Simulation-based approaches to characterize the effect of sequencing depth on the quantity and quality of metagenome-assembled genomes

Taylor Royalty[a], Andrew D. Steen[a*]

[a]Department of Earth and Planetary Sciences, University of Tennessee – Knoxville

Running Title: Modeling metagenome-assembled genome properties

* Corresponding: asteen1@utk.edu

Abstract Word Count: 230

Importance Word Count: 108

Text Word Count: 4999

Abstract

We applied simulation-based approaches to characterize how sequencing depth influences the properties of genomes identified in metagenomes assembled from short read sequences. An initial analysis evaluated the quantity, completion, and contamination of metagenome-assembled genomes (MAGs) as a function of sequencing depth on four preexisting sequence read datasets taken from four environments: a maize soil, an estuarine sediment, the surface ocean, and the human gut. These were subsampled to varying degrees in order to simulate the effect of sequencing depth on MAG binning. The property, MAG quantity fit the Gompertz curve, which has been used to describe microbial growth curves. A second analysis explored the relationship between sequencing depth and the proportion of available metagenomic DNA sequenced during a sequencing experiment as a function of community richness, evenness, and genome size. Typical sequencing depths in published experiments (1 to 10 Gb) reached the point of diminishing returns for MAG creation. Simulations from the second analysis demonstrated that both community richness and evenness influenced the amount of sequencing required to sequence a metagenome to a target fraction of exhaustion. The most abundant genomes required comparable quantities of bases sequenced regardless of community evenness, while more uneven communities required considerably more sequences to fully sequence rarer members. Future whole-genome shotgun sequencing studies can use an approach comparable to the one described here to estimate the quantity of sequences required to achieve scientific objectives.

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 are assembled into metagenome-assembled genomes which allow metabolic processes influencing health,

agriculture, and biogeochemical cycles to be assigned to microbial clades. At present, no reliable guidelines exist to select sequencing depth as a function of experimental goals in metagenome-assembled genomes creation projects. The work presented here provides a framework for obtaining a constrained estimate on the number of short read sequences needed for sequencing microbial communities. Results suggested that both the microbe community richness and evenness influence the amount of sequencing in a predictable matter.

Introduction

The assembly of high-accuracy short read sequences into metagenome-assembled genomes (MAGs) is a recent 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 helped constrain the relationships between microbial ecology, microbial metabolisms, and biogeochemistry. At present, there is little information to guide how much sequencing is appropriate for metagenomic shotgun sequencing experiments (5). For the year 2017, estimates compiled by Quince et al. (5) suggest that up till now, metagenomic shotgun sequencing experiments usually sequence between 1 Gb and 10 Gb DNA nucleotides. Nonetheless, more guidance is necessary for selecting an appropriate metagenomic shotgun sequencing depth for one's experimental question which balances the maximization of information and minimization of cost.

Illumina sequencing technology is currently the most popular platform to generate metagenomic shotgun sequences (5). Here we present two distinct analyses which constrain the relationship between the quantity of Illumina metagenomic shotgun sequences and the quantity and quality of retrieved MAGs. First, we performed *in silico* experiments simulating the effect of how sequencing depth on Illumina sequence read datasets impacted the retrieved MAG properties for these datasets. Second, we applied a theoretical model and numerical simulations to estimate the minimum sequencing depth needed to sequence a metagenome to a target fraction of exhaustion. The work presented here illustrates how community evenness and richness control the sequencing depth necessary to sequence a metagenome to a target fraction of exhaustion. These patterns can be used to

guide sequencing depth decisions for future sequencing efforts in which MAG creation is a primary goal.

Results

MAG ASSEMBLY AS A FUNCTION OF SEQUENCING DEPTH IN EXISTING METAGENOMIC DATASETS

The number of "effective MAGs" (equivalents to100%-complete MAGs, as defined in the Methods section) as a function of high quality bases empirically fit the Gompertz equation (equation 1; Fig 1B; parameters in Table 1). For each environment, the data fit the Gompertz equation better than a linear least-squares fit based on Akaike Information Criterion (AIC) (6). This equation is formulated for applications with microbial growth curves, such that the parameters A, μ, and λ correspond to maximum cell density, growth rate, and lag time (Fig 1A). Here, A, μ, and λ correspond to the maximum number of effective MAGs assembled with the pipeline, the maximum rate which effective MAGs form as with more sequencing, and the "lag bases," or the bases which must be sequenced prior to rapid retrieval in effective MAGs. For the estuary, maize, and human gut datasets, MAG yield began to asymptote at higher sequencing depths, which indicates that further sequencing would yield diminishing returns with our pipeline. The Tara Ocean dataset followed a similar pattern at <25 Gb. However, when the number of sequenced bases was >25Gb, the number of effective MAGs decreased and became insensitive to sequencing depth. Since we have expressed MAG creation in terms of effective MAGs, the actual number of MAGs created in each example was considerably higher.

Mean MAG completeness also increased towards an asymptote with increasing sequencing depth (Fig 1C). Completeness was highest for the human gut dataset, with a maximum of 23.9%, and increased continuously as sequencing depth increased. The mean MAG completeness reached an asymptote of ~10-15% for the other three datasets with sequenced bases >10 Gb. Note that when >10 Gb were sequenced, the number of effective MAGs created still increased as new sequences were

88 added. For all datasets, mean MAG contamination was <2% (Fig 1D) and did not depend strongly on
89 sequencing depth.

90 SIMULATION EXPERIMENTS

91 Using equation 7, we calculated the number of *k*-length sequence reads required to sequence all
92 unique DNA sequences of length, *k* (*k*-mers), in four hypothetical metagenomes. Three of the
93 community structures are ecologically unrealistic but represented a community in which taxa are
94 distributed perfectly evenly, highly unevenly, and at an intermediate level of evenness (Fig 2A-C). The
95 fourth community structure, which is lognormally distributed, is ecologically realistic (Fig 2D; (7, 8)).
96 The expectation value of the log number of sequences required to fully sequence metagenomes of those
97 hypothetical communities was linear with respect to log-transformed size of the metagenome (i.e.,
98 number of unique *k*-mers in the population, approximate number of unique base pairs in a
99 metagenome); this suggests a power-law relationship between metagenome size and expectation value
100 of sequence reads required to sequence the metagenome to exhaustion (Fig 2E). For all community
101 structures, the slope of the relationship between log-transformed sequenced reads and log-transformed
102 unique number of sequenced reads was within 1% of 1.06. The structure of the population strongly
103 influenced the number of reads required such that more even community structures required far fewer
104 reads than less even structures.

105 As equation 7 only estimates the number of reads to sequence a metagenome to exhaustion, we
106 used a numerical simulation to estimate the number of *k*-sized reads to sequence a metagenome to a
107 target fraction of exhaustion. Numerical simulation results predicted the same number of sequences
108 reads to sequence 100% of a given metagenome as the numerically integrated expected sequences from
109 equation 7 (Fig 3); this supported the use of this simulation. The log-transformation of both total
110 unique *k*-sized reads ($|K_{MG}|$ and sequenced reads showed a linear response for all target fractions and

all community structures. The amount of sequences required to achieve a given target of $|K_{MG}|$ was variable for the different communities shown in Fig 2A. For instance, the lognormally-distributed community required the most amount of sequencing to sequence a metagenome to a target fraction of exhaustion but required similar amount of sequencing to sequence the metagenome to a target fraction of 50% as the other communities.

We applied the simulation to semi-quantitatively demonstrate the effect that community evenness has on the number of reads required to sequence a community to a target fraction of completion. These communities ranged from perfectly even (a=0, eq. 9) 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). Computational limits precluded simulating communities with Pielou evenness less than 0.977 given the richness and size of genomes within the communities. The number of sequence reads required to sequence genomes to a target fraction of completion depended strongly on both the evenness and the target fraction of completion (Fig 4B). Again, more even communities required more sequence reads than less even communities. The strength of this relationship also depended on the target fraction of completion. A community with Pielou evenness of 0.97 required 3 orders of magnitude more sequence reads to sequence a metagenome to a target fraction of exhaustion than a perfectly even community while the same community only required about 42% more reads to sequence 50% of the metagenome.

The minimum number of sequence reads required to sequence a microbe genome given a combination of target fraction, genome size, and fraction of the metagenome community was modeled with a generalized additive model. The smooth dimensions for target fraction, genome size, and fraction of the metagenome community was 7, 3, and 9, respectively, to achieve a normal distribution of residuals. To normalize for different sequence read length, sequence reads were converted to bases

and ranged from 1×10^{7} to 1×10^{13}. More bases were required to sequence microorganisms when 1) the genome was relatively rarer in the community, 2) to achieve better coverage of the genome, and 3) when the genome increased in size.

Discussion

We sought to establish evidence-based guidelines for selecting a sequencing depth during shotgun metagenomic sequencing experiments with the goal of creating MAGs of a given quantity and quality. Random subsamples of existing short read datasets, which were each individually assembled and binned, simulated the effect of creating MAGs from datasets of different sizes and environments. The datasets analyzed here are argued to be representative of both the order of magnitude of sequencing depth (1 to 10 Gb) (5) and the types of target environments microbial ecologists often investigate (10). A variety of software is available for all steps of MAG creation pipelines, and the quantity/quality of MAGs will depend on software selection, software configuration, and sequenced environment (5). Furthermore, it is best-practice to manually curate algorithmically-created MAG bins (11). We do not argue that the pipeline used here is objectively optimal for generating "true" MAGs (i.e., represent true genomes). Thus, MAG quantity was not directly reported but expressed as effective MAGs. The metric, effective MAGs, represents the integrated completeness (12) divided by 100 for MAGs retrieved with a taxonomic rank of at least phylum. In effect, effective MAGs represents phylogenetic signal, as defined by the presence of marker genes in assembled contigs (necessary for constructing MAGs). Thus, increases in effective MAGs should scale proportionally with increases in the quantity of true MAGs.

As sequencing depth increased, there was at first a "lag time" (more precisely a lag depth, or number of bases before effective MAGs began to increase) followed by a rapid increase in effective MAG quantity, and then diminishing returns at higher sequencing depths. Previous investigators

157 modeled the response of 16S RNA gene (13–15), Hill's number diversity (16), taxon-resolved
158 abundance (17), and gene abundance (17) as a function of sequencing depth using rarefaction curves,
159 or collectors curves. The effective number of MAGs created did not match a traditional collector's
160 curve, which does not contain any initial lag. The Gompertz function, conversely, fit the data well,
161 suggesting that MAG construction as a function of sequencing depth behaves similarly to microbial
162 growth in a constrained medium, in concept if not in precise mechanism. The Gompertz function is
163 defined in terms of three parameters, A, μ, and λ. These parameters correspond to the maximum
164 effective MAGs at infinite sequencing depth (A), maximum rate that effective MAGs increased with
165 increases in sequencing depth (μ), and a minimum threshold of sequencing necessary prior to rapid
166 effective MAGs retrieval (λ) (Fig 1A). The Gompertz equation achieves the same asymptotic behavior
167 of conventional rarefaction models while also modeling the apparent lag (λ) in effective MAGs
168 observed during this work (Fig 1B).

169 The four environments analyzed demonstrated different responses to increases in sequencing
170 depth. Specifically, the predicted maximum effective MAGs varied from ~17 to ~97, the predicted
171 maximum rate that effective MAGs increased varied from ~1.4 to ~5.8, and the minimum threshold of
172 sequencing necessary prior to seeing effective MAGs varied from ~0.6 to ~6.7. The Tara Ocean
173 dataset, where effective MAGs decreased at sequencing depth >20 Gbp, was an exception. We
174 speculate that our choice of pipeline, and specifically the fact that we discarded contigs <3kb, caused
175 poor performance at higher sequencing depth for the Tara Ocean dataset.

176 As mean MAG completeness converged to an asymptote considerably less than 100% (Fig 1B),
177 MAG yields (Table 1) were close to 100%. This suggests the maximum effective MAGs (A) likely
178 represents sequence reads associated with abundant MAGs. Thus, we asked how much sequencing was
179 necessary to sequence a community to exhaustion. The expected number of sequence reads required to

sequence an entire metagenome was estimated using equation 7 for four hypothetical communities (Fig 2A-D). The total unique *k*-sized reads (i.e., richness) and community structure influenced how much sequencing is necessary to sequence an entire metagenome (Fig 2E). For a given community structure, increases in community richness lead to linear increases the sequencing depth necessary to exhaust the metagenome. All regressions had similar slopes, indicating that community structure did not exert a major influence on that relationship. Interestingly, the sequencing depth necessary to sequence an entire metagenome depended strongly on the structure of the target microbial community (Fig 2E). As sequencing depth was log-transformed in Fig 2E, the differences in model intercepts indicate orders of magnitude differences in the necessary sequencing depth. The primary implication of Fig 2 is that the sequencing depth increased in a predictable trend in response to richness, regardless of the community structure.

One limitation to equation 7 is that it only provides an estimate of the sequencing depth required to sequence a metagenome to exhaustion. For practical applications, a continuous increase in sequencing depth eventually leads to diminishing returns in identifying unique sequence reads while also leading to a disproportional increase in monetary resources needed to find these unique sequence reads (18). Thus, it is desirable to constrain the fraction of unique sequence reads (e.g., 50%, 70%, 90%, etc.) sequenced from a metagenome in relation to monetary investment necessary to achieve that fraction of a metagenome. Simulations show that as target metagenome completeness increases, the sequencing depth required increases dramatically (Fig 3). Simulation results were validated by comparing the sequencing depth necessary to sequence 100% of a metagenome with predictions from equation 7. While the numerical approach successfully reproduced and extended equation 7, communities with large values of richness ($|K_{MG}| > 1 \times 10^8$ became computationally burdensome. Nonetheless, when the target fraction and community structures were held constant, the linear increase

in sequencing depth as a function of increased richness suggests linear regression may be sufficient to estimate sequencing depth for communities with large values of richness.

One observation from the numerical simulations was the impact that community structure had on the required depth of sequencing (Fig 2E and 3). Even communities required less sequencing to achieve a fraction of $|K_{MG}|$. Conceptually this makes sense, as abundant taxa (i.e., large *n* values in equation 3) should be sequenced more deeply compared to rarer taxa. To further explore the influence that community evenness had on required sequencing depth, communities with similar and more realistic lognormal structures (7, 16) at different levels of evenness were compared to one another (Fig 4A). Decreasing evenness (increasing *a*; equation 9) led to both increases in the sequencing depth required to sequence a given target fraction of $|K_{MG}|$ (Fig 4B). For communities with more uneven species distributions, rarer community members required more sequencing. While only semi-quantitative, this analysis demonstrates that community evenness can have a significant impact on the sequencing depth necessary to characterize an entire community.

In practice, information about a target community structure may not be available for estimating sequencing depth. The spline model built here illustrates the minimum number of sequences necessary to sequence a given fraction of a target genome, assuming genome size and proportion that the genomic content represents in the community metagenome (G_{MG}) (Fig 5). This proves useful for constraining the observed MAG properties from one's bioinformatic pipeline (e.g., Fig 1B-D) in the context of what proportion of a given microbe's metagenome (g_{MG}; equation 4) has been sequenced to exhaustion. For example, taking the 5 Gb human gut dataset analyzed here (Table 2), if a microbe with a genome size of ~5 Mbp existed from this environment, then Fig 5C suggests that a 5 Mbp genome representing >10% of the whole metagenome (G_{MG}; equation 5) will be sequenced to a minimum of 50% to exhaustion. More so, one has constrained perspective of how a given genome may be

represented in the retrieved MAGs. Although the simple nature of sequencing a genome may not necessarily translate into the production of more MAGs, one can safely say that additional sequencing of that 5 Mbp genome which represents >10% of the community will not lead to the addition of more MAGs. More so, the bioinformatic pipeline would act as the limiting step (opposed to sequencing) in the production of MAGs.

Materials and Methods

Sequence data sources

All sequence data were downloaded from NCBI's Sequence Read Archive (SRA) using the SRA Toolkit (fastq-dump –split-files) (19). Exact duplicate reads for both forward and reverse reads were removed using PRINSEQ (-derep 1; v0.20.4) (20). All sequencing datasets were limited to Illumina shotgun metagenomic paired-end reads. Four datasets were analyzed for this analysis. The first dataset was from oceanic surface water collected at 5m depth in the Caribbean Sea as a part of the Tara Oceans expedition (21). The second dataset was from sediment from a depth of 8-10 cm below the surface (sulfate-rich zone) and collected at the White Oak River Estuary, Station H, North Carolina, USA (4). The third dataset was collected from maize soil (22). The last dataset was collected from human fecal samples and represented a human gut microbiome (23). All datasets analyzed in this study are summarized in Table 1.

MAG Assembly Pipeline

The pipeline developed here followed similar pipelines described by other authors (3, 24). All sequence datasets were analyzed as follows. Trimmomatic (v0.36) (25) removed adapters as well as trimmed low-quality bases from the ends of individual reads. Read leading and trailing quality scores were required to be >3. The sliding window was set to 4 base pairs and filtered base pair windows with a

mean score <15. Quality controlled reads were assembled into contigs using MEGAHIT (v1.1.2; --presets meta-large) (26). Due to RAM limitations, assembled contigs <3000 bp in length were excluded from the analysis. Redundant contigs were removed using CD-HIT (v4.6.8; cd-hi-est -c 0.99 -n 10) (27). Similarity among the remaining contigs was further evaluated via intra-contig sequence alignments using Minimus2 (-D OVERLAP=100 MINID=95). The quality-controlled reads (i.e., after using Trimmomatic) were then mapped to the remaining contigs using Bowtie 2 (v2.3.3) (28) to generate a coverage score for individual contigs.

Resultant contigs were iteratively clustered into MAGs using the unsupervised clustering algorithm Binsanity (v0.2.6) (24). Similar to Tully et al. (3), six initial clustering iterations were performed with the parameter, *preference* (-p), set to -10 (iteration 1), -5 (iteration 2), -3 (iteration 3-6). Between iterations, a refinement step (Binsanity-refine) was performed on the putative MAGs with constant *preference* (-p) of -25. The refined putative MAGs were evaluated for contamination and completeness using the software CheckM (v1.0.6) (12), which uses HMMER (v3.1) and Prodigal (v2.6.3) (29). Contigs associated with putative MAGs meeting one of the following criteria: 1) had a completeness > 90% and contamination < 10%, 2) had a completeness > 80% and contamination < 5%, or 3) had a completeness > 50% and contamination < 5% were treated as high-quality. All other MAGs were considered low-quality MAGs. MAGs defined as high-quality were not modified any further. Contigs associated with the high-quality MAGs were not used in the subsequent reclustering and refinement steps. The contigs associated with low-quality MAGs were pooled together and reclustered during the next iteration of Binsanity clustering. After the sixth iteration, the remaining MAGs which did not fall into one of the three categories underwent additional refinement using Binsanity-refine. During this step, MAGs were iteratively refined with *preference* set to -10 (iteration 1), -3 (iteration 2), and -1 (iteration 3). Between each refinement step, metrics of contamination and completeness were

evaluated using CheckM. Again, MAGs which met the criteria of one of the high-quality categories described above were not further modified. The respective contigs to the putative MAGs were not used in proceeding refinement steps. After the last iteration of refinement, all MAGs were reevaluated for completeness and contamination as well as assigned a final taxonomic rank using CheckM. Completeness and contamination values for MAGs with the resolved taxonomic rank of phylum were integrated together. The integrated completeness was then divided by 100 to produce effective number of MAGs.

Subsampling Sequence Read Datasets

The effect of decreased sequencing depth was simulated by subsampling the initial sequence read datasets described above. Downloaded sequence read datasets were randomly sampled at set fractions of 1%, 10%, 20%, 40%, 60%, 80%, 90%, 95%, and 100%. To account for variability in the reads sampled at a given fraction, each fraction was resampled, assembled, and binned in triplicate. All triplicates were analyzed using the MAG assembly pipeline described above.

Modeling MAG Response to Sequencing Depth

Effective MAGs as a function of sequencing depth was modeled for environmental sequence datasets using the Gompertz equation, as reformulated by Zweitering et al. (30) for use with microbial growth curves:

$$\text{Effective MAGs} = A \times e^{-e^{\frac{\mu \times e}{A}(\lambda - b)+1}} \quad (1$$

where A, μ, and λ are fit coefficients and b is high-quality bases. To assess the validity of this function, AIC (6) was calculated for all Gompertz equation fits and compared to AIC values for linear regressions models for same dataset.

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 application of probability theory for predicting the expected number sequences to sequence a metagenome became founded by defining a metagenome as the set of available metagenomic DNA that can be sequenced in a sequencing experiment. Fig 6A-E provides a cartoon example illustrating the application of this set theory on a hypothetical microbial population, *G*. *G* is a community of genomes (*g*) with finite abundances (*n*). As the definition of microbial species is somewhat contentious (31), *g* is taken as the average genome for all individual genomes defined as a meeting some criteria defining a taxonomic rank. Thus, the richness (*s*) of *G*, or the total number of *g*, depends on the definition of *g*. In the example *G* (Fig 6A-E), *s*=6 and the total *n*=13. Thus, *G* can be represented as (Fig 6A):

$$G = \{n_1 g_1, n_2 g_2 \dots n_s g_s | n \in N\} \quad (2$$

where *s* is the total number of unique species within the community (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 6B). Thus, the number of unique *k*-sized reads, *K*, associated with the i^{th} genome, g_i, within *G* is equal to:

$$K_{g_i} = l(g_i) - k + 1 \quad (3$$

From equation 3, the metagenome, g_{MG}, for g_i is defined as the set of all unique possible *k*-sized reads (Fig 6C) or:

$$g_{MG,i} = \{g_{i,1,1+k}, g_{i,2,2+k} \dots, g_{i,Kg_i,Kg_i+k}\} \quad (4$$

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 2 (Fig 6D), the metagenome for a microbial community, G_{MG}, is derived to be:

$$G_{MG} = \{n_1 g_{MG,1}, n_2 g_{MG,2} \dots n_s g_{MG,s} | n \in N\} \quad (5$$

while the population of unique *k*-sized reads in the metagenome, G_{MG} (Fig 6E), is represented as:

$$K_{MG} = \{g_{MG,1}, g_{MG,2} \dots g_{MG,s}\} \quad (6$$

From equation 4, one can determine the cardinality, or the total number, of unique *k*-sized reads in associated with G_{MG} (expressed as $|K_{MG}|$). When attempting to fully sequence G_{MG} using shotgun metagenomics, we assume that sampling events (sequence reads) are independent and are sampled with replacement. In fact, Illumina sequencing technology sequences reads in parallel via the individual DNA fragments binding to individual clusters. Furthermore, the fragmented DNA cannot be sequenced twice as the sequencing process is destructive (32). Nonetheless, the mass of DNA extracted from a target environment will represent a negligible fraction of the total DNA which exists in that environment. As the relative abundance of the *k*-sized reads in K_{MG} does not change when DNA is extracted from an environment, sampling events can be treated as independent and thus, DNA sampling reduces to sampling with replacement. If the proportion DNA mass extracted had a significant impact on the remaining mass of DNA in the environment, then one would be more suited to sequence all the DNA versus a smaller proportion of the DNA. The sequencer should have no impact on sampling assuming no sequencing errors due to misreading or spatial sampling issues (i.e., clonal density issues). Obviously, these issues do exist, but for the sake of a first order, general approximation, these biases can be ignored.

By making the above assumptions, the probability of sequencing all elements in G_{MG} reduces to a coupon collectors problem (33). Using the general functional form for calculating expected samples for sampling all unique elements in a set (equation 13b in 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:

$$E(G_{MG}) = \int_0^\infty \left(1 - \prod_{j \in K_{MG}} (1 - e^{-p_j t})\right) dt \quad (7$$

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:

$$p_j = \frac{n_i \times j \in K_{MG}}{|G_{MG}|} \quad (8$$

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 7 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 2) on the expected number of sequences was tested for four hypothetical communities. The first community had an even structure such that all the metagenomic DNA segments were equally distributed across all K_{MG}. In the second community, 90% of the metagenomic DNA segments were equally distributed in 50% of K_{MG}, and the remaining 10% of the metagenomic DNA segments 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 metagenomic DNA segments were equally distributed across 10% of K_{MG}, and the remaining 10% of the metagenomic DNA segments 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, or octaves (i.e., s was the same in all groups). The abundance of the metagenomic DNA segments in each group followed a lognormal distribution which has been observed in true microbial populations (e.g., (7, 16)). The functional form for modeling abundances was based on the functional form of a lognormal community (34):

$$S(R) = S_0 e^{-a^2 R^2} \quad (9$$

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 octave. For the lognormal abundance distribution in Fig 2D, a was set to a value of 0.2. Each hypothetical community started with a unique number of k-sized reads $|K_{MG}| = 1 \times 10^2$.

$|K_{MG}|$ was incrementally increased at 10 equally-spaced, linear steps to a maximum of $|K_{MG}| = 1 \times 10^6$. As $|K_{MG}|$ increased, all community structures remained constant. Graphical representation of rank abundance in Fig 2a 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

$$r_n = \frac{r}{s} \quad (10$$

where r and s are untransformed rank abundance and richness, respectively. Thus, the most abundant k-mer in a in a metagenome population has a normalized rank abundance of $1/s$ and the least abundant has a normalized rank abundance of 1. For each community, at each step, the expected number of sequences was calculated using equation 7. The expected number of sequences as a function of $|K_{MG}|$ were modeled with linear regressions.

Equation 7 gives the expected number of sequences required to sequence any sized community to exhaustion. Numerical sequencing simulations were performed to determine the number of sequences necessary to sequence a subset of all unique DNA (K_{MG}). These numerical sequencing simulations were applied to four hypothetical community structures described above. Numerical simulations were performed such that $|K_{MG}| = 3 \times 10^7, 4 \times 10^7, 5 \times 10^7, 7 \times 10^7, 9 \times 10^7$, and 1×10^8. During each of these simulations, the parameters read length (k) and average genome size (l) were set to 100 and 1×10^6, respectively, for all g. Random elements from K_{MG} were selected with replacement to simulate a sequencing event. Numerical simulations were performed until the fraction 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 fraction 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 depth by performing numerical sequencing simulations on 6 different lognormally-distributed communities. The numerical sequencing simulations were similar to the simulations described above. The 6 lognormal communities were modeled such that each community had S_0 =1, 10 equally-sized octaves, and $|K_{MG}| = 1 \times 10^7$. The difference between the 6 lognormal distributions was due to variations in a where a=0, a=0.005, a=0.008, a=0.01, a=0.015, and a=0.02. Evenness was represented using Pielou evenness index (9), which is the ratio of the Shannon diversity index (35) for a given community to that of an even community of the same richness. Shannon diversity was calculated in the context of a metagenomes such that:

$$H_{MG} = \sum_{j \in K_{MG}} -p_j log(p_j) \quad (11$$

where p_j is the proportion that the j^{th} k-sized read represents among all unique DNA sequences in the metagenome. Thus, the Pielou evenness index (9) was calculated such that:

$$J = \frac{H_{MG}{}'}{H_{MG,max}} \quad (12$$

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

Lastly, numerical simulations were performed to determine the sequencing depth necessary to achieve a target fraction for an individual metagenome (g_{MG}). Target fractions were 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 (l) varied such that l=0.5×10^6, l=1×10^6, l=2×10^6, l=3×10^6, l=5×10^6, l=10×10^6, l=15×10^6, and l=20×10^6. The sequencing depth for a given combination of target fraction, genome size, and fraction of the metagenome community was modeled using the gam function (mgcv R package; (36)). For modeling

purposes, target fraction was raised to the 12^{th} power and both genome size and sequences were log-transformed. The number of smooth dimensions for fraction of community, genome size, and target fraction were heuristically varied till the resulting fit demonstrated residuals with a normal distribution. Note that the objective here was not build a predictive model but simply a first order approximation for simulations performed here.

Data Availability

All simulations and codes used for modeling sequencing depth are freely available on Github at: https://github.com/taylorroyalty/sequence_simulation_code.

Acknowledgements

This research was supported by the National Science Foundation and a C-DEBI subaward (contribution number to be determined).

References

1. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik A V, 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.
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 903:1–10.
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.
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.
5. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844.
6. Zumel N, Mount J. 2014. Practical Data Science with R, 1st ed. Manning Publications Co., Greenwich, CT, USA.
7. Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. 2009. Ecology of the rare microbial biosphere of the Arctic Ocean. Proc Natl Acad Sci 106:22427–22432.
8. Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proc Natl Acad Sci 113:5970–5975.
9. Pielou EC. 1966. The measurement of diversity in different types of biological collections. J Theor Biol 13:131–144.
10. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G,

443 Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome
444 sequencing. Funct Integr Genomics 15:141–161.
445 11. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F,
446 Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft
447 ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA,
448 Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema
449 TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon
450 KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus
451 A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T.
452 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-
453 assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731.
454 12. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the
455 quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome
456 Res 25:1043–55.
457 13. Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, Daroub SH, Camargo
458 FAO, Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial
459 diversity. ISME J 1:283–290.
460 14. Feng BW, Li XR, Wang JH, Hu ZY, Meng H, Xiang LY, Quan ZX. 2009. Bacterial diversity of
461 water and sediment in the Changjiang estuary and coastal area of the East China Sea. FEMS
462 Microbiol Ecol 70:236–248.
463 15. Rintala A, Pietilä S, Munukka E, Eerola E, Pursiheimo JP, Laiho A, Pekkala S, Huovinen P.
464 2017. Gut microbiota analysis results are highly dependent on the 16s rRNA gene target region,
465 whereas the impact of DNA extraction is minor. J Biomol Tech 28:19–30.

16. 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.

17. 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.

18. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: Key considerations in genomic analyses. Nat Rev Genet 15:121–132.

19. Leinonen R, Sugawara H, Shumway M. 2010. The Sequence Read Archive. Nucleic Acids Res 39:2010–2012.

20. Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864.

21. 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 9:7–11.

22. 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 111:4904–4909.

23. Schirmer M, Smeekens SP, Vlamakis H, Jaeger M, Oosting M, Franzosa EA, 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.

24. Graham ED, Heidelberg JF, Tully BJ. 2017. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. PeerJ 5:e3035.

25. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–20.

26. Li D, Luo R, Liu C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102:3–11.

27. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.

28. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–9.

29. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

30. Zwietering MH, Jongenburger I, Rombouts FM, Van't Riet K. 1990. Modeling of the bacterial growth curve. Appl Environ Microbiol 56:1875–1881.

31. Rosselló-Móra R, Amann R. 2015. Past and future species definitions for Bacteria and Archaea. Syst Appl Microbiol 38:209–216.

32. 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 S V., 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, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi MCE, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo K V., Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov D V., Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

33. Flajolet P, Gardy D, Thimonier L. 1992. Birthday paradox, coupon collectors, caching

algorithms and self-organizing search. Discret Appl Math 39:207–229.

34. Magurran AE. 1988. Ecological Diversity and Its Measurement, 1st ed. Croom Helm Ltd.

35. Shannon CE. 1948. A mathematical theory of communication. Bell Syst Tech J 27:379–423.

36. Wood S. 2017. mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation.

CRAN https://cran.r-project.org/package=mgcv Retrieved 14 May 2018

Tables

Table 1. Estimates of fit coefficients for the Gompertz equation (equation 1) for the effective MAGs as a function of sequencing depth in published datasets from ocean surface water, estuarine sediment, maize soil, and the human gut. *p* values for all coefficients were <<0.05.

Environment	A (±SE)	μ (±SE)	λ (±SE)	MAG Yield*
Ocean Surface Water	97.67 (4.15)	5.84 (0.31)	1.16 (0.33)	0.88
Estuary Sediment	26.25 (2.11)	1.63 (0.06)	3.70 (0.24)	0.86
Maize Soil	43.65 (1.98)	1.43 (0.07)	6.13 (0.60)	0.70
Human Gut	17.49 (1.02)	5.01 (0.46)	0.67 (0.14)	0.90

*Calculated as the ratio of the maximum effective MAGs experimentally observed to maximum effective MAGs (A)

547 Table 2. Summary of sequence datasets analyzed with the MAG pipeline.

Environment	NCBI SRA Accession	Sequencing Platform	Total Reads*	High Quality Bases*	General Notes	Citation
Ocean Surface Water	ERR599029	Illumina HiSeq 2000	337,228,196	33,396,930,215	Caribbean Sea (5 mbsl)	(21)
Estuary Sediment	SRR5248164	Illumina HiSeq 2000	113,025,112	15,887,161,501	Sulfate Zone (8-10 cmbsf)	(4)
Maize Soil	SRR351473	Illumina HiSeq 2000	472,686,494	38,246,948,858	Surface Soil	(22)
Human Gut	SRR5127631	Illumina HiSeq 2000	50,951,710	4,846,948,241	--	(23)

548 *Combination of forward and backwards pair-end reads.

549 **Figures**

550 **Fig 1.** The influence that the parameters A, μ, and λ had on the Gompertz equation (A). The property of
551 the Gompertz equation that each parameter influences is colored red. Mean MAG completeness (B),
552 and mean MAG contamination as a function of simulated sequencing depth (Gb) for sequence datasets
553 of the human gut, maize soil, estuarian sediment, and surface ocean microbiomes, using the pipeline
554 described in the methods section. Translucent lines in (A) correspond to nonlinear least squares fits of
555 the Gompertz equation to the respective environmental dataset.

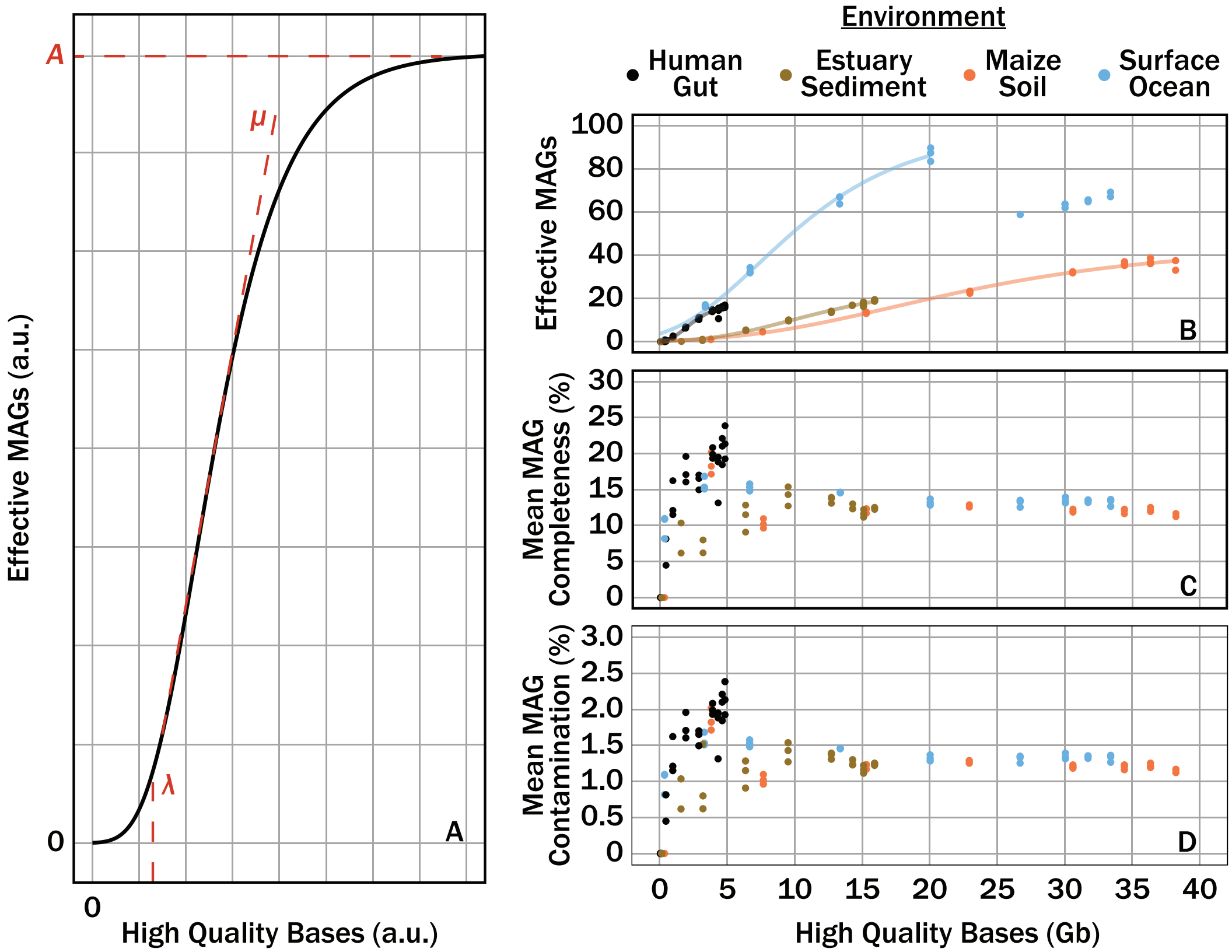
556 **Fig 2**. Average expected sequences required to fully sequence four different community structures, one
557 with relatively high community evenness (A), relatively moderate community evenness (B), relatively
558 low community evenness (C), and one with a lognormal community structure (D), were predicted using
559 linear regressions (E) and the log of $|K_{MG}|$ from equation 6 as a predictor.

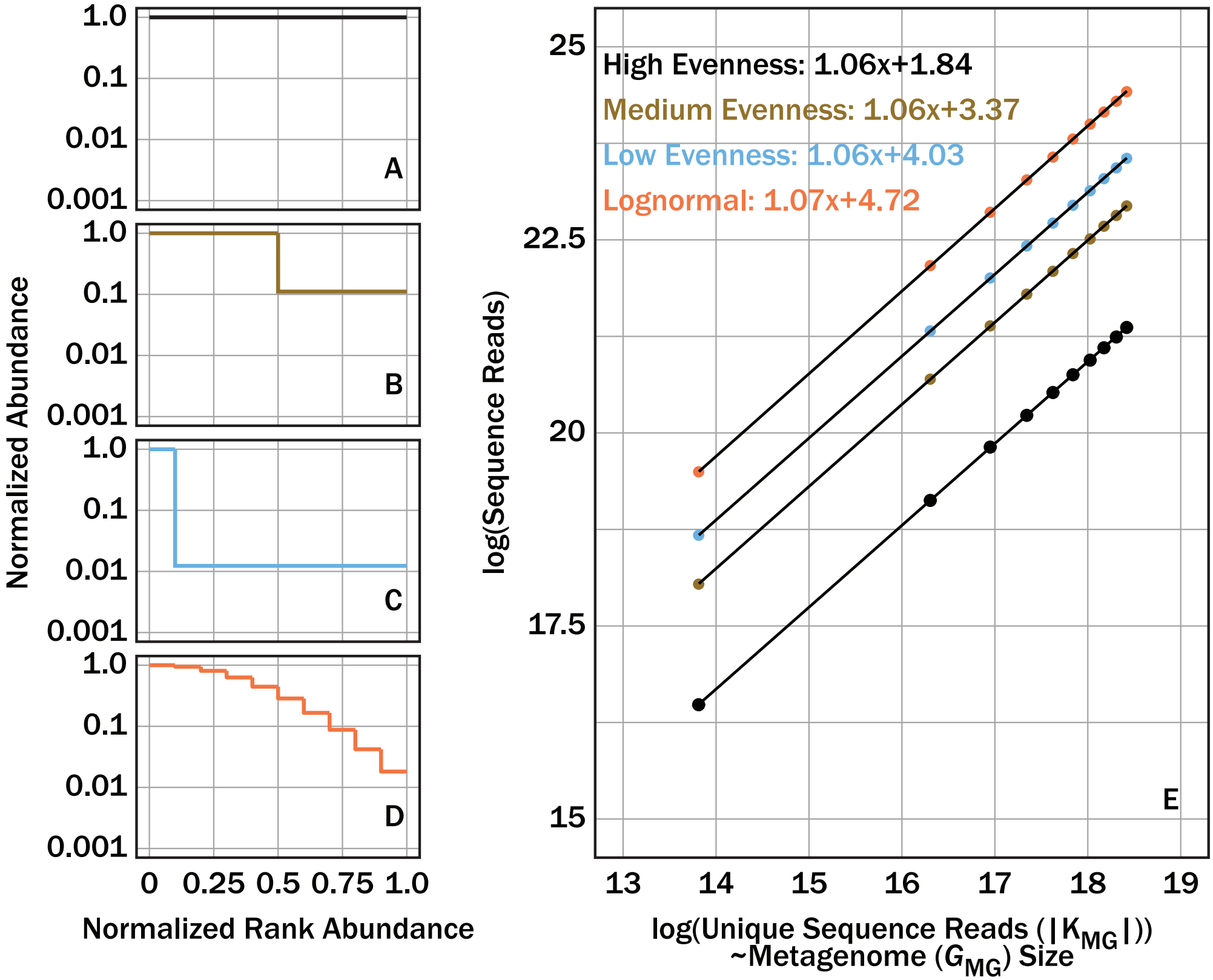
560 **Fig 3**. Sequences necessary to reach variable target sequencing depths (colors) for four different
561 community structures, one with relatively high community evenness (A), relatively moderate
562 community evenness (B), relatively low community evenness (C), and one with a lognormal
563 community structure (D). Red translucent lines correspond with linear regression curves for the
564 respective community in Fig 2E.

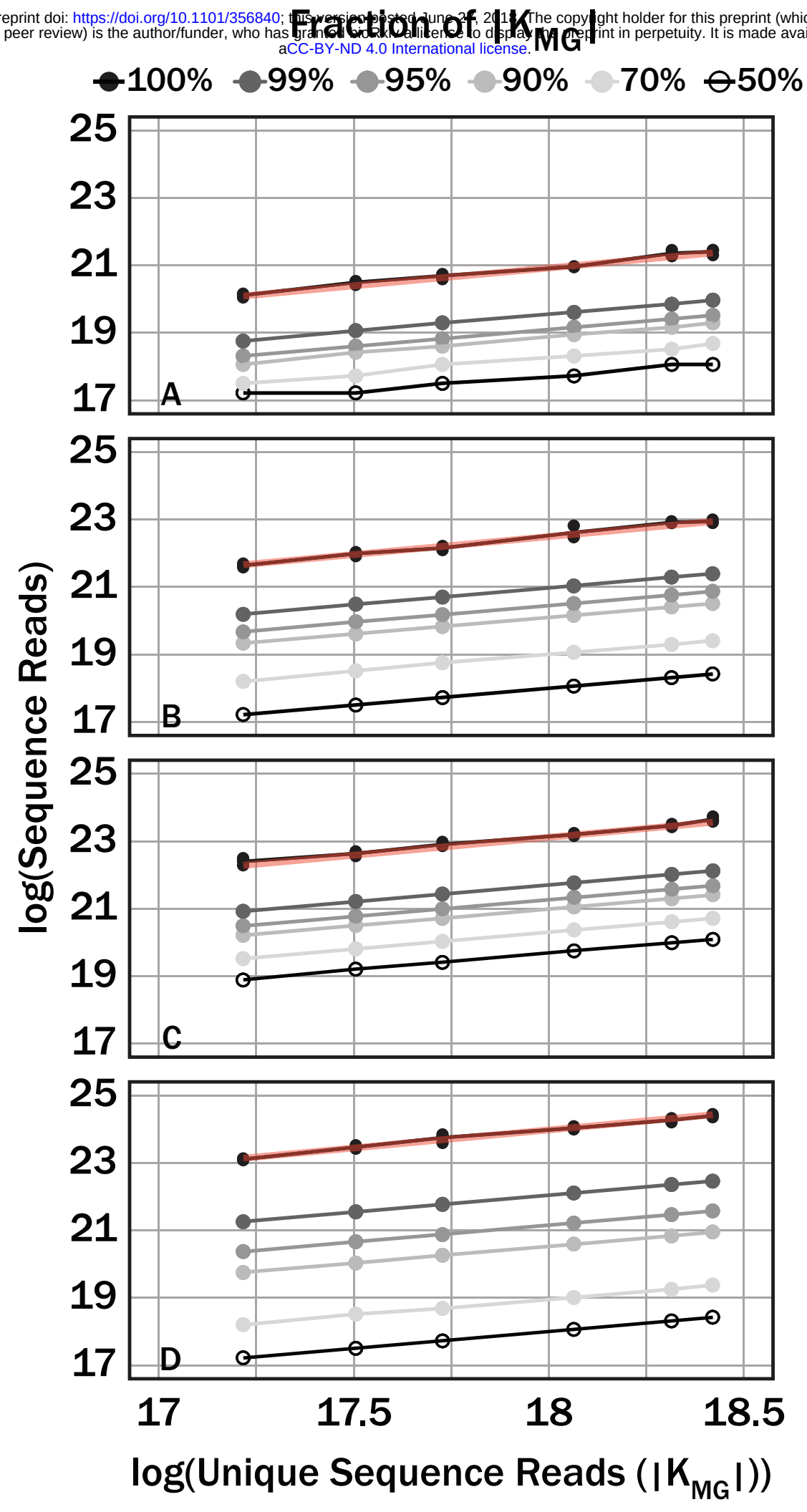
565 **Fig 4.** Numerical sequencing simulations applied to 6 hypothetical communities with different
566 lognormal distributions that were defined by the parameter, *a*, from equation 9 (A). The number of
567 sequences necessary to sequence a target fraction of $|K_{MG}|$ (dashed contours) as a function of the
568 Pielou evenness index, *J*, for a given lognormal community structure (B).

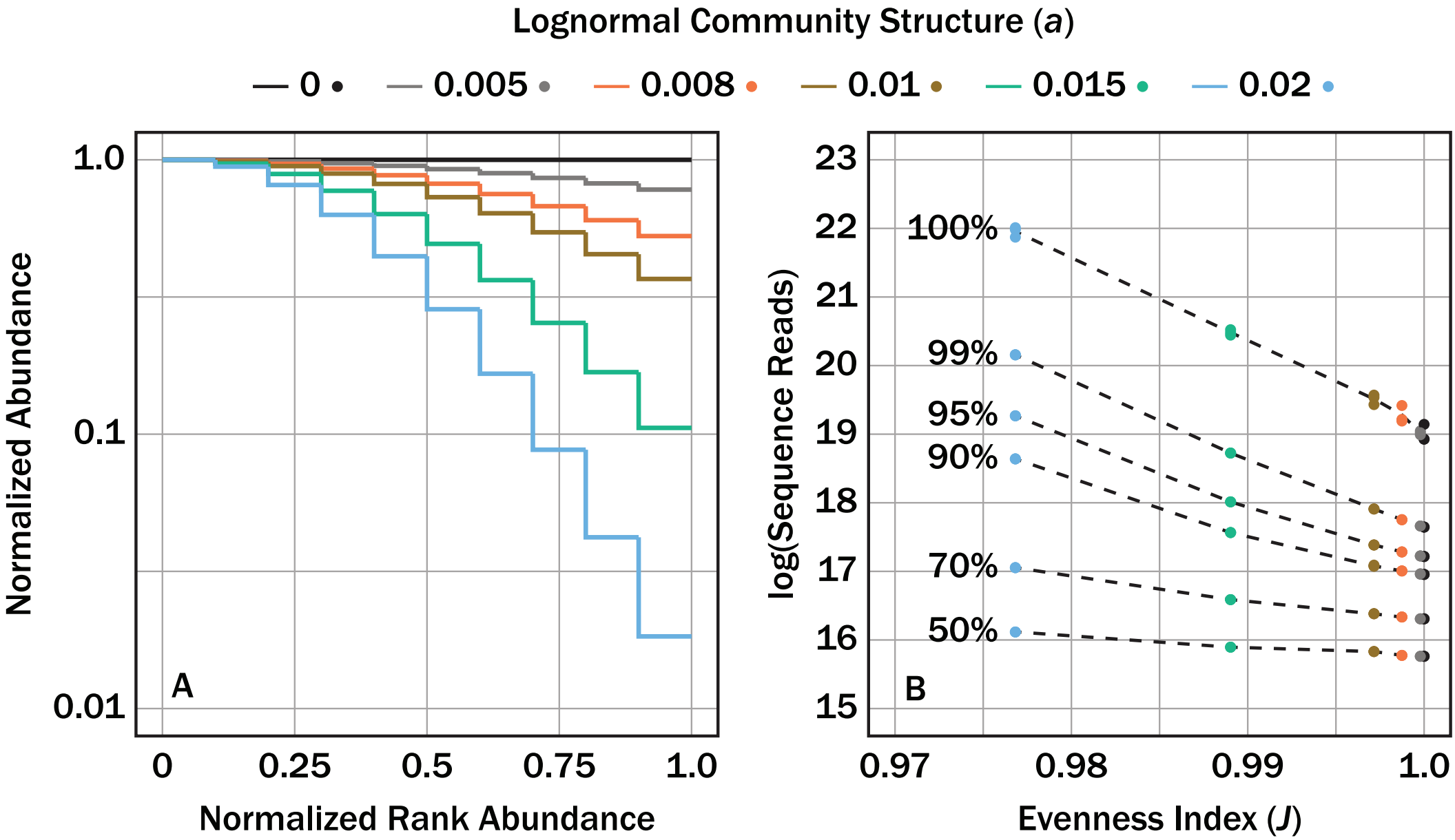
569 **Fig 5.** Numerical sequencing simulations show the number of bases (color bar) required to sequence a
570 target fraction of a genome which represents a given fraction of a community metagenome. Genomes
571 evaluated were 0.5×10^6 (A), 2×10^6 (B), 5×10^6 (C), 10×10^6 (D), and 20×10^6 (E) base pairs long.

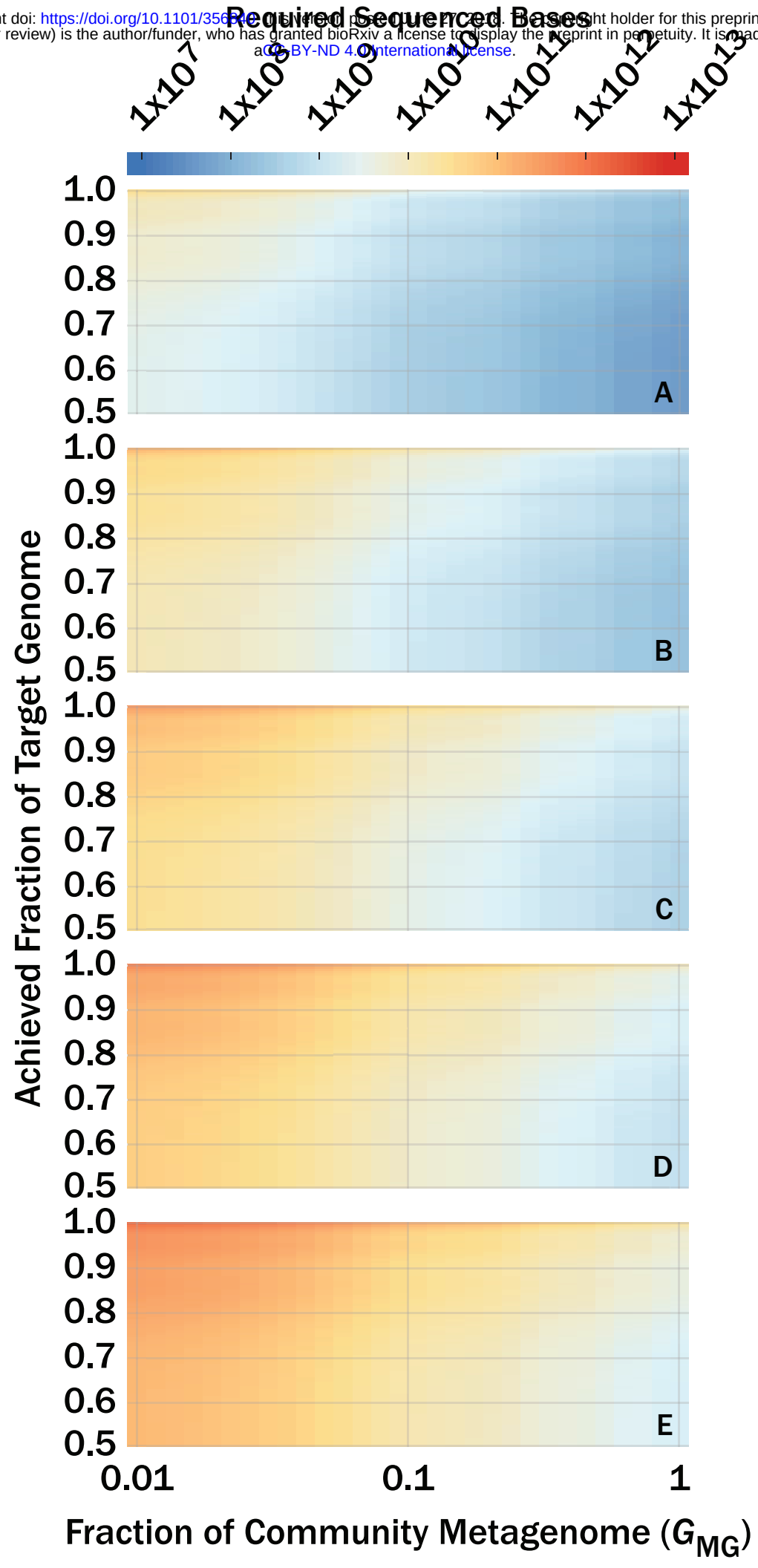
572 **Fig 6.** A cartoon illustrating an example microbial community (G), metagenomes for genomes ($g_{MG,i}$)
573 within G, and the overall metagenome for the given microbial community (G_{MG}). In this example, there
574 are 6 MAGs (s=6) and a total of 13 microbes. (A) Black circles represent individual microbes whose
575 genomes are averaged together, g. The average genome, g, are indicated by different color inner-
576 circles. (B) Individual average genomes can be sequenced at K unique positions depending on the
577 characteristic read length, k, of a sequencer. (C) All unique positions that can be sequenced for a given
578 genome, g, defines the metagenome, g_{MG}, for the i^{th} genome, g_i. (D) Replacing all individual genomes
579 in (A) with metagenomes, g_{MG}, gives the metagenome of the microbial community, G_{MG}.
580
581

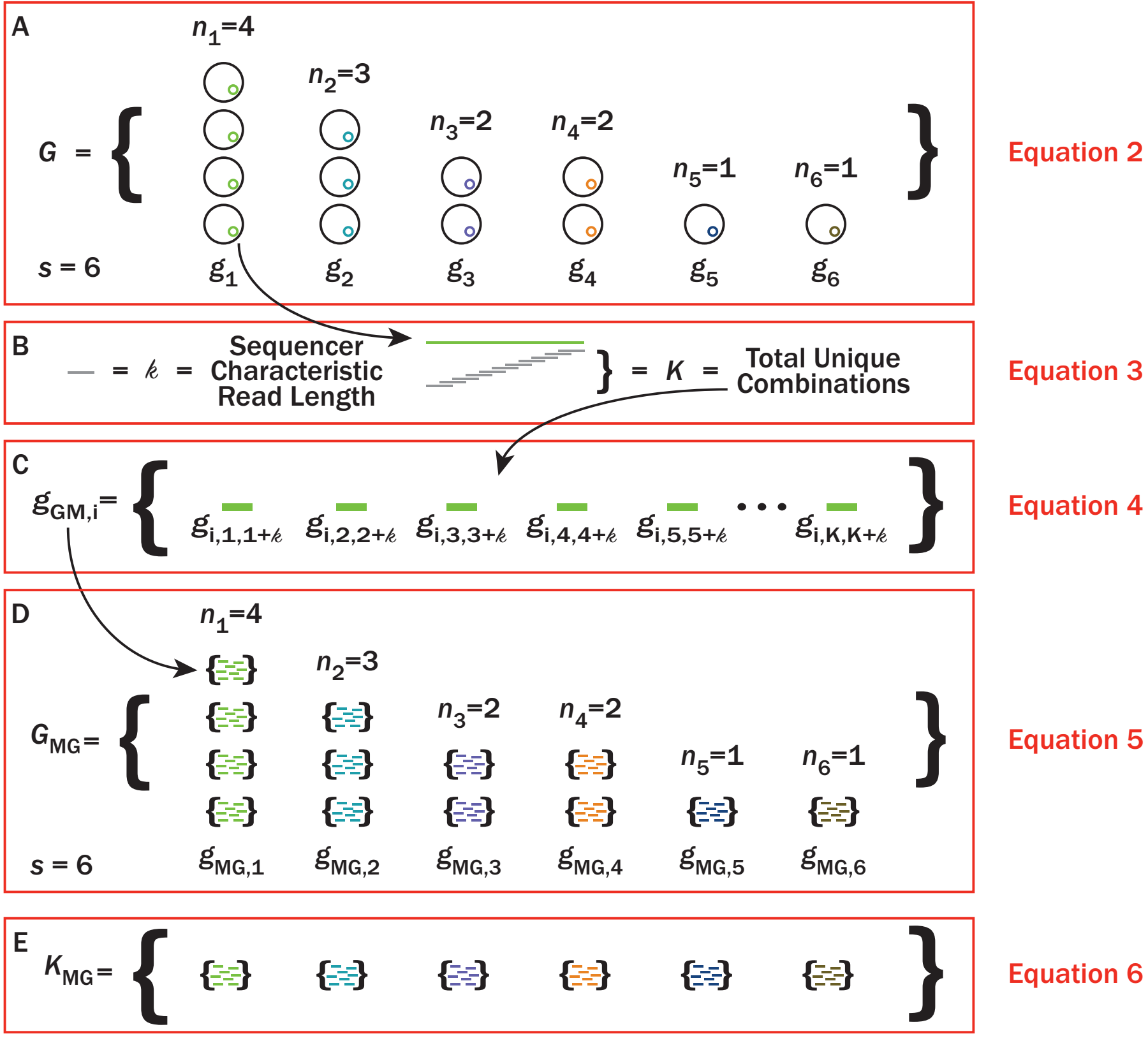

A
μ
λ
0
A
Effective MAGs (a.u.)
0
High Quality Bases (a.u.)
Environment
Human Gut
Estuary Sediment
Maize Soil
Surface Ocean
Effective MAGs
100
80
60
40
20
0
B
Mean MAG Completeness (%)
30
25
20
15
10
5
0
C
Mean MAG Contamination (%)
3.0
2.5
2.0
1.5
1.0
0.5
0
D
0
5
10
15
20
25
30
35
40
High Quality Bases (Gb)


1.0
0.1
0.01
0.001
A
B
C
D
Normalized Abundance
0
0.25
0.5
0.75
1.0
Normalized Rank Abundance
25
22.5
20
17.5
15
High Evenness: 1.06x+1.84
Medium Evenness: 1.06x+3.37
Low Evenness: 1.06x+4.03
Lognormal: 1.07x+4.72
log(Sequence Reads)
E
13
14
15
16
17
18
19
log(Unique Sequence Reads (|K_MG|))
~Metagenome (G_MG) Size




Lognormal Community Structure (a)
0
0.005
0.008
0.01
0.015
0.02
Normalized Abundance
1.0
0.1
0.01
A
0
0.25
0.5
0.75
1.0
Normalized Rank Abundance
log(Sequence Reads)
23
22
21
20
19
18
17
16
15
100%
99%
95%
90%
70%
50%
B
0.97
0.98
0.99
1.0
Evenness Index (J)




A
n_1=4
n_2=3
n_3=2
n_4=2
n_5=1
n_6=1
G =
s = 6
g_1
g_2
g_3
g_4
g_5
g_6
Equation 2
B
= k = Sequencer Characteristic Read Length
= K = Total Unique Combinations
Equation 3
C
$g_{GM,i}$=
$g_{i,1,1+k}$
$g_{i,2,2+k}$
$g_{i,3,3+k}$
$g_{i,4,4+k}$
$g_{i,5,5+k}$
$g_{i,K,K+k}$
Equation 4
D
n_1=4
n_2=3
n_3=2
n_4=2
n_5=1
n_6=1
G_{MG}=
s = 6
$g_{MG,1}$
$g_{MG,2}$
$g_{MG,3}$
$g_{MG,4}$
$g_{MG,5}$
$g_{MG,6}$
Equation 5
E
K_{MG}=
Equation 6