0.6$; Fig 3b,d,f).

Some Assembly is Required: gRodon's Performance on Synthetic Metagenomes

So far, the benchmarking analyses we have described do not account for complications that could arise from the sequencing or assembly process. In order to better simulate a realistic dataset we generated sequencing reads from genome mixtures [36], focusing on organisms found in the human gut [33], in order to generate 100 synthetic gut metagenomes. We then tested MMv2's performance on (1) the genome mixtures used to generate these synthetic metagenomes, (2) the set of genes called from contigs assembled from the synthetic metagenomes, and (3) genes called directly from reads from the synthetic metagenomes. For the third analysis, we used gRodon's "fragments" option, which filters out genes shorter than 40 codons long (gRodon typically filters genes shorter than 80 codons long based on guidance from the authors of the CUB statistic used by the program [19, 21, 37]). This short-filter option was necessary to accommodate the short length of the synthetic reads (125bp), and gives us an idea whether or not it is wise to skip the assembly process altogether for short-read datasets.

We found that MMv2 performed equally well on assembled metagenomes as the original genome mixtures, but that performance suffered greatly when MMv2 was applied to genes called directly from reads (Fig 4). We suspected that this drop in performance was due to the short length of the genes when called directly from reads, which makes it difficult to confidently estimate CUB. Looking at individual genomes, we found that when genes were concatenated to progressively shorter lengths the variance of their estimated CUB increased greatly, and for very short lengths (typically < 120bp) the CUB was underestimated (S13 Fig). This effect was stronger for ribosomal proteins in fast-growing organisms, including for genes longer than 120bp (S14 Fig). This suggests that minimum doubling times would be overestimated for genomes or metagenomes with many short genes (i.e., genes inferred from reads), which

is consistent with our results on synthetic gut metagenomes (Fig 4). We also observed that using genes called from reads from our synthetic metagenomes led to a direct reduction in the estimated CUB of the ribosomal proteins in a sample relative to estimates derived from genes called from assembled contigs (S15 Fig). This analysis suggests that assembly is a necessary step for growth prediction, and that short genes should be excluded when making such predictions (as done in gRodon's default settings). While growth prediction from short reads (< 240bp) is a desirable feature for future development, it is not recommended using the existing gRodon model.

Interestingly, the assembly process itself did not seem to affect gRodon's prediction performance, with MMv2 performing just as well on synthetic metagenomic assemblies from simulated reads as on mixtures of the original genes used to generate those reads. One might expect that when applied to assemblies gRodon's performance would suffer marginally due to poor assembly of rare community members, but because those community members contribute so little to the average maximum growth rate calculation (their contribution being inversely related to their relative abundance), the assembly process seems to have little effect. It is possible that in a hyper-diverse community where even abundant organisms assemble very poorly these results would be different. Nevertheless, by subsampling our genome mixtures, as well as deeply-sequenced metagenomes gathered from diverse habitats, we found that even if only a small fraction of the total genes in a community are sampled (around 5% of all genes, and capturing at least 30-50 unique ribosomal proteins) MMv2 will provide reliable community-level predictions (S16 and S17 Figs; also see unweighted predictions with no correction for relative abundance in S18 and S19 Figs). While the actual cutoff for accurate prediction will vary across communities (the 30-50 gene cutoff is apparently appropriate for those with a lognormal species abundance distribution, more may be required for more uniform abundance distributions), these simulations suggest that gRodon can be applied to relatively shallow metagenomes.

Growth Prediction on Large Metagenomic Datasets from the Global Oceans and the Human Gut

We ran MMv2 on two large metagenomic datasets to get a sense of how the model would perform on real data: (1) the BioGEOTRACES dataset which consists of 610 globally distributed marine metagenomes [26], and (2) the Human Microbiome Project (HMP) Illumina WGS Assemblies (HMASM) dataset which consists of 749 metagenomes sampled from a large number of individuals and body sites [8, 27]. These datasets describe dramatically different microbial communities living in dramatically different environments, with marine systems typically being dominated by slower-growing organisms living in oligotrophic conditions, and the human microbiome generally being composed of much faster growing organisms living in nutrient-rich conditions [9, 13, 21].

First, we found that our simulated genome mixtures used for benchmarking did a reasonably good job of recapitulating natural distributions of Ψ (i.e., codon-similarity patterns across organisms) from the same environments. That is, the distribution of Ψ seen in our "Marine Surface" mixtures overlapped well with the distribution of Ψ seen in the BioGEOTRACES metagenomes (though the distributions were still statistically distinguishable; Kolmogorov-Smirnov test, $p = 1.17 \times 10^{-7}$; Fig 5a). Similarly, our "Human Gut" mixtures had a Ψ distribution that overlapped with the stool samples in the HMASM dataset (statistically indistinguishable means; Kolmogorov-Smirnov test, $p = 0.056$; Fig 6a). In both cases, the Ψ distribution of environmental mixtures and actual metagenomes was shifted towards much lower values than seen in the "Refseq" mixtures. This implies that environment-specific genome mixtures do a much better job of replicating at least one feature of natural communities than do sets of genomes drawn

at random from genomic databases.

Second, we found that MMv2 allowed us to make specific hypotheses about variation in community-wide average maximum growth rates in marine and human systems. Among the BioGEOTRACES samples, we noted a distinct decrease in the community-wide maximum growth rate with depth after 100 meters, consistent with predictions from other groups and with decreasing energy inputs to ocean systems with decreasing light (Fig 5b; [9]). Among the HMASM samples, we found differences in the community-wide maximum growth rate across body sites, with the oral microbiome of the tongue and cheek having the fastest maximum growth and the gut and tooth plaque having the slowest (Fig 6b). These likely have to do with differences in the kinds of substrates available across environments, where more complex carbon sources are broken down in the gut relative to the mouth (e.g., [38]). Finally, we found clear differences in the average community-wide maximal growth rates across environments (mean BioGEOTRACES doubling time of 11.1 hours and mean HMASM doubling time of 1.6 hours; Wilcox rank sum test $p < 2.2 \times 10^{-16}$; Figs 5c and 6c), with 92% of HMASM samples having a faster community-wide maximal growth rate than the fastest seen in a BioGEOTRACES sample. Importantly, many of these ecological patterns may have been masked if we applied MMv1, which in general was strongly biased towards faster inferred maximum growth rates (Figs 5c and 6c).

Conclusions

We developed gRodon MMv2, an improved predictor of the community-wide average maximum growth rate for mixed-species metagenomes and implemented this predictor in an open-source R package. We provide extensive benchmarking to demonstrate that MMv2 outperforms previous growth models, and to provide guidance on when our predictor's performance will suffer. For example, we showed that weighting predictions by species abundances improved overall performance, and that assembly was a necessary step for reliable prediction on short-read datasets. The development of predictors that work well on un-assembled short-read data is an active area of development, and would greatly expand the utility of our tool. We emphasize that MMv2 shows improvement only for community-level prediction from mixed species communities, and that for individual genomes users should still use gRodon's "full" mode [21], which is the best method available for those data types. In the rare case where reference genomes are available for all species present in a metagenome, applying gRodon to the reference database and using coverage to estimate relative abundances in order to calculate the weighted average maximum growth rate, rather than applying metagenome mode directly, is likely to give the best results. Nevertheless, the ability of MMv2 to resolve cross-species differences, even across closely-related organisms (S20 Fig), appears to be robust.

By applying MMv2 to large-scale metagenomic datasets, we are able to derive insights about the variation of the community-wide average maximum growth rate across marine and and human-associated environments. For example, we saw clear decreases in the community-wide maximum average growth rate with depth, in agreement with other studies [9]. At the same time, we saw strong differentiation among body sites in their average maximal growth rates, potentially related to the availability of different growth substrates across sites. Together, these results demonstrate the broad utility of our tool and it's potential to enable rapid hypothesis generation from metagenomic datasets.

Finally, we note that our benchmarking approach of comparing single-genome inferences to community-level inferences is inherently limited in that the "ground truth" is an inference rather than a datapoint in itself (though see S6 Fig). Nevertheless, we

found that community-level predictions can clearly recapitulate individual-level predictions without having to extract individual genomes from the community. As more extensive curated trait datasets become available, alternative benchmarking approaches for mixed-species communities may become feasible, and there is a critical need for such databases.

Materials and Methods

Code for all analyses presented can be found at: `https://github.com/jlw-ecoevo/gRodon2-benchmarking`. The R package gRodon with MMv2 implemented can be found at: `https://github.com/jlw-ecoevo/gRodon2`.

Sequence handling in R was performed using the Biostrings package [39]. Data visualization was performed using ggplot2 and ggpubr [40, 41].

Data

Out of the set of RefSeq annotated assemblies [31, 31] incorporated in the original EGGO database [21] we randomly sampled one genome per genus (following the NCBI taxonomy [42]) to generate a set of non-redundant genomes spanning 2976 genera. These comprised our "RefSeq" genome set. Genomes of human gut isolates were obtained from the Zou et al. [33] set of 1520 sequenced cultured isolates. Marine SAGs were obtained from the Pachiadaki et al. [34] GORG-Tropics database of 12,715 ogranisms. All genomes were screened and genomes with < 10 annotated ribosomal proteins were removed.

Genomes for Mycobacteria and Vibrio species were taken from the set of all GTDB207 representative genomes assigned to these genera [43].

Metabolic oxygen use data for each species in the training dataset were obtained from Madin et al [44].

Raw reads,assemblies, and temperature metadata for the BioGEOTRACES dataset were obtained from Biller et al. [26]. We then ran EukRep v0.6.6 on these assemblies to classify contigs as eukaryotic or prokaryotic, and retained only prokaryotic contigs (using settings `-m strict --tie prok`; [45]). Raw reads and assemblies for the HSASM dataset [8] were obtained from `https://www.hmpdacc.org/hmp/HMASM/#data`, and sample temperature was assumed to be 37 degrees Celsius. Assemblies were annotated with prokka (using options `--norrna --notrna --metagenome --centre X --compliant`; [46]). For reads, adapters and low-quality reads were trimmed using fastp v0.21.0 [47], and then cleaned reads were mapped to inferred genes using bwa mem v0.7.12 (default settings ; [48]) and coverage was quantified using bamcov v0.1.1 (available at `https://github.com/fbreitwieser/bamcov`).

Genome Mixtures

To simulate a genome mixture we randomly sampled 10 genomes from the relevant set of genomes and concatenated annotated coding sequences from these genomes into a single fasta file for that mixture. This was repeated 10,000 times for each data source. Maximum growth rates were inferred for all source genomes using gRodon's "full" prediction mode to derive ground-truth against which to benchmark predictions on the genome mixtures.

To simulate relative abundances, the individual abundance of each genome in a mixture was drawn from a lognormal distribution. These relative abundances were then

used to weight the calculation of average community-wide maximum growth rate for the gRodon "full" mode benchmark and were additionally passed as gene relative abundances to gRodon when running MMv2 in abundance-corrected mode.

We generated optimal growth temperatures for genome mixtures using the following two-step procedure. First, for each genome mixture a sample temperature between 0 and 60°C was drawn from a uniform distribution. Then, the simulated optimal growth temperature of each species in the mixture was taken as the sum of this sample-wide value and a draw from a normal distribution with mean zero and standard deviation of 10. These organism-level temperatures were used to predict the individual growth rates of genomes (for the benchmark), and the sample temperature was used to predict the community level growth rate. This approach was used to account for possible variation in an organism's OGT relative to the conditions it may be found in.

Training data mixtures were generated from the genomes matching species in the Madin et al. [44] training set that were used to train gRodon [21].

Synthetic Metagenomes

Synthetic reads were generated from genome mixtures using inSilicoSeq [36]. We (using options `--n_reads 100M --model hiseq --coverage lognormal`) generated synthetic metagenomes with a Hi-Seq sequencing error model and 100×2 million 125bp paired reads. Adapters and low-quality reads were trimmed using fastp v0.21.0 and reads were then either assembled with MEGAHIT v1.2.9 (default parameters; [49]) and genes annotated with prokka (using options `--norrna --notrna --metagenome --centre X --compliant`; [46]) or genes were called directly from reads using FragGeneScanRs (using option `--training-file illumina_5`; [50, 51]). For assemblies, cleaned reads were mapped to inferred genes using bwa mem v0.7.12 (default settings ; [48]) and coverage was quantified using bamcov v0.1.1 (available at `https://github.com/fbreitwieser/bamcov`). We annotated genes called directly from reads as ribosomal proteins using blastn (Evalue cutoff of 10^{-5} and a 99% identity; [52]) and the ribosomal protein database from growthpred [13]. The sets of genes predicted directly from reads were very large (160 million genes on average), so that we subsampled 1% of genes per sample for gRodon prediction (increasing this to 10% did not change out result; S21 Fig).

Subsampled Synthetic Communities and Metagenomes

Subsampled community level data was generated via sampling genes from genome mixtures or metagenomes with probabilities proportional to their relative abundances – simulated relative abundances in the case of mixtures, coverage in the case of metagenomes – without replacement.

Bias-Corrected Growth Model

See Table S1 for a list of models used in the manuscript.

The original gRodon MMv1 model was fit as:

$$\Phi_\lambda(\text{Doubling Time}) \sim \overline{\text{CUB}}_{\text{HE}} \tag{1}$$

where Φ_λ is the Box-Cox transformation with parameter λ, fitted using the MASS package [53], and $\overline{\text{CUB}}_{\text{HE}}$ is the median CUB of the highly expressed genes (taken to be the set of ribosomal proteins) relative to the average codon usage patterns across the genome, calculated as the MILC statistic [19] using the coRdon package [37]. See Weissman et al. [21] for more discussion of the model fitting process.

The bias-corrected model was fit as:

$$\Phi_\lambda(\text{Doubling Time}) \sim \frac{\overline{\text{iCUB}}_{\text{All}} - \overline{\text{iCUB}}_{\text{HE}}}{\overline{\text{iCUB}}_{\text{All}}} + |0.5 - \text{GC}| \tag{2}$$

using a normalization approach implemented in the original growthpred software by Vieira-Silva and Rocha [13]. GC is the genome- or metagenome-wide GC content. Here $\overline{\text{iCUB}}$ refers to the average CUB calculated using a per-gene background to estimate the expected codon frequencies rather than the whole genomic background. That is to say, the MILC statistic of a gene is normally calculated by comparing observed codon frequencies to a null expectation estimated from genome-wide frequencies. Here, we calculate the expected codon frequency based on the gene itself, controlling for nucleotide frequencies. This is done by repeatedly shuffling the gene sequence to produce a set of simulated genes on which the expected frequencies are calculated (by default 100 such "null" genes are simulated for each sequence). Because this process is slow, we sample 100 genes from the genomic background (not in the highly expressed set) for which these per gene estimates are made to calculate $\overline{\text{iCUB}}_{\text{All}}$. This model was fit using the Madin et al. [44] training set used to train the original gRodon model.

Finally, MMv2 applies either the MMv1 or bias-corrected model based on a Ψ threshold of 0.6, with values lower than 0.6 leading to the application of the MMv1 model and values larger than that threshold leading to the application of the bias-corrected model. Ψ is calculated as the mean CUB of the ribosomal proteins using the set of ribosomal proteins to calculate the background expectation of codon frequencies. This essentially measures how distant the ribosomal proteins are from each other in codon-usage space, and thus is an indicator of whether or not organisms in a metagenomic sample share the same codon biases (see Weissman et al. [21], where Ψ is defined as "consistency").

For comparisons, growthpred v1.0.8 was run from a docker image from Long et al. [25] (`https://hub.docker.com/r/shengwei/growthpred`) in metagenome mode (-m) using the same set of ribosomal proteins as gRodon in all cases.

Acknowledgments

J.L.W. was supported by a postdoctoral fellowship in marine microbial ecology from the Simons Foundation (Award 653212). 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. Green JL, Bohannan BJ, Whitaker RJ. Microbial biogeography: from taxonomy to traits. science. 2008;320(5879):1039–1043.

2. Krause S, Le Roux X, Niklaus PA, Van Bodegom PM, Lennon JT, Bertilsson S, et al. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. Frontiers in microbiology. 2014;5:251.

3. Martiny JB, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: a phylogenetic perspective. Science. 2015;350(6261):aac9323.

4. Lajoie G, Kembel SW. Making the most of trait-based approaches for microbial ecology. Trends in Microbiology. 2019;27(10):814–823.

5. Pace NR. A molecular view of microbial diversity and the biosphere. Science. 1997;276(5313):734–740.

6. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457–463.

7. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. Nature. 2018;560(7717):233–237.

8. Consortium HMP, et al. Structure, function and diversity of the healthy human microbiome. nature. 2012;486(7402):207.

9. 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.

10. Ustick LJ, Larkin AA, Garcia CA, Garcia NS, Brock ML, Lee JA, et al. Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. Science. 2021;372(6539):287–291.

11. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nature Biotechnology. 2020;38(6):685–688.

12. Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome. 2022;10(1):1–22.

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

14. Roller BR, Stoddard SF, Schmidt TM. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nature microbiology. 2016;1(11):1–7.

15. 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.

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

17. 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.

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

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

20. 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.

21. 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).

22. Kirchman DL. Growth rates of microbes in the oceans. Annual review of marine science. 2016;8:285–309.

23. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010;330(6007):1099–1102.

24. Weissman JL, Dimbo ERO, Krinos AI, Neely C, Yagues Y, Nolin D, et al. Estimating the maximal growth rates of eukaryotic microbes from cultures and metagenomes via codon usage patterns. bioRxiv. 2021;.

25. 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.

26. 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.

27. Gevers D, Pop M, Schloss PD, Huttenhower C. Bioinformatics for the human microbiome project. PLoS Computational Biology. 2012;8(11):e1002779.

28. Roller M, Lucić V, Nagy I, Perica T, Vlahoviček K. Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic acids research. 2013;41(19):8842–8852.

29. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–929.

30. Felsenstein J. Phylogenies and the comparative method. The American Naturalist. 1985;125(1):1–15.

31. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic acids research. 2014;42(D1):D553–D559.

32. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic acids research. 2016;44(14):6614–6624.

33. Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nature biotechnology. 2019;37(2):179–185.

34. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, et al. Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. Cell. 2019;179(7):1623–1635.

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

36. Gourlé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics. 2019;35(3):521–522.

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

38. Weissman JL, Dogra S, Javadi K, Bolten S, Flint R, Davati C, et al. Exploring the functional composition of the human microbiome using a hand-curated microbial trait database. BMC bioinformatics. 2021;22(1):1–21.

39. Pages H, Aboyoun P, Gentleman R, DebRoy S, Pages MH, IRanges L, et al. Package 'Biostrings'. R. 2013;.

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

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

42. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database. 2020;2020.

43. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic acids research. 2022;50(D1):D785–D794.

44. Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, et al. A synthesis of bacterial and archaeal phenotypic trait data. Scientific Data. 2020;7(1):1–8.

45. 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.

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

47. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–i890.

48. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013;.

49. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674–1676.

50. Van der Jeugt F, Dawyndt P, Mesuere B. FragGeneScanRs: better and faster gene prediction for short reads. bioRxiv. 2021;.

51. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic acids research. 2010;38(20):e191–e191.

52. 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.

53. 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.

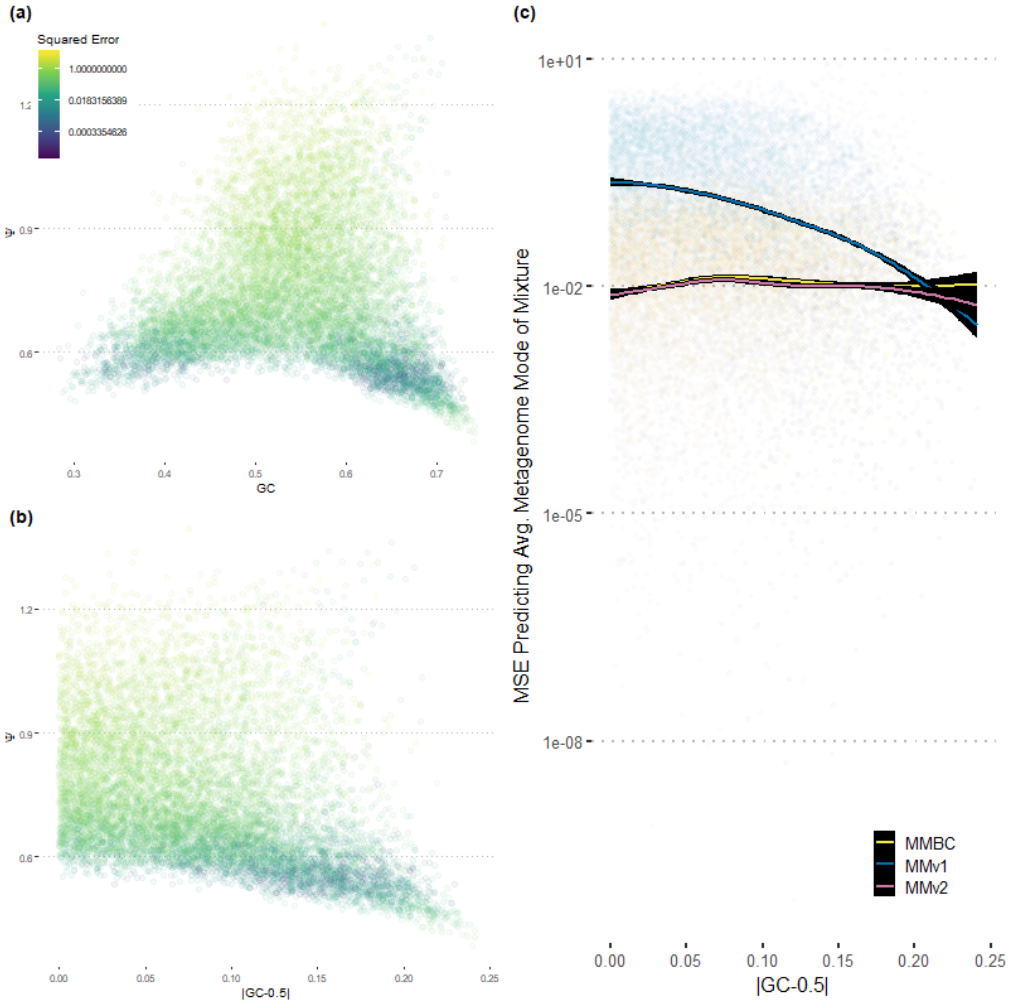

Fig 1. Samples with highly skewed GC content have low Ψ and low prediction error with MMv1. (a-b) More extreme GC values (far from 0.5) correspond to low Ψ values (< 0.6).(c) This panel shows how well each predictor converges to "itself" for various levels of GC-skew for genome pairs. That is, if we predict the max. growth rate of each genome in a mixture using that model and take the average, how well does the prediction on the mixed community estimate that value? Observe that MMv1 only performs comparably to (or better than) MMBC and MMv2 when GC content is highly skewed, corresponding to a low Ψ.

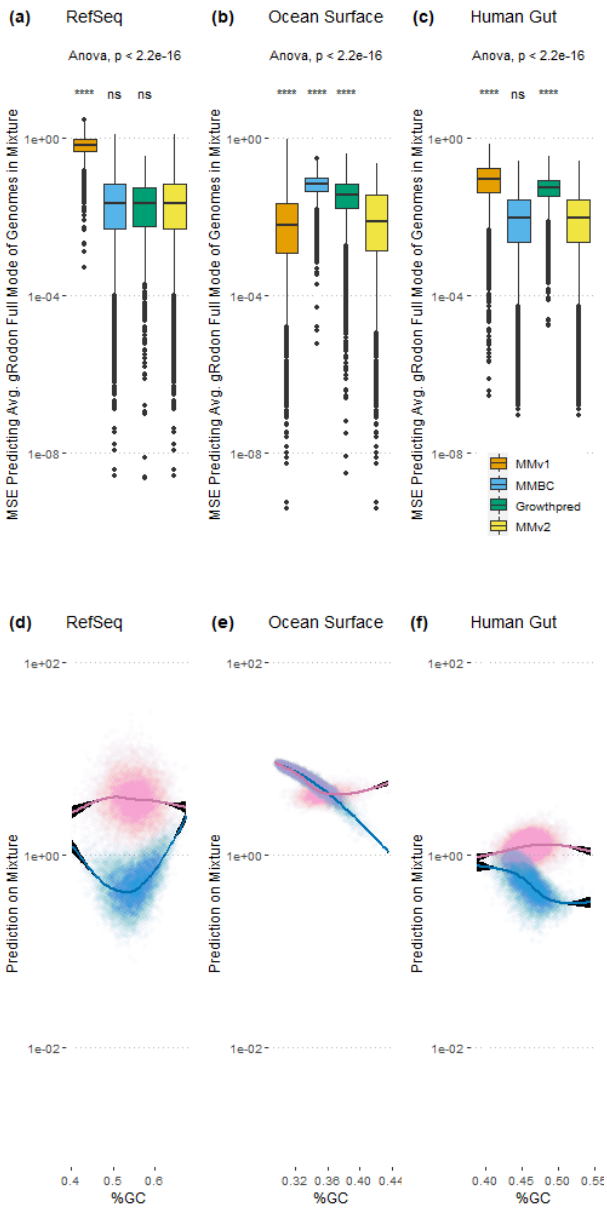

Fig 2. Metagenome mode v2 (MMv2) outperforms MMv1 and corrects its apparent GC bias. (a-c) For each set of source genomes we calculated the MSE against the prediction for individual genomes using gRodon's "full mode", which gives us the best available baseline for "actual" maximal growth rates for each mixture [21]. ANOVA for each sample source shown, with pairwise significance values relative to MMv2 obtained from a Wilcox signed rank test. (d-f) Predicted doubling times from MMv2 are generally insensitive to the GC content of a sample (see main text for discussion of ocean surface mixtures), whereas MMv1's predictions are strongly affected by GC content.

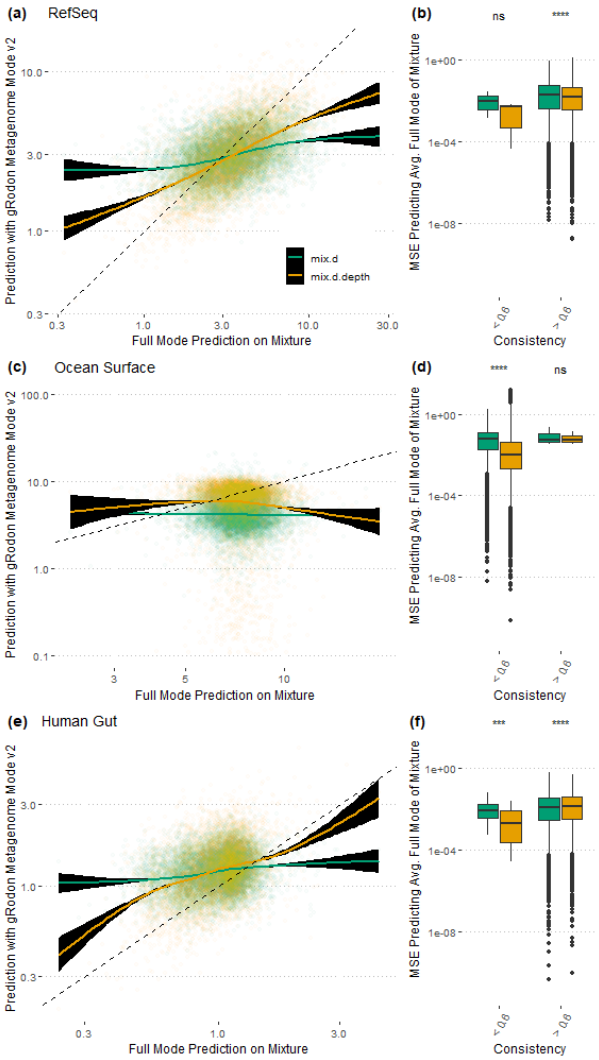

Fig 3. Abundance-weighted growth predictions with MMv2 are more accurate than predictions that do not account for abundance. (a,c,e) Doubling time predictions that are weighted for the relative abundance of an organism are more accurate than unweighted predictions. (b,d,f) The advantages of abundance-weighting are seen primarily in mixtures in which organisms have similar codon usage patterns ($\Psi < 0.6$). Pairwise significance values from a Wilcox signed rank test.

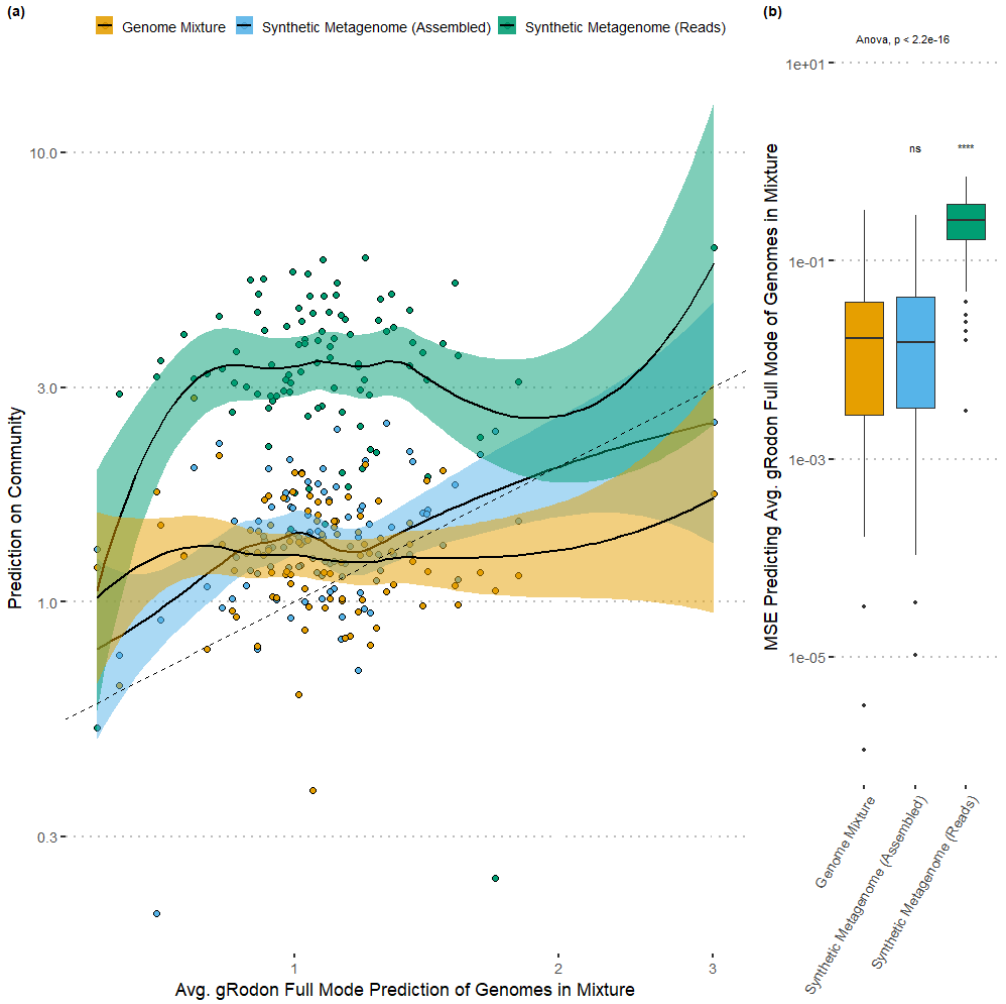

Fig 4. For short read datasets, assembly is required for confident maximum growth rate predictions. (a-b) Doubling time predictions on genome mixtures, synthetic metagenomic assemblies from simulated reads, and directly on the simulated reads of synthetic metagenomes for the same sets of organisms were compared. Predictions using genes inferred directly from reads had much higher error rates than predictions on the assembled contigs or genome mixtures, which were not statistically distinguishable. ANOVA across sample types shown, with pairwise significance values relative to genome mixture obtained from a Wilcox signed rank test.

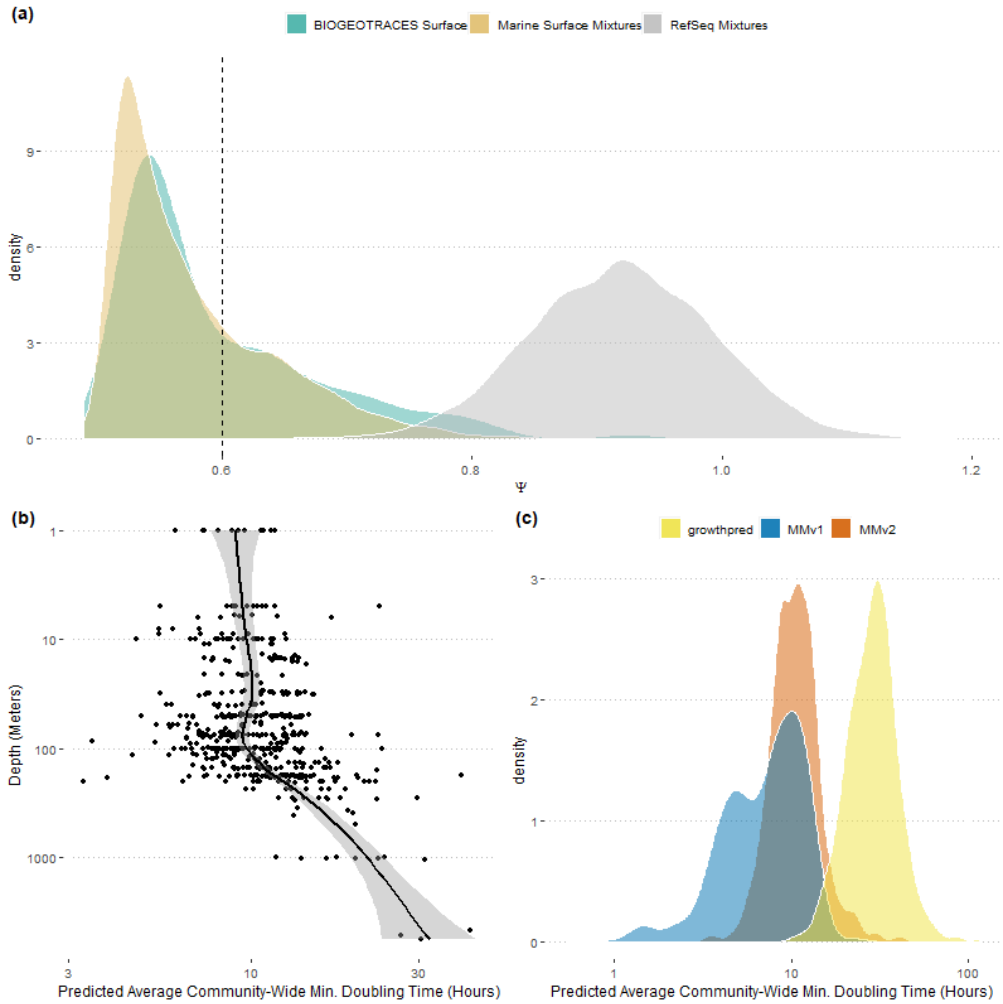

Fig 5. Natural marine communities as a test-case for MMv2. (a) The distribution of Ψ across organisms among BioGEOTRACES samples taken from a depth of < 100 meters largely overlaps with the Ψ distribution from our "Ocean Surface" mixtures, but not our "Refseq" mixtures. Ψ is a measure of codon-usage dissimilarity across highly-expressed genes in a sample, and low values of Ψ indicate high similarity (see Results and Discussion). (b) MMv2 applied to the BioGEOTRACES samples recovers a pattern of increasing average minimum doubling time of marine communities with depth after 100 meters. (c) Relative to MMv2, MMv1 predictions are strongly biased towards shorter doubling times for BioGEOTRACES samples.

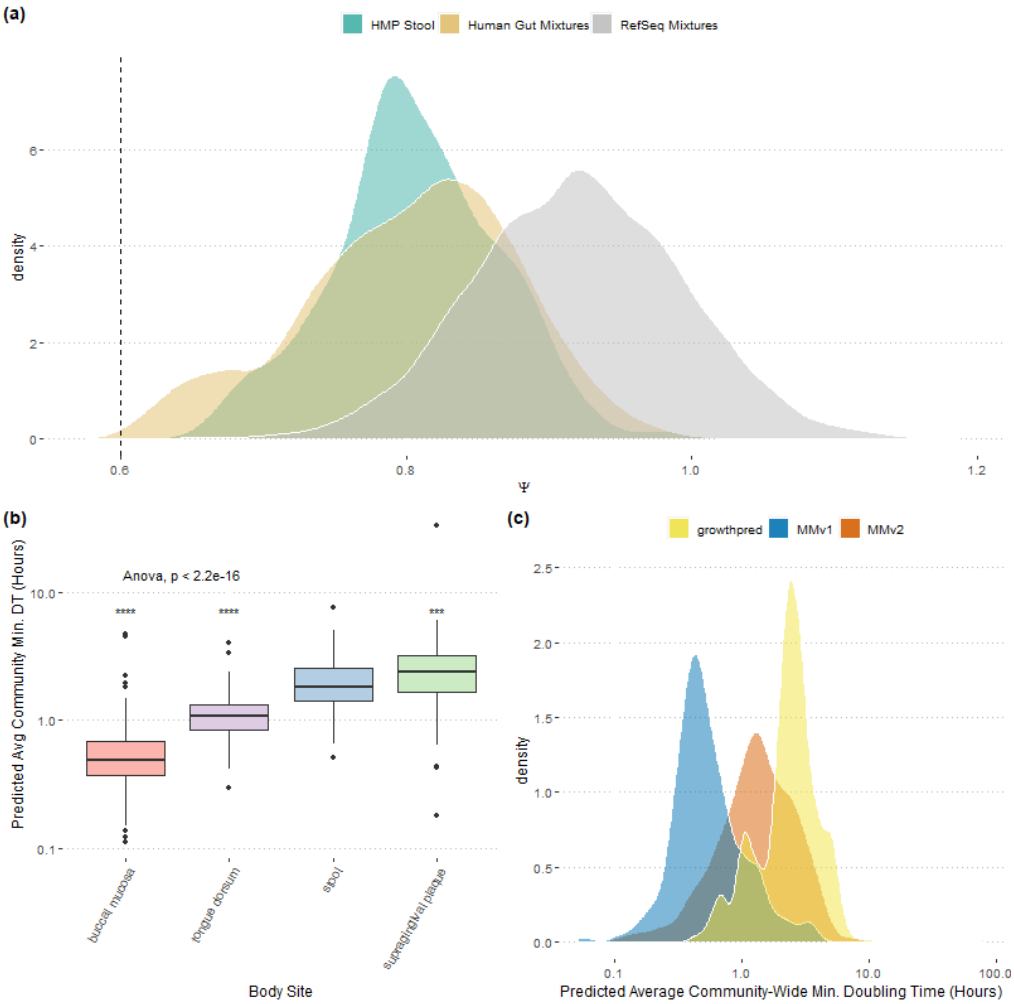

Fig 6. Natural human-associated communities as a test-case for MMv2. (a) The distribution of Ψ across HMASM samples taken from a human stool largely overlaps with the Ψ distribution from our "Human Gut" mixtures, but not our "Refseq" mixtures. Ψ is a measure of codon-usage dissimilarity across highly-expressed genes in a sample, and low values of Ψ indicate high similarity (see Results and Discussion). (b) MMv2 applied to the HMASM samples recovers differences in growth rates across human body sites (sites with > 100 samples). ANOVA across sites shown, with pairwise significance values relative to stool obtained from a Wilcox signed rank test. (c) Relative to MMv2, MMv1 predictions are strongly biased towards shorter doubling times for HMASM samples.