2019

Theoretical and Simulation-Based Investigation of the Relationship between Sequencing Effort, Microbial Community Richness, and Diversity in Binning Metagenome-Assembled Genomes

Royalty Taylor M
University of Tennessee, Knoxville

Andrew D. Steen
University of Tennessee, Knoxville

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Novel Systems Biology Techniques

Theoretical and Simulation-Based Investigation of the Relationship between Sequencing Effort, Microbial Community Richness, and Diversity in Binning Metagenome-Assembled Genomes

Taylor M. Royalty,[a] Andrew D. Steen[a,b]

[a]Department of Earth and Planetary Sciences, University of Tennessee, Knoxville, Tennessee, USA
[b]Department of Microbiology, University of Tennessee, Knoxville, Tennessee, USA

ABSTRACT We applied theoretical and simulation-based approaches to characterize how microbial community structure influences the amount of sequencing effort to reconstruct metagenomes that are assembled from short-read sequences. First, a coupon collector equation was proposed as an analytical model for predicting sequencing effort as a function of microbial community structure. Characterization was performed by varying community structure properties such as richness, evenness, and genome size. Simulations demonstrated that while community richness and evenness influenced the sequencing effort required to sequence a community metagenome to exhaustion, the effort necessary to sequence an individual genome to a target fraction of exhaustion depended only on the relative abundance of the genome and its genome size. A second analysis evaluated the quantity, completion, and contamination of metagenome-assembled genomes (MAGs) as a function of sequencing effort on four preexisting sequence read data sets from different environments. These data sets were subsampled to various degrees of completeness to simulate the effect of sequencing effort on MAG retrieval. Modeling suggested that sequencing efforts beyond what is typical in published experiments (1 to 10 Gbp) would generate diminishing returns in terms of MAG binning. A software tool, Genome Relative Abundance to Sequencing Effort (GRASE), was created to assist investigators to further explore this relationship. Reevaluation of the relationship between sequencing effort and binning success in the context of genome relative abundance, as opposed to base pairs, provides a constraint on sequencing experiments based on the relative abundance of microbes in an environment rather than arbitrary levels of sequencing effort.

IMPORTANCE Short-read sequencing with Illumina sequencing technology provides an accurate, high-throughput method for characterizing the metabolic potential of microbial communities. Short-read sequences can be assembled and binned into metagenome-assembled genomes, thus shedding light on the function of microbial ecosystems that are important for health, agriculture, and Earth system processes. The work presented here provides an analytical framework for selecting sequencing effort as a function of genome relative abundance. As such, experimental goals in metagenome-assembled genome creation projects can select sequencing effort based on the rarest target genome as a constrained threshold. We hope that the results presented here, as well as GRASE, will be valuable to researchers planning sequencing experiments.

KEYWORDS DNA sequencing, MAG, ecology, mathematical modeling, metagenomics, microbial communities

Citation Royalty TM, Steen AD. 2019. Theoretical and simulation-based investigation of the relationship between sequencing effort, microbial community richness, and diversity in binning metagenome-assembled genomes. mSystems 4:e00384-19. https://doi.org/10.1128/mSystems.00384-19.

Editor Janet K. Jansson, Pacific Northwest National Laboratory

Address correspondence to Andrew D. Steen, asteen1@utk.edu.

This article is contribution 491 from C-DEBI.

Received 25 June 2019
Accepted 26 August 2019
Published 17 September 2019

The reconstruction of high-accuracy short-read sequences into metagenome-assembled genomes (MAGs) is a powerful approach to characterize microbial metabolisms within complex communities (1). The recent creation of ~8,000 MAGs from largely uncultured organisms across the tree of life (2), the spatial characterization of microbial metabolisms and ecology across Earth's oceans (3), and the characterization of the potential impact that fermentation-based microbial metabolisms have on biogeochemical cycling in subsurface sediment environments (4) provide a few examples of how MAGs have helped to elucidate the relationships between microbial ecology, microbial metabolisms, and biogeochemistry.

Sampling environmental microbial DNA involves selecting a target environment, sequencing effort, bioinformatic pipeline software and parameters, metabolism characterization software (i.e., for gene identification and similarity searches), and databases (Fig. 1). At present, there is little information to guide how much sequencing is appropriate to achieve scientific goals in such experiments (5). This gap in knowledge is partly attributed to the unknown structure of target microbial communities. A further challenge is that the accuracy and efficiency of bioinformatic pipelines are often difficult to characterize, and thus obscure the relationship between sequencing effort and MAG retrieval. Recent estimates compiled by Quince et al. (5) suggest that typical metagenomic shotgun sequencing experiments usually sequence between 1 Gbp and 10 Gbp. Researchers require more precise guidance to select an appropriate shotgun sequencing effort in order to maximize information and minimize cost for their specific experimental question.

Illumina sequencing technology is currently the most popular platform to generate metagenomic shotgun sequences (5). Previous investigators established theoretical relationships between contig formation rate (6) and single genome coverage (7) as a function of short-read sequencing effort. On the community level, heuristic approaches have been proposed for evaluating community-level coverage in respect to increases in sequencing effort. For example, it has been proposed to utilize short-read redundancies as a function of sequencing effort to estimate community-level coverage (8). Without *a priori* knowledge of the microbial community structure, practical application of these methods to estimate MAG retrieval as a function of sequencing effort is hindered.

Here, we present two distinct analyses which constrain the relationship between the quantity of Illumina metagenomic shotgun sequences and the community-level sequence coverage. First, we applied a theoretical model and numerical simulations to estimate the minimum sequencing effort needed to sequence a metagenome to a target fraction of exhaustion. Our theoretical model is unique compared to previous models (6, 7) in that we characterize sequencing effort in the context of community structure and consider all sequenceable combinations of *k*-mers in a metagenome. Second, we performed *in silico* experiments to simulate the effect of sequencing effort on retrieved MAGs for Illumina sequence data sets. Coupling results from the two analyses provides a framework for investigators to define sequencing experiments in the context of selecting a rarity and fraction of exhaustion for a desired target genome when sequencing a community. The patterns presented here can be used to guide sequencing effort decisions in future sequencing projects when MAG reconstruction is a primary goal.

RESULTS

Theoretical and numerical sequencing effort simulations. Using equation 6, we calculated the number of sequence reads required to sequence four hypothetical microbial communities to exhaustion: a perfectly even, moderately uneven, highly uneven, and a lognormally distributed structured community (Fig. 2a to d). Here, we define sequencing a community to exhaustion as sequencing all possible combinations of DNA *k*-mers within a genome. Note that, to simplify the model, this model treats identical *k*-mers (i.e., same DNA sequence) at different locations in the genome as mathematically different *k*-mers. The expected number of sequence reads to fully

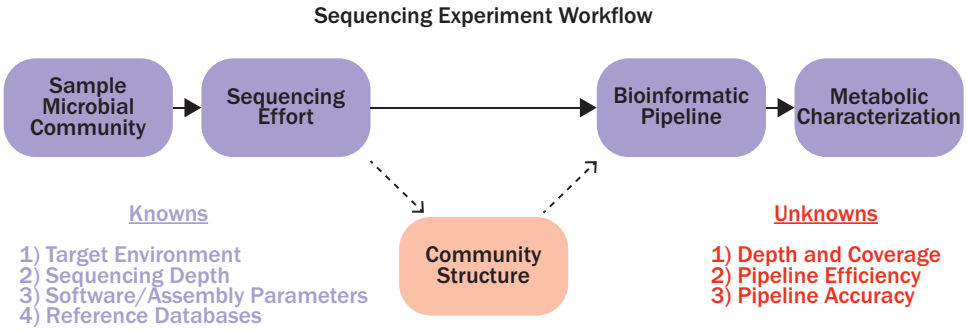

FIG 1 A flow diagram illustrating the workflow for sequencing experiments.

sequence the hypothetical communities was linear after log-transforming both expected sequences and metagenome size (unit of unique *k*-mers), suggesting a power-law relationship between metagenome size and the number of sequence reads required to sequence the metagenome to exhaustion (Fig. 2e). For all community structures, the slope of the relationship between log-transformed sequencing effort (sequenced reads) and metagenome size (unique *k*-mers) was within 1% of 1.06. The structure of the population strongly influenced the number of reads required such that more-even community structures required far fewer reads than less-even structures.

A limitation from using equation 6 for modeling sequencing effort is that it estimates only the sequencing effort for sequencing a metagenome to exhaustion. We circumnavigated this limitation by applying a numerical simulation to estimate the sequencing effort necessary to sequence a metagenome to a fraction of exhaustion for the same community structures analyzed earlier. The numerical simulation was validated by comparing the sequencing effort predicted from the numerical simulation and those from equation 6 when sequencing a metagenome to exhaustion (Fig. 3).

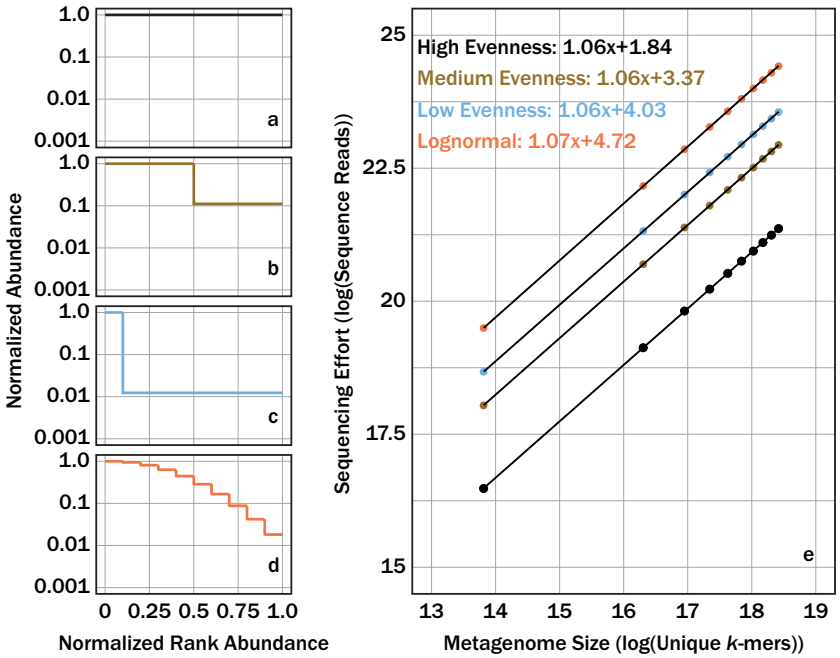

FIG 2 Sequencing effort, with units of log of sequence reads, required to fully sequence four different community structures, one with relatively high community evenness (a), relatively moderate community evenness (b), relatively low community evenness (c), and one with a lognormal community structure (d), were predicted using linear regressions (e) and the log of metagenome size (total possible *k*-mers) as a predictor.

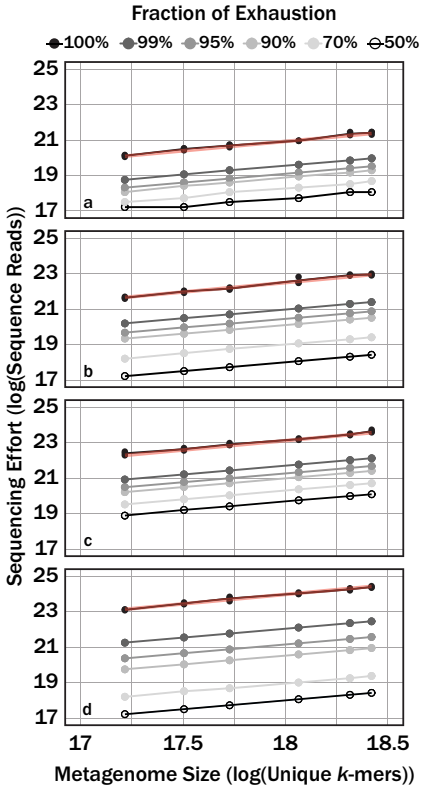

FIG 3 Sequencing effort, with units of log sequence reads, necessary to reach variable target sequencing depths (colors) for four different community structures, one with relatively high community evenness (a), relatively moderate community evenness (b), relatively low community evenness (c), and one with a lognormal community structure (d). Red translucent lines correspond with linear regression curves for the respective community in Fig. 2e.

Again, the relationship between metagenome size and sequencing effort fit well to a power law. This observation was independent of the selected target fraction of exhaustion. Nonetheless, an increase in target fraction of exhaustion resulted in uniform increases in sequencing effort (log units) when the community structure was fixed; however, the rate of this increase varied with community structure evenness.

We were interested in quantitatively relating community evenness to sequencing effort. These communities ranged from perfectly even ($a = 0$, equation 8) to more uneven ($a = 0.02$, Fig. 4a). Evenness was quantified using the Pielou evenness index, which expresses Shannon diversity relative to the diversity of a perfectly even community (9). The sequencing effort required to characterize genomes depended strongly on both the evenness and the target fraction of exhaustion (Fig. 4b). Again, less-even communities required more sequence reads than more-even communities did. The strength of this relationship also depended on the target fraction of completion. A community with a Pielou evenness of 0.97 required 3 orders of magnitude more sequencing effort to sequence a metagenome to a target fraction of exhaustion than a perfectly even community, while the same community required only about 42% more reads to sequence 50% of the metagenome.

The inherent limitation to the theoretical and numerical constraints presented above is that community structure is not known *a priori*. Nonetheless, a simple line of rationalization can be applied to circumnavigate this issue for practical applications of our model. Equation 6 constrains expected sequencing effort based on the proportion of the population under consideration. That is, we can limit the model to consider only DNA *k*-mers associated with a set of genomes that represent a certain fraction of the community. Here, we limit the population such that our model predicts the expected sequencing effort of the rarest genome that we wish to sequence. Inherently, any

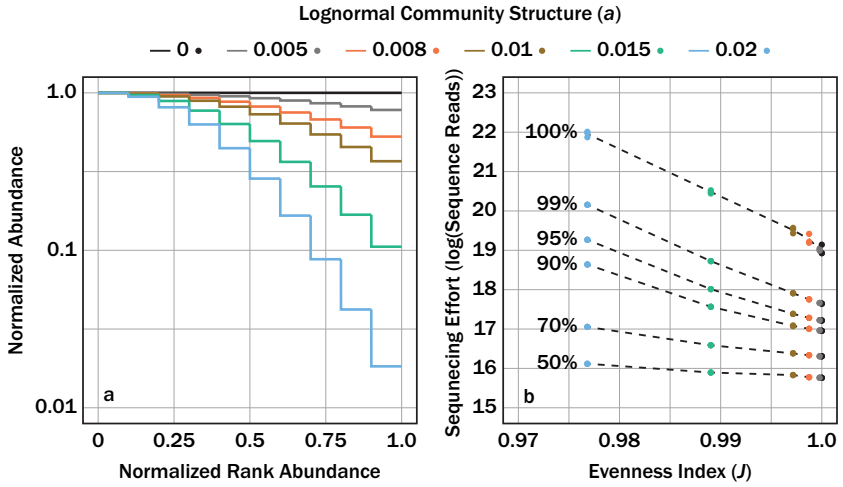

FIG 4 Numerical sequencing simulations applied to six hypothetical communities with different lognormal distributions that were defined by the parameter *a* from equation 7 (a). The sequencing effort, with units of log of sequence reads, necessary to sequence a target fraction of exhaustion (dashed contours) as a function of the Pielou evenness index *J* for a given lognormal community structure (b).

member more abundant than the selected rarest genome will also be sequenced to the minimum coverage and depth of the selected rarest genome. To determine the rareness of a given genome within a metagenome, the total number of unique *k*-mers within a genome (equation 2) is scaled by the true abundance of the microbe and divided by the size of the metagenome (equation 7). Thus, the proper application of this rationale requires a desired target fraction of exhaustion, an assumed genome size, and the relative abundance of the rarest genome to sequence. Numerical simulations, like those described earlier, were performed to determine the sequencing effort necessary to achieve a target fraction of exhaustion. The difference here is that these simulations analyzed the sequencing effort necessary to sequence a genome of a certain rareness to a target fraction of exhaustion, whereas the simulations above analyze the effort necessary to sequence the entire community. A generalized additive model (GAM) was built from simulation outputs to extrapolate sequencing effort required for relative abundances of less than 1% as simulations with lower abundances became computationally too intensive (GAM) (Fig. 5). The GAM shows expected sequencing effort required for microbial genome sizes of 0.5, 2, 5, 10, and 20 Mbp, target genome completeness fractions from 0.5 to 1.0, and genome relative abundances from 1 to 0.0001. The smooth dimensions for target fraction, genome size, and fraction of the metagenome community were 50, 6, and 29, respectively. To normalize for different sequence read length, sequence reads were converted to bases and ranged from 1×10^7 to 1×10^{15} total bases. More bases were required to sequence microorganisms (i) when the genome was relatively rarer in the community, (ii) to achieve better coverage of the genome, and (iii) when average genome sizes were larger.

Rarefying MAG binning as a function of sequencing effort. We rarefied four sequence read data sets to nine different fractions of the complete sequence data sets in triplicate. The rarefied data sets were then assembled and binned for a total of 108 analyzed metagenomes. Raw MAG data are provided in Data Set S1 in the supplemental material. The sum of medium- and high-quality MAGs as a function of high-quality bases empirically fit the Gompertz equation (equation 12; Fig. 6b to e; parameters in Table 1; Data Set S1). The sum of medium- and high-quality MAGs (henceforth referred to as quality MAGs for brevity) reduces sensitivity to whether bioinformatic pipelines tend to lump contigs into fewer, more-complete MAGs versus split them into more, less-complete MAGs. This was important for our analysis due to the large number of metagenomes ($n = 108$) which required assembly and an unsupervised binning algo-

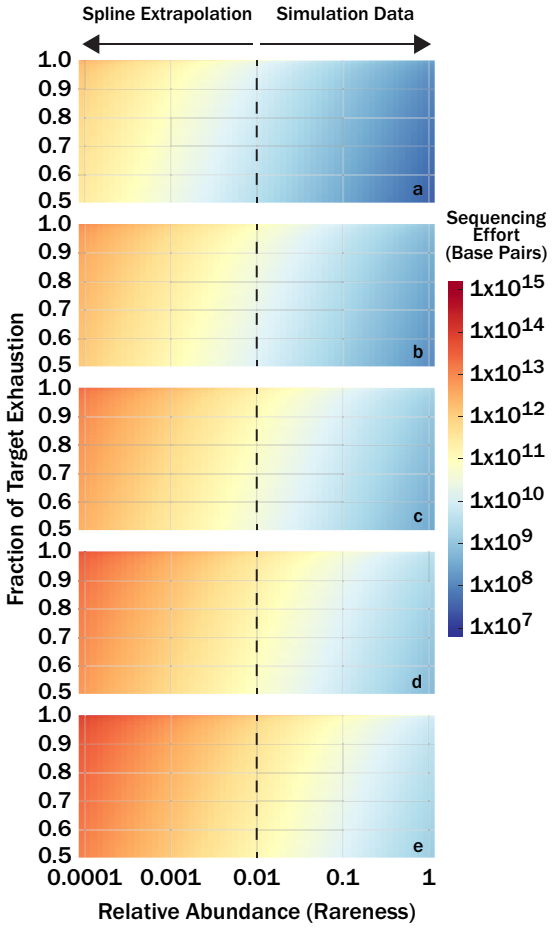

FIG 5 Numerical sequencing simulations show the number of bases (color bar) required to sequence a target fraction of a genome as a function of that genome's relative abundance in the community metagenome. Genomes evaluated were 0.5×10^6 (a), 2×10^6 (b), 5×10^6 (c), 10×10^6 (d), and 20×10^6 (e) base pairs long.

rithm. The large number of metagenomes made manual bin curation impractical during rarefication simulations.

For each environment, the quality MAGs as a function of simulated sequencing effort followed a sigmoidal relationship. In order to make the parameters of the fit intuitive to interpret, we fit the data to the Gompertz equation as rewritten by Zwietering et al. (10) (equation 12). Here, A, μ, and λ correspond to the maximum quality MAGs assembled with the pipeline, the maximum rate which the quality MAGs form with more sequencing, and the "lag bases," or the bases which must be sequenced before a sufficient number of sequence reads exist to generate overlap and form contigs (6). The predicted maximum quality MAGs varied from ~9.6 in the estuary sediment community to ~19 in the maize soil community. The predicted maximum rate that the quality MAGs increased varied from ~0.9 to ~3.8 MAGs Gbp^{-1}. Last, the minimum threshold of sequencing necessary prior to seeing quality MAGs varied from ~0.6 to ~6.1 Gbp. The *Tara Oceans* data set, where the quality decreased at a sequencing effort of >20 Gbp, was an exception. For the estuary, maize, and human gut data sets, the quality MAGs yield began to asymptote with increasing sequencing efforts. The *Tara Oceans* data set followed a similar pattern at <25 Gbp. However, when the number of sequenced bases was >25 Gbp, the quality MAGs decreased and became insensitive to sequencing effort.

Constraining MAG rarefaction analyses to community structure. Using the relationship shown in Fig. 5, we can convert sequencing effort to an abundance

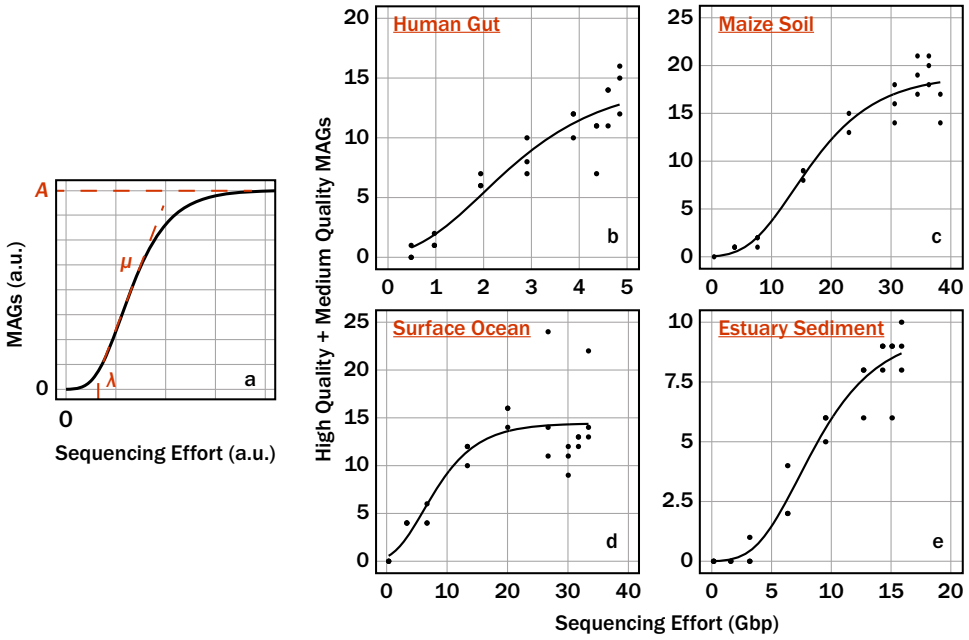

FIG 6 (a) The influence that the parameters *A*, μ, and λ had on the Gompertz equation. The sequencing effort and MAGs are shown in arbitrary units (a.u.). The property of the Gompertz equation that each parameter influences is colored red. The sum of medium-quality and high-quality MAGs (quality MAGs) as a function of sequencing effort for the gut (b), soil (c), surface ocean (d), and sediment (e) communities. The solid lines in panels b to e correspond to nonlinear least-squares fits of the Gompertz equation to the respective environmental data set. Note the different scales for the axes in panels b to e.

(rareness) cutoff if we assume a genome size. Genome sizes for genomes in RefSeq v92 (11) have 25th, 50th, and 75th quantiles of 2.73 Mbp, 4.30 Mbp, and 5.14 Mbp, respectively, and provide reasonable constraints for assumptions of genome size. Figure 7 shows quality MAGs retrieved for a genome of a given level of abundance that is sequenced to a target fraction of completeness for the human gut, maize soil, estuarine sediment, and surface ocean microbiomes analyzed earlier. Correlation coefficients for the regressions used to relate log-transformed sequencing effort (in base pairs) to genome relative abundance were $R = 1$ for all three genome sizes evaluated (1 Mbp, 5 Mbp, and 20 Mbp). Evaluation of 1 Mbp and 20 Mbp define the range of uncertainty in predicting quality MAGs as the true size of genomes is unknown. Unlike the asymptotic response of quality MAGs to sequencing effort (Fig. 6b to e), quality MAGs increase exponentially as the abundance cutoff decreases (note that the abundance cutoff is on a log scale). The target genome completeness fraction was held at a constant of 0.5 for all regressions, and the sensitivity of quality MAGs will change with respect to the genome abundance cutoff with different values of target genome completeness fraction.

TABLE 1 Estimates of fit coefficients for the Gompertz equation (equation 12) for the sum of high-quality and medium-quality MAGs as a function of sequencing depth in published data sets from ocean surface water, estuarine sediment, maize soil, and the human gut[a]

Environment	*A* (SE)	μ (SE)	λ (SE)	Yield
Ocean surface water	14.4 (1.0)	1.1 (0.4)	1.2 (2.1)	1.00
Estuary sediment	9.6 (0.9)	1.0 (0.2)	3.7 (0.7)	0.91
Maize soil	19.0 (1.1)	0.9 (0.1)	6.1 (1.5)	0.96
Human gut	14.6 (2.2)	3.8 (0.7)	0.6 (0.3)	0.88

[a]*P* values for all coefficients were $<<0.05$ except for gut and ocean surface water λ values. Yield is defined in equation 13 as the number of medium- and high-quality MAGs relative to the number predicted at infinite sequencing depth.

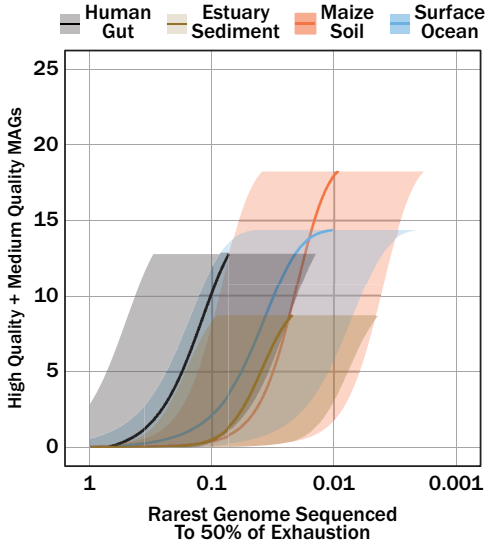

FIG 7 Quality MAGs (medium and high quality) as a function of the rarest genome sequenced to 50% exhaustion for human gut, maize soil, estuarian sediment, and surface ocean sequence data sets. Sequencing effort was converted to genome relative abundance sequenced to target fraction using the GAM presented in Fig. 5 The translucent shaded areas correspond to uncertainty from the target genome size (1 Mbp, or the lower bound, to 20 Mbp, or the upper bound), while the solid lines correspond to genome sizes of 5 Mbp.

DISCUSSION

We sought to establish evidence-based guidelines for selecting a sequencing effort during shotgun metagenomic sequencing experiments. The model proposed here (equation 6) addresses this goal. Our model establishes an intrinsic relationship between a community structure and the sequencing effort necessary to sequence members with different rareness, or relative abundance in a community (equation 1; Fig. 8). It is important to emphasize that the proposed model treats individual *k*-mers within a genome as members of the total population of DNA in an environment (equation 4; Fig. 8). Thus, the relative abundance of any given *k*-mer in a population is equal to the abundance of the host genome divided by the total number of *k*-mers in the entire metagenome (equation 7). Summing the probabilities of sequencing all individual *k*-mers from the same genome (equation 2; Fig. 8), should equal the relative abundance of the genome within the population of genomes. The theoretical model utilizes these individual probabilities of sampling *k*-mers to determine how much sequencing effort is required to sequence all possible *k*-mers in a community (Fig. 2 and 3) and for a member of some rareness in a community (Fig. 5). Practical sequencing challenges associated with strain-level microdiversity, extracellular DNA, and sequencer contamination are not problematic for the proposed model. However, these issues can still lead to problems during assembly. In particular, abundant lineages with a large amount of strain-level microdiversity can be entirely missing from the assembled data set despite a high coverage. Strains could be treated as independent taxonomic units, and ultimately, sequencing effort should be selected based on a target rareness of DNA being sequenced in the sequencer. Proper measures such as removing extracellular DNA (e.g., reference 12) and properly removing contaminated DNA are essential to avoid skewing the sequenced DNA population. Last, in practice, homologous DNA regions across genomes (e.g., shared genes) will be sequenced faster than unique regions. For simplicity of our model, we cannot predict the degree of homologous DNA and simply treat unique loci at DNA as a unique *k*-mer. Nonetheless, if this information is known *a priori*, the proposed model can account for homologous DNA. Equation 6 calculates sequencing effort for individual *k*-mer regions based on their relative abundance (p_j) with respect to all *k*-mers in the community. Theoretically, the relative abundances for

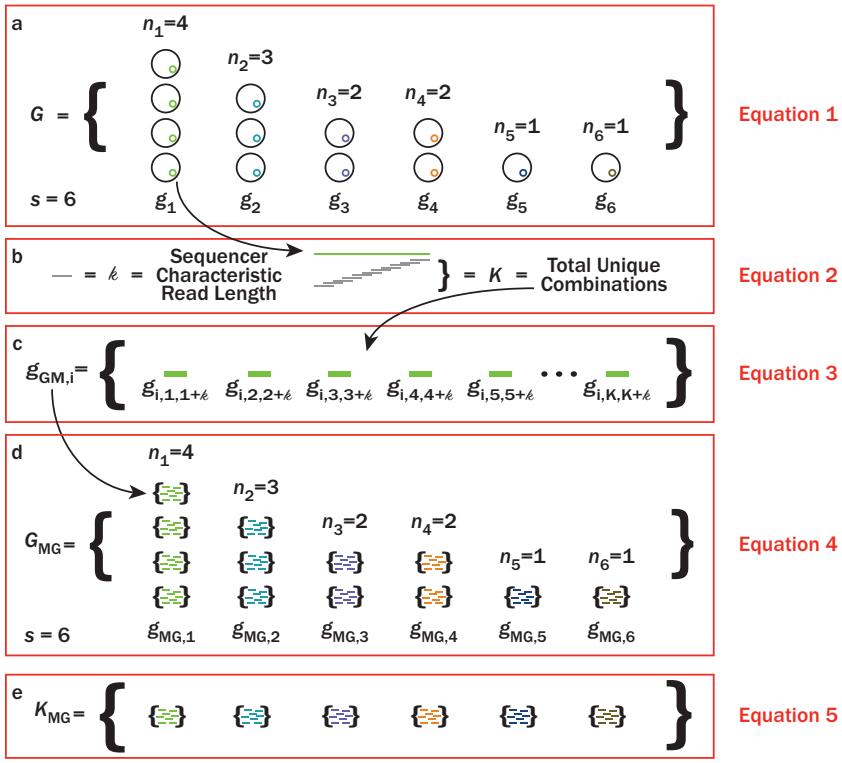

FIG 8 Cartoons illustrating an example microbial community (G), the metagenomes for genomes ($g_{MG,i}$) (as defined in equation 2) within G, and the overall metagenome for the given microbial community (G_{MG}). In this example, there are six genomes ($s = 6$) and a total of 13 individual microbes. (a) Black circles represent individual microbes whose genomes are averaged together, g. The average genome values, g, are indicated by different color inner circles. (b) Individual average genomes can be sequenced at K unique positions depending on the characteristic read length, k, of a sequencer. (c) All unique positions that can be sequenced for a given genome, g, defines the metagenome, g_{MG}, for the ith genome, g_i. (d) Replacing all individual genomes in panel a with metagenomes, g_{MG}, gives the metagenome of the microbial community, G_{MG}.

k-mers in homologous regions could be treated as an independent "reservoir," or genome ($g_{GM,i}$; equation 3), such that the relative abundance of the jth k-mer in this reservoir equals the sum of relative abundances from all host genomes that the homologous DNA actually exists in.

The theoretical probability model (equation 6) demonstrated that the sequencing effort to sequence a metagenome to exhaustion was predictable, regardless of community structure (Fig. 2). Furthermore, our characterization of community complexity demonstrates that less-complex communities require less sequencing to sequence all available DNA (Fig. 4). A limitation to equation 6 is that it predicts only the sequencing effort to sequence a metagenome to completion. In practice, assemblers do not require every unique k-mer to generate an accurate contig. As long as sufficient overlap exists between sequenced k-mers, assemblers can generate contigs. We therefore used a numerical simulation to predict the sequencing effort necessary to sequence an individual metagenome to a target fraction of exhaustion. The numerical simulation agreed with the theoretical model when predicting the sequencing effort to sequence a metagenome to exhaustion (Fig. 3). A last analysis explored the semiquantitative relationship between community evenness and richness and the necessary sequencing effort to achieve a target fraction of exhaustion for a metagenome (Fig. 4).

In practice, information about a target community structure may not be available for estimating sequencing effort. The GAM built here predicts the minimum number of sequences necessary to sequence a given fraction of a target genome as a function of average genome size and the relative abundance of the target genome in the community (Fig. 5). Even without knowledge of a target community's structure, the GAM

provides a useful constraint for designing whole-genome shotgun metagenomic sequencing experiments. The size of prokaryotic genomes is fairly constrained, with 50% of the genomes in RefSeq v92 spanning from 2.74 Mbp to 5.15 Mbp (11). The limited range in genome size allows for reasonable assumptions about genome size for prokaryotes.

We wanted to relate our model to actual sequencing experiments. As such, the subsampling analysis on existing short-read data sets (individually sampled, assembled, and binned) simulated the effect of creating MAGs from data sets of different sizes from different environments. The data sets analyzed here are representative of both the sequencing effort (1 to 10 Gbp) (5) and the types of target environments that bacterial and archaeal ecologists often investigate (13). We want to emphasize that the data sets analyzed here do not necessarily reflect all the variability in characteristics of their parent communities (i.e., these data sets do not reflect global/temporal variations of these environments). Furthermore, a wide variety of software packages are available for all steps of MAG creation pipelines, and the quantity/quality of MAGs will depend on software selection, software configuration, and sequenced environment (Fig. 1) (5). The best practice is to manually curate algorithmically created MAG bins (14). Due to the large number of metagenomes that were assembled in this work ($n = 108$), manual curation was an impractical approach. As with most sequencing experiments, viral/eukaryotic DNA is likely included during assembly and binning. Thus, we do not argue that the pipeline used here is objectively optimal for generating high-quality (high-completeness and low-contamination) MAGs. Rather, our pipeline was configured to minimize contamination (MAGs with $<$10%) associated with retrieved MAGs at the expense of reduced completeness (see Text S1 in the supplemental material). For this reason, we reported the number of quality MAGs (medium- and high-quality MAGs) rather than the actual number of MAGs. Reporting quality MAGs reduces bias associated with the generalized approach used with the binning software. As such, we interpret the data sets analyzed here as reflecting general community properties (richness, abundances, and phylogeny) which are generally known to be different from one another (4, 15–17).

All communities demonstrated similar responses to rarefaction. All subsampled depths were performed in triplicate to account for possible variation in assembly and binning. As sequencing effort increased, there was an initial lag in quality MAGs followed by a rapid increase in quality MAGs, and then diminishing returns at higher sequencing efforts. Previous investigators modeled the response of 16S RNA gene (18–20), Hill's number diversity (21), taxon-resolved abundance (22), and gene abundance (22) as a function of sequencing effort with rarefaction curves, or collector's curves. The quality MAGs as a function of sequencing effort did not match a traditional collector's curve, which lacks an initial lag. Rather, the data appear sigmoidal. We modeled the data using the Gompertz function (equation 12), because its parameters can be interpreted in terms of quantities that are familiar from microbial growth curves (lag time, growth rate, and maximum density) (10). The fits to the Gompertz function illustrate that there is an optimal sequencing effort for MAG creation efforts corresponding to the upper shoulder of the Gompertz curve ("late log phase"). When sequencing effort is too close to λ, MAGs bin poorly, and when sequencing effort is too great, the number and quality of MAGs per unit sequencing effort (and therefore cost) are low. We speculate that our choice of pipeline, and specifically the fact that we discarded contigs of $<$3 kb, caused poor performance at higher sequencing effort for the *Tara Oceans* data set. Species-level microdiversity and interspecies homologous DNA can cause "bubbles," which impair assembly in larger data sets (23, 24). Improved assembly would likely have yielded more quality MAGs for our assembly of the largest subsets of the *Tara Oceans* data.

Metagenomic shotgun sequencing experimental designs should be rationally designed such that sequencing effort is selected to capture a desired fraction of a target microbial genome. Investigators should be cognizant of the rarest microbial genome desired to be characterized as well as the degree of characterization of that microbial

TABLE 2 Summary of sequence data sets analyzed with the MAG pipeline[a]

Environment	NCBI SRA accession no.	Total no. of reads[b]	No. of high-quality bases[b]	General notes[c]	Reference
Ocean surface water	ERR599029	3.372×10^8	3.340×10^{10}	Caribbean Sea (5 mbsl)	15
Estuary sediment	SRR5248164	1.137×10^8	1.589×10^{10}	Sulfate zone (8-10 cmbsf)	4
Maize soil	SRR351473	4.727×10^8	3.824×10^{10}	Surface soil	16
Human gut	SRR5127631	5.095×10^8	4.847×10^9		17

[a]The sequencing platform used for sequence data sets from all the environments shown was Illumina HiSeq 2000.
[b]Combined forward and reverse paired-end reads.
[c]mbsl, meters below sea level; cmbsf, centimeters below sea floor.

metagenome when designing a sequencing experiment. To that end, we have built a tool, Genome Relative Abundance to Sequencing Effort (GRASE), to report estimated sequencing effort required to capture a defined fraction of a genome as a function of the relative abundance of the corresponding microorganism in the community and average genome size. This R-based graphical user interface (GUI) app can be accessed online at http://adsteen.shinyapps.io/grase and is archived at http://github.com/adsteen/GRASE, from which it can be downloaded and run locally.

When the sequence read data sets analyzed here (Table 2 and Fig. 6) are reevaluated in the context of the relative abundance of a microbial metagenome (g_{MG}) sequenced to a target fraction of exhaustion (0.5), quality MAGs increase appreciably in response to minor increases in deeper characterization of the community metagenome (Fig. 7). This observation contrasts the quality MAGs response to sequencing effort (in base pairs), where substantial increases in sequencing effort (by contemporary standards) leads to diminishing returns in quality MAGs. It is important to note that abundance cutoff and sequencing effort are interchangeable; however, the responses of quality MAGs to changes in the respective predictor (i.e., base pairs versus abundance cutoff) alter the optics of the collector's curve. Modest increases in sequencing effort contribute minor amounts to extending abundance cutoff. A substantial amount of genomic and metabolic data can be gained from targeting rarer microbes (metagenomic abundances of <0.005), with the caveat that whole-genome shotgun sequencing technology (as well as computational power) requires significant increases in either the number or length of reads generated per run.

MATERIALS AND METHODS

Defining the microbial metagenome and sequencing probability. Here, we draw on set theory to provide a theoretical grounding for our *in silico* simulations described below. The expected number of sequences to sequence a fraction of an individual microbe's genome can be modeled with probability theory by defining a community metagenome with set theory. Figure 8a to e provide cartoons illustrating the application of this set theory on a hypothetical microbial population, *G*. *G* contains unique genomes (*g*) with finite abundances (*n*). The definition of microbial species is somewhat contentious (25). Here, we define *g* as a genome that is unique in length and composition for all loci compared to all other genomes in a community. As such, the species richness (*s*; unique *g*) of *G* will vary on how it is defined and should be consistent with the objectives of the investigator. In the example communities (Fig. 8a to e), *s* is 6 and the total *n* is 13. Thus, *G* is represented as follows (Fig. 8a):

$$G = \left\{ n_1 g_1, n_2 g_2 \ldots n_s g_s \middle| n \in N \right\} \quad (1)$$

where *s* is the species richness. When characterizing *G* via shotgun metagenomics, the *i*th genome, g_i, can be sequenced at *K* unique sections given a characteristic read length, *k*, and average genome size, *l*, in number of base pairs (Fig. 8b). Thus, the number of possible *k*-mers, *K*, associated with the *i*th genome, g_i, within *G* is equal to:

$$K_{g_i} = l(g_i) - k + 1 \quad (2)$$

Note that equation 2 considers homologous DNA as unique *k*-mers. This is for model simplification. From equation 2, the metagenome, g_{MG}, for g_i is defined as the set of all *k*-mers (Fig. 8c) or:

$$g_{MG,i} = \left\{ g_{i,1,1+k}, g_{i,2,2+k} \cdots, g_{i,Kg_i,Kg_i+k} \right\} \quad (3)$$

where the subscripts for g_i represent a given *k*-sized read spanning from an arbitrary starting base pair to the arbitrary starting base pair plus *k*. By substituting $g_{MG,i}$ into all *g* for equation 1 (Fig. 8d), the

metagenome for a microbial community, G_{MG}, is derived to be:

$$G_{MG} = \{n_1 g_{MG,1}, n_2 g_{MG,2} \ldots n_s g_{MG,s} | n \in N\} \quad (4)$$

while the population of *k*-mers in the metagenome, G_{MG} (Fig. 8e), is represented as:

$$K_{MG} = \{g_{MG,1}, g_{MG,2} \ldots g_{MG,s}\} \quad (5)$$

From equation 5, one can determine the cardinality, or the total number, of *k*-mers associated with G_{MG} (expressed as $|K_{MG}|$). To an effect, $|K_{MG}|$ is analogous with "metagenomic richness" of an environment. When attempting to fully sequence G_{MG} using shotgun metagenomics, we assume that sampling events (sequence reads) are independent and are sampled with replacement (26).

The probability of sequencing all elements in G_{MG} reduces to a coupon collector problem (27) by making the above assumptions. Using the general functional form for calculating expected samples for sampling all unique elements in a set (equation 13b in reference 8), one can predict the number of sequences necessary to sequence all elements in K_{MG}, such that the expected number of sequences, $E(G_{MG})$, is:

$$E(G_{MG}) = \int_0^\infty \left(1 - \prod_{j \in K_{MG}} \left(1 - e^{-p_j t}\right)\right) dt \quad (6)$$

where *j* is a given element within K_{MG}, *t* is the number of sampling events, and p_j is equal to the proportion of the *j*th *k*-sized read within a given population of *k*-sized reads. p_j can be expressed as follows:

$$p_j = \frac{n_i \times j \in K_{MG}}{|G_{MG}|} \quad (7)$$

where n_i is the respective abundance for the species whose MAG contains the *j*th *k*-sized read within K_{MG}, and $|G_{MG}|$ is the cardinality of G_{MG}, or the total number of *k*-sized reads in the metagenome, G_{MG}.

Modeling expected sequences. Equation 6 provides an estimate for the total number of sequences to sequence all K_{MG}. The influence of increasing species richness (i.e., *s* in equation 1) on the expected number of sequences was tested for four hypothetical communities. The first community had an even structure such that all *k*-mers were equally distributed across all K_{MG}. In the second community, 90% of the *k*-mers were equally distributed in 50% of K_{MG}, and the remaining 10% of the *k*-mers were distributed equally across the remaining 50% of K_{MG}. This community represented a community with relatively moderate species evenness. In the third community, 90% of the *k*-mers were equally distributed across 10% of K_{MG}, and the remaining 10% of the *k*-mers were distributed equally across the remaining 90% of K_{MG}. This community represented a community with relatively low species evenness. The last community had 10 equally sized groups. The abundance of the *k*-mers in each group was based on the function form of a lognormal community (28) which has been observed in microbial populations (e.g., references 21 and 29), such that:

$$S(R) = S_0 e^{-a^2 R^2} \quad (8)$$

where S_0 was treated as the maximum relative of abundance ($S_0 = 1$), *a* was the inverse width of the distribution, *R* was treated as the positive octave range spanning 0 to 9, and *S*(*R*) represented the abundance for a given group. For the lognormal abundance distribution in Fig. 2d, *a* was set at a value of 0.2. Each hypothetical community started with $|K_{MG}| = 1 \times 10^6$. $|K_{MG}|$ incrementally increased at 10 equally spaced, linear steps to a maximum of $|K_{MG}| = 1 \times 10^8$. As $|K_{MG}|$ increased, all community structures remained constant. Graphical representation of rank abundance in Fig. 2a to d was normalized by a given $|K_{MG}|$ to reflect that populations retained the same structure even as population size varied. We defined a normalized rank abundance r_n such that

$$r_n = \frac{r}{s} \quad (9)$$

where *r* and *s* are untransformed rank abundance and richness, respectively. For each community, at each step, the expected number of sequences was calculated using equation 6. The expected number of sequences as a function of $|K_{MG}|$ were modeled with linear regressions.

Equation 6 gives the expected number of sequences required to sequence any size community to exhaustion. Numerical sequencing simulations were performed to determine the number of sequences necessary to sequence a subset of all *k*-mers (K_{MG}). These numerical sequencing simulations were applied to four hypothetical community structures described above. Numerical simulations were performed such that $|K_{MG}| = 3 \times 10^7$, 4×10^7, 5×10^7, 7×10^7, 9×10^7, and 1×10^8. During each of these simulations, the parameters read length (*k*) and average genome size (*l*) were set at 100 and 1×10^6, respectively, for all *g*. Random elements from K_{MG} were selected with replacement to simulate sequencing events. Numerical simulations were performed until the percentage of $|K_{MG}|$ sequenced was 50%, 70%, 90%, 95%, 99%, or 100%. A weight distribution was applied to elements in a given K_{MG}. The weight distribution biased sequencing to reflect the relative abundances of the four hypothetical communities described above. The percentage of $|K_{MG}|$ sequenced was evaluated every 1×10^7 sequences. Numerical simulations were performed in triplicate for all $|K_{MG}|$ and all target fractions of $|K_{MG}|$.

We explored the influence of community evenness on required sequencing effort by performing sequencing simulations on six different lognormally distributed communities. The numerical sequencing simulations followed the simulations described above. The six lognormal communities were modeled such that each community had $S_0 = 1$, $R = 10$, and $|K_{MG}| = 1 \times 10^7$. The values of *a* for the six lognormal

distributions were as follows: $a = 0$, $a = 0.005$, $a = 0.008$, $a = 0.01$, $a = 0.015$, and $a = 0.02$. Evenness was represented using the Pielou evenness index (9), or the ratio of the Shannon diversity index (30) for an observed community to an even community of equal richness. Shannon diversity was calculated in the context of a metagenomes such that:

$$H_{\mathrm{MG}} = \sum_{j \in K_{\mathrm{MG}}} - p_j \log(p_j) \tag{10}$$

where p_j is the proportion that the jth k-sized read represents among all k-mers in the metagenome. Thus, the Pielou evenness index (9) was calculated such that:

$$J = \frac{H_{\mathrm{MG}}'}{H_{\mathrm{MG,max}}} \tag{11}$$

where J was the Pielou evenness index, H_{MG}' was the metagenome Shannon diversity index, and $H_{\mathrm{MG,max}}$ represented the metagenome Shannon diversity index when all p_j were equal ($a = 0$).

Last, numerical simulations were performed to determine the sequencing effort necessary to achieve a target fraction for an individual metagenome (g_{MG}). Target fractions increased from 0.5 to 1 at 100 linearly spaced intervals. The fraction of the metagenome community (G_{MG}) that g_{MG} represented varied from 1% to 100% in 30 lognormally spaced intervals. The target genome sizes varied such that $l = 0.5 \times 10^6$, $l = 1 \times 10^6$, $l = 2 \times 10^6$, $l = 3 \times 10^6$, $l = 5 \times 10^6$, $l = 10 \times 10^6$, $l = 15 \times 10^6$, and $l = 20 \times 10^6$. The sequencing effort for a given combination of target fraction, genome size, and fraction of the metagenome community was modeled using the GAM function (mgcv R package [31]). Further description of the GAM regression is provided in Text S1 in the supplemental material.

Sequence data sources. A more detailed description of the data sources is provided in Text S1. All data sets analyzed in this study are summarized in Table 1.

MAG assembly pipeline. The pipeline developed here followed similar pipelines described by other authors (3, 32). A more detailed description of the pipeline is provided in Text S1.

Subsampling sequence read data sets. A description of the sampling methodology is provided in Text S1.

Modeling MAG response to sequencing effort. Medium-quality and high-quality MAGs were defined on the basis of completeness and contamination from CheckM (33). Medium-quality MAGs were defined as MAGs with >50% completeness and <10% contamination. High-quality MAGs were defined as MAGs with >90% completeness and <5% contamination (14). The sum of medium- and high-quality MAGs as a function of sequencing effort was modeled for environmental sequence data sets using the Gompertz equation, as reformulated by Zweitering et al. (10) for use with microbial growth curves:

$$g_{\mathrm{eq}}(A, \mu, \lambda, b) = A \times e^{-e^{\frac{\mu \times e}{A}(\lambda - b)+1}} \tag{12}$$

where A, μ, and λ are fit coefficients and b is high-quality bases (Gbp). MAG yield could be defined as:

$$\text{MAG yield} = \frac{n_{\mathrm{MQ}} + n_{\mathrm{HQ}}}{A} \tag{13}$$

where n_{MQ} is the total medium-quality MAGs derived from the subsampling experiment, n_{HQ} is the total high-quality MAGs derived from the subsampling experiment, and A is from equation 12.

Relating MAG response to the theoretical sequencing model. Details of how we related MAG response to the theoretical sequencing model are provided in Text S1.

Data availability. All simulations and codes used for modeling sequencing effort are freely available on Github at https://github.com/taylorroyalty/sequence_simulation_code. All data generated during the subsampling experiment is available in Data Set S1 in the supplemental material. The GRASE GUI application is available at http://adsteen.shinyapps.io/grase.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00384-19.

TEXT S1, DOCX file, 0.02 MB.

DATA SET S1, CSV file, 6.7 MB.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation grant OCE-1431598 to A.D.S. and C-DEBI subawards to T.M.R. and A.D.S. Funding for open access to this research was provided by the University of Tennessee's Publishing Support Fund.

REFERENCES

1. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. Nat Rev Microbiol 16:410. https://doi.org/10.1038/s41579-018-0029-9.
2. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-

assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542. https://doi.org/10.1038/s41564-017-0012-7.
3. Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 5:1–8. https://doi.org/10.1038/sdata.2017.203.
4. Baker BJ, Lazar CS, Teske AP, Dick GJ. 2015. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. Microbiome 3:14. https://doi.org/10.1186/s40168-015-0077-6.
5. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844. https://doi.org/10.1038/nbt.3935.
6. Lander ES, Waterman MS. 1988. Genomic mapping by end-characterized random clones: a mathematical analysis. Genomics 2:231–239. https://doi.org/10.1016/0888-7543(88)90007-9.
7. Wendl MC, Kota K, Weinstock GM, Mitreva M. 2013. Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. J Math Biol 67:1141–1161. https://doi.org/10.1007/s00285-012-0586-x.
8. Rodriguez-R LM, Konstantinidis KT. 2014. Estimating coverage in metagenomic data sets and why it matters. ISME J 8:2349–2351. https://doi.org/10.1038/ismej.2014.76.
9. Pielou EC. 1966. The measurement of diversity in different types of biological collections. J Theor Biol 13:131–144. https://doi.org/10.1016/0022-5193(66)90013-0.
10. Zwietering MH, Jongenburger I, Rombouts FM, van 't Riet K. 1990. Modeling of the bacterial growth curve. Appl Environ Microbiol 56:1875–1881.
11. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189.
12. Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N. 2016. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. Nat Microbiol 2:1–6. https://doi.org/10.1038/nmicrobiol.2016.242.
13. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161. https://doi.org/10.1007/s10142-015-0433-4.
14. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731. https://doi.org/10.1038/nbt.3893.
15. Karsenti E, Acinas SG, Bork P, Bowler C, de Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie JM, Follows M, Gorsky G, Hingamp P, Iudicone D, Jaillon O, Kandels-Lewis S, Krzic U, Not F, Ogata H, Pesant S, Reynaud EG, Sardet C, Sieracki ME, Speich S, Velayoudon D, Weissenbach J, Wincker P. 2011. A holistic approach to marine eco-systems biology. PLoS Biol 9:e1001177-11. https://doi.org/10.1371/journal.pbio.1001177.
16. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. 2014. Tackling soil diversity with the assembly of large, complex metagenomes. Proc Natl Acad Sci U S A 111:4904–4909. https://doi.org/10.1073/pnas.1402564111.
17. Schirmer M, Smeekens SP, Vlamakis H, Jaeger M, Oosting M, Franzosa EA, ter Horst R, Jansen T, Jacobs L, Bonder MJ, Kurilshikov A, Fu J, Joosten LAB, Zhernakova A, Huttenhower C, Wijmenga C, Netea MG, Xavier RJ. 2016. Linking the human gut microbiome to inflammatory cytokine production capacity. Cell 167:1125–1136.e8. https://doi.org/10.1016/j.cell.2016.10.020.
18. Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, Daroub SH, Camargo FAO, Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1:283–290. https://doi.org/10.1038/ismej.2007.53.
19. Feng BW, Li XR, Wang JH, Hu ZY, Meng H, Xiang LY, Quan ZX. 2009. Bacterial diversity of water and sediment in the Changjiang estuary and coastal area of the East China Sea. FEMS Microbiol Ecol 70:236–248. https://doi.org/10.1111/j.1574-6941.2009.00772.x.
20. Rintala A, Pietilä S, Munukka E, Eerola E, Pursiheimo JP, Laiho A, Pekkala S, Huovinen P. 2017. Gut microbiota analysis results are highly dependent on the 16s rRNA gene target region, whereas the impact of DNA extraction is minor. J Biomol Tech 28:19–30. https://doi.org/10.7171/jbt.17-2801-003.
21. Kang S, Rodrigues JLM, Ng JP, Gentry TJ. 2016. Hill number as a bacterial diversity measure framework with high-throughput sequence data. Sci Rep 6:1–4. https://doi.org/10.1038/srep38263.
22. Zaheer R, Noyes N, Ortega Polo R, Cook SR, Marinier E, Van Domselaar G, Belk KE, Morley PS, McAllister TA. 2018. Impact of sequencing depth on the characterization of the microbiome and resistome. Sci Rep 8:5890. https://doi.org/10.1038/s41598-018-24280-8.
23. Ghurye JS, Cepeda-Espinoza V, Pop M. 2016. Metagenomic assembly: overview, challenges and applications. Yale J Biol Med 89:353–362.
24. Lonardi S, Mirebrahim H, Wanamaker S, Alpert M, Ciardo G, Duma D, Close TJ. 2015. When less is more: 'slicing' sequencing data improves read decoding accuracy and de novo assembly quality. Bioinformatics 31:2972–2980. https://doi.org/10.1093/bioinformatics/btv311.
25. Rosselló-Móra R, Amann R. 2015. Past and future species definitions for Bacteria and Archaea. Syst Appl Microbiol 38:209–216. https://doi.org/10.1016/j.syapm.2015.02.001.
26. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. https://doi.org/10.1038/nature07517.
27. Flajolet P, Gardy D, Thimonier L. 1992. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. Discret Appl Math 39:207–229. https://doi.org/10.1016/0166-218X(92)90177-C.
28. Magurran AE. 1988. Ecological diversity and its measurement, 1st ed. Croom Helm Ltd., London, United Kingdom.
29. Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. 2009. Ecology of the rare microbial biosphere of the Arctic Ocean. Proc Natl Acad Sci U S A 106:22427–22432. https://doi.org/10.1073/pnas.0908284106.
30. Shannon CE. 1948. A mathematical theory of communication. Bell Syst Tech J 27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.
31. Wood S. 2017. mgcv: mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation.
32. Graham ED, Heidelberg JF, Tully BJ. 2017. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. PeerJ 5:e3035. https://doi.org/10.7717/peerj.3035.
33. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.