Simulation-based approaches to characterize the effect of sequencing effort 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: 250

Importance Word Count: 92

Text Word Count: 5365 words

Abstract

We applied simulation-based approaches to characterize how sequencing effort influences the properties of genomes identified in metagenomes assembled from short read sequences. An initial analysis evaluated the quantity, completion, and contamination of complete-genome equivalents, a bioinformatic-pipeline normalized metric for metagenome-assembled genomes (MAGs), as a function of sequencing effort on four preexisting sequence read datasets taken from a maize soil, an estuarine sediment, the surface ocean, and the human gut. These datasets were variously subsampled 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 second analysis explored the theoretical relationship between sequencing effort and the proportion of available metagenomic DNA sequenced during a sequencing experiment as a function of community richness, evenness, and genome size. Simulations from this analysis demonstrated that while community richness and evenness influenced the amount of sequencing required to sequence a community metagenome to exhaustion, the effort necessary to sequence an individual genome to a target fraction of exhaustion was only dependent on the relative abundance and genome size. The software tool, GRASE, was created to assist investigators further explore this relationship. Re-evaluation of the relationship between sequencing effort and binning success in the context of the relative abundance of genomes, as opposed to base pairs, provides a framework to design 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

37 sequences can be assembled and binned into metagenome-assembled genomes, thus shedding light on
38 the function of microbial ecosystems that are important for health, agriculture, and Earth system
39 processes. The work presented here provides an analytical framework to select sequencing effort as a
40 function of experimental goals in metagenome-assembled genomes creation projects. We hope that the
41 results presented here, as well as GRASE will be valuable to researchers planning sequencing
42 experiments.

43 **Introduction**

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

52 Sequencing experimental workflows involve selecting a target environment, sequencing effort,
53 bioinformatic pipeline software and parameters, metabolism characterization software (i.e., for gene
54 identification and similarity searches) and databases (Fig 1). At present, there is little information to
55 guide how much sequencing is appropriate to achieve scientific goals in such experiments (5). This gap
56 in knowledge is partly attributed to the unknown structure of target microbial communities. Further
57 complications arise in that the accuracy and efficiency of bioinformatic pipelines are often difficult to
58 characterize, and thus obscure the relationship between sequencing effort and MAG retrieval. For the
59 year 2017, estimates compiled by Quince et al. (5) suggest that up till now, metagenomic shotgun
60 sequencing experiments usually sequence between 1 Gbp and 10 Gbp DNA nucleotides. Researchers
61 require more precise guidance to select an appropriate shotgun sequencing effort in order to maximize
62 information and minimize cost for their specific experimental question.

63 Illumina sequencing technology is currently the most popular platform to generate
64 metagenomic shotgun sequences (5). Previous investigators established theoretical relationships
65 between contig formation rate (6) and single genome coverage (7) as a function of short read

66 sequencing effort. On the community level, heuristic approaches have been proposed for evaluating
67 community-level coverage to increases in sequencing effort. For example, it has been proposed to
68 utilize short read redundancies as a function of sequencing effort to estimate community level coverage
69 (8). Without *a priori* knowledge of the microbial community structure, practical application of these
70 methods to estimate MAG retrieval as a function of sequencing effort is hindered. This is particularly
71 true as heuristic methods will only be sensitive towards highly abundant genomes.

72 Here we present two distinct analyses which constrain the relationship between the quantity of
73 Illumina metagenomic shotgun sequences and the community-level sequence coverage. First, we
74 performed *in silico* experiments simulating the effect of sequencing effort on retrieved MAG properties
75 for Illumina sequence datasets. For these experiments, we propose a new metric, complete-genome
76 equivalents, to reduce sensitivity to the unknown accuracy and efficiency of our bioinformatic pipeline
77 (Fig 1). Second, we applied a theoretical model and numerical simulations to estimate the minimum
78 sequencing effort needed to sequence a metagenome to a target fraction of exhaustion. Our theoretical
79 model is unique compared to previous models (6, 7) in that we characterize sequencing effort in the
80 context of all physically-distinct DNA fragments in the metagenome. This is important, as complete
81 genome coverage does not necessarily translate into accurate read reconstruction due to interspecies
82 homologous DNA and intraspecies microdiversity (9). Coupling results from the two analyses provides
83 a framework for investigators to define sequencing experiments in the context of selecting a rarity and
84 fraction of exhaustion for a desired target genome when sequencing a community. Overall, the patterns
85 presented here can be used to guide sequencing effort decisions for future sequencing project when
86 MAG creation is a primary goal.

87 **Results**

88 MAG RECONSTRUCTION AS A FUNCTION OF SEQUENCING EFFORT IN EXISTING METAGENOMIC DATASETS

89 The number of 'complete-genome equivalents' (equation 1) as a function of high-quality bases
90 empirically fit the Gompertz equation (equation 2; Fig 2B; parameters in Table 1). Complete-genome
91 equivalents represent the sum of MAGs retrieved, weighted by each MAG's checkM completeness
92 score (10). This metric is less sensitive to whether bioinformatic pipelines tend to lump contigs into
93 fewer, more-complete MAGs versus split them into fewer, less-complete MAGs.

94 For each environment, the number of complete-genome equivalents as a function of simulated
95 sequencing effort fit the Gompertz equation. Here, A, μ, and λ correspond to the maximum number of
96 complete-genome equivalents assembled with the pipeline, the maximum rate which complete-genome
97 equivalents form as with more sequencing, and the "lag bases," or the bases which must be sequenced
98 prior to a sufficient number of sequence reads existing to generate overlap and form contigs (6). The
99 predicted maximum complete-genome equivalents varied from ~17 in the gut community to ~97 in the
100 surface ocean, the predicted maximum rate that complete-genome equivalents increased varied from
101 ~1.4 to ~5.8, and the minimum threshold of sequencing necessary prior to seeing complete-genome
102 equivalents varied from ~0.6 to ~6.7. The *Tara Oceans* dataset, where complete-genome equivalents
103 decreased at sequencing effort >20 Gbp, was an exception. For the estuary, maize, and human gut
104 datasets, the complete-genome equivalents yield began to asymptote with increasing sequencing
105 efforts. The *Tara Oceans* dataset followed a similar pattern at <25 Gbp. However, when the number of
106 sequenced bases was >25Gbp, the number of complete-genome equivalents decreased and became
107 insensitive to sequencing effort. The total number of MAGs retrieved was considerably higher than the
108 number of complete-genome equivalents.

109 Mean MAG completeness also increased towards an asymptote with increasing sequencing
110 effort (Fig 2C). Completeness was highest for the human gut dataset, with a maximum of 23.9%, and

111 increased continuously as sequencing effort increased. The mean MAG completeness reached an
112 asymptote of ~10-15% for the other three datasets with sequenced bases >10 Gbp. Note that when >10
113 Gbp were sequenced, the number of complete-genome equivalents created still increased as new
114 sequences were added. For all datasets, mean MAG contamination was <2% (Fig 2D) and did not
115 depend strongly on sequencing effort.

116 SIMULATION EXPERIMENTS

117 Using equation 8, we calculated the number of *k*-length sequence reads required to sequence all
118 physically-distinct DNA sequences of length, *k* (*k*-mers), in four hypothetical metagenomes, a perfectly
119 even, moderately uneven, highly uneven, and a lognormally-distributed structure (Fig 3A-D). Note that
120 equation 8 treats homologous DNA uniquely from one another. The expectation value of the log
121 number of sequences required to fully sequence metagenomes of the hypothetical communities was
122 linear with respect to log-transformed size of the metagenome (i.e., physically-distinct *k*-mers in the
123 population, or unique base pairs in a metagenome); this suggests a power-law relationship between
124 metagenome size and expectation value of sequence reads required to sequence the metagenome to
125 exhaustion (Fig 3E). For all community structures, the slope of the relationship between log-
126 transformed sequenced reads and log-transformed physically-distinct number of sequenced reads was
127 within 1% of 1.06. The structure of the population strongly influenced the number of reads required
128 such that more-even community structures required far fewer reads than less-even structures.

129 As equation 8 only estimates the number of reads to sequence a metagenome to exhaustion, we
130 used a numerical simulation to estimate the number of *k*-sized reads to sequence a metagenome to a
131 target fraction of exhaustion. The numerical simulation and equation 8 agreed on the expected
132 sequences reads to sequence 100% of a given metagenome (Fig 4); this supported the use of this
133 simulation. The log-transformation of both total physically-distinct *k*-sized reads ($|K_{\mathrm{MG}}|$ in equation 7)

134 and sequenced reads showed a linear response for all target fractions and all community structures. The
135 amount of sequences required to achieve a given target of physically-distinct *k*-sized reads was variable
136 for the different communities shown in Fig 3A. For instance, the lognormally-distributed community,
137 when compared to other communities, required the most sequencing effort to sequence a metagenome
138 to exhaustion but required similar sequencing effort to sequence a metagenome to a target fraction of
139 50%.

140 We applied the simulation to demonstrate the effect of community evenness on the number of
141 reads required to sequence a community to a target fraction of completion. These communities ranged
142 from perfectly even (a=0, equation 10) to more uneven (a=0.02, Fig 5A). Evenness was quantified
143 using the Pielou evenness index, which expresses Shannon diversity relative to the diversity of a
144 perfectly even community (11). The sequencing effort required to characterize genomes depended
145 strongly on both the evenness and the target fraction of completion (Fig 5B). Again, less-even
146 communities required more sequence reads than more-even communities. The strength of this
147 relationship also depended on the target fraction of completion. A community with Pielou evenness of
148 0.97 required 3 orders of magnitude more sequence reads to sequence a metagenome to a target
149 fraction of exhaustion than a perfectly even community while the same community only required about
150 42% more reads to sequence 50% of the metagenome.

151 The minimum sequencing effort required to sequence a microbe genome given a combination
152 of target fraction, genome size, and fraction of the metagenome community was modeled with a
153 generalized additive model (GAM; Fig 6). The GAM shows expected sequencing effort required for
154 microbial genome sizes 0.5, 2, 5, 10, and 20 Mbp, target genome completeness fractions from 0.5 to
155 1.0, and genome relative abundances from 1 to 0.0001. The smooth dimensions for target fraction,
156 genome size, and fraction of the metagenome community were 50, 6, and 29. To normalize for

157 different sequence read length, sequence reads were converted to bases and ranged from 1×10^{7} to
158 1×10^{13} total bases. More bases were required to sequence microorganisms 1) when the genome was
159 relatively rarer in the community, 2) to achieve better coverage of the genome, and 3) when average
160 genome sizes were larger.

161 Using the relationship shown in Fig 6, we can convert sequencing effort to an abundance cutoff.
162 Fig 7 shows complete-genome equivalents retrieved for a genome of a given level of abundance that is
163 sequenced to a target depth for the human gut, maize soil, estuarine sediment, and the surface ocean
164 microbiomes analyzed earlier. Correlation coefficients for the regressions used to relate log-
165 transformed sequencing effort (base pairs) to genome relative abundance were R=1 for all three
166 genome sizes evaluated (1 Mbp, 5 Mbp, and 20 Mbp). Evaluation of 1 Mbp and 20 Mbp define the
167 range of uncertainty in predicting complete-genome equivalents as the true size of genomes is
168 unknown. Unlike the asymptotic response of complete-genome equivalents to sequencing effort (Fig
169 2B), complete-genome equivalents increase exponentially as the abundance cutoff decreases (note that
170 the abundance cutoff is on a log-scale). The target genome completeness fraction was held at a constant
171 of 0.5 for all regressions and the sensitivity of complete-genome equivalents will change with respect
172 to the genome abundance cutoff with different values of target genome completeness fraction.

173 **Discussion**

174 We sought to establish evidence-based guidelines for selecting a sequencing effort during
175 shotgun metagenomic sequencing experiments. Random subsamples of existing short read datasets
176 (individually sampled, assembled, and binned) simulated the effect of creating MAGs from datasets of
177 different sizes and environments. The datasets analyzed here are representative of both the sequencing
178 effort (1 to 10 Gbp) (5) and the types of target environments that microbial ecologists often investigate
179 (12). A variety of software is available for all steps of MAG creation pipelines, and the quantity/quality

180 of MAGs will depend on software selection, software configuration, and sequenced environment (Fig
181 1; (5)). Furthermore, it is best-practice to manually curate algorithmically-created MAG bins (13). We
182 do not argue that the pipeline used here is objectively optimal for generating high-quality (high
183 completeness and low contamination) MAGs. Rather, our pipeline was configured to minimize
184 contamination (Fig 2D) associated with retrieved MAGs at the expense of reduced completeness. Thus,
185 MAG quantity was not directly reported but expressed as complete-genome equivalents (equation 1).
186 Reporting complete-genome equivalents reduces bias associated with binning software.

187 As sequencing effort increased, there was an initial lag in complete-genome equivalents
188 followed by a rapid increase in complete-genome equivalents, and then diminishing returns at higher
189 sequencing efforts. Previous investigators modeled the response of 16S RNA gene (14–16), Hill's
190 number diversity (17), taxon-resolved abundance (18), and gene abundance (18) as a function of
191 sequencing effort with rarefaction curves, or collector's curves. The complete-genome equivalents as a
192 function of sequencing effort did not match a traditional collector's curve, which lacks an initial lag.
193 Rather, the data appear sigmoidal. We modeled the data using the Gompertz function (equation 8),
194 because its parameters can be interpreted in terms of quantities that are familiar from microbial growth
195 curves (lag time, growth rate, and maximum density) (19). The three Gompertz parameters, A, μ, and λ
196 are interpreted as the maximum possible number of complete-genome equivalents at infinite
197 sequencing effort, maximum rate that complete-genome equivalents increased with increasing
198 sequencing effort, and a minimum sequencing threshold prior to rapid complete-genome equivalents
199 retrieval (Fig 2A), respectively. Consequently, the Gompertz equation fit parameters may provide
200 means for inter-comparisons between different metagenomic datasets.

201 The fits to the Gompertz function illustrate that there may be an optimal sequencing effort for
202 MAG creation efforts corresponding to the upper shoulder of the Gompertz curve ("late log phase").

203 When sequencing effort is too close to λ, MAGs bin poorly; when sequencing effort is too great, the
204 number and quality of MAGs per unit sequencing effort is low. We speculate that our choice of
205 pipeline, and specifically the fact that we discarded contigs <3kb, caused poor performance at higher
206 sequencing effort for the *Tara Oceans* dataset. Species-level microdiversity and interspecies
207 homologous DNA can cause "bubbles" which impair assembly in larger data sets (9, 20). Improved
208 assembly would likely have yielded more complete-genome equivalents for our assembly of the largest
209 subsets of the *Tara Oceans* data.

210 As mean MAG completeness converged to an asymptote well below 100% (Fig 2B), complete-
211 genome equivalent yields, or the ratio of the maximum observed complete-genome equivalents to the
212 theoretical maximum complete-genome equivalents (A) (Table 1) were close to 100%. The asymptote
213 in mean MAG completeness (Fig 2C) and negligible values of mean MAG contamination (Fig 2D)
214 demonstrates that the retrieved complete-genome equivalents at higher sequencing efforts represent
215 true signal opposed to false positive signal. If the complete-genome equivalents were false signal, then
216 either 1) the mean MAG completeness would decrease with the addition of false positive marker genes
217 which fail to cluster (i.e., more MAGs with low completeness) or 2) false positive signal would cluster
218 with preexisting MAGs leading to an increase in mean MAG contamination. Neither of these scenarios
219 occurred. More so, the fact that mean MAG completeness was at steady state suggests more
220 sequenceable DNA exists in these communities. Thus, we asked how much sequencing was necessary
221 to sequence a community to exhaustion.

222 The theoretical probability model (equation 8) demonstrated that the sequencing effort to
223 sequence a metagenome to exhaustion was predictable, regardless of community structure (Fig 3). One
224 limitation to equation 8 is predicting sequencing effort to sequence a metagenome to a fraction of
225 exhaustion. As such, a numerical simulation was used for predicting the sequencing effort necessary to

226 sequence an individual metagenome to a fraction of exhaustion. The numerical simulation agreed with
227 the theoretical model when predicting the sequencing effort to sequence a metagenome to exhaustion
228 (Fig 4). A further analysis explored the semi-quantitative relationship between community evenness
229 and richness and the necessary sequencing effort to achieve a target fraction of exhaustion for a
230 metagenome (Fig 5).

231 In practice, information about a target community structure may not be available for estimating
232 sequencing effort. The GAM built here predicts the minimum number of sequences necessary to
233 sequence a given fraction of a target genome as a function of average genome size and the relative
234 abundance of the target genome in the community (Fig 6). Even without knowledge of a target
235 community's structure, the GAM provides a useful constraint for designing whole-genome shotgun
236 metagenomic sequencing experiments.

237 Metagenomic shotgun sequencing experimental designs should be rationally designed such that
238 sequencing effort is selected to capture a desired fraction of a target microbial metagenome.
239 Investigators should be cognizant of the rarest microbial metagenome desired to be characterized as
240 well as the degree of characterization of that microbial metagenome when designing a sequencing
241 experiment. To that end, we have built a tool, `Genome Relative Abundance to`
242 `Sequencing Effort (GRASE)`, to report estimated sequencing effort required to capture a
243 defined fraction of a genome as a function of the relative abundance of the corresponding
244 microorganism in the community and average genome size. This R-based GUI app can be accessed
245 online at http://adsteen.shinyapps.io/grase and is archived at http://github.com/adsteen/GRASE, from
246 which it can be downloaded and run locally.

247 When the sequence read datasets analyzed here (Table 2; Fig 2) are re-evaluated in the context
248 of the relative abundance of a microbial metagenome (g_{MG}) sequenced to a target fraction (0.5),

complete-genome equivalents increase appreciably in response to minor increases in deeper characterization of the community metagenome (Fig 7). This observation contrasts the complete-genome equivalents response to sequencing effort (in base pairs), where substantial increases in sequencing effort (by contemporary standards) leads to diminishing returns in complete-genome equivalents. It is important to note that abundance cutoff and sequencing effort are interchangeable; however, complete-genome equivalents response to changes in the respective predictor (i.e., base pairs versus abundance cutoff) alters 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 <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

SEQUENCE DATA SOURCES

All sequence data were downloaded from NCBI's Sequence Read Archive (SRA) using the SRA Toolkit (fastq-dump –split-files) (21). Exact duplicate reads for both forward and reverse reads were removed using PRINSEQ (-derep 1; v0.20.4) (22). 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 (23). 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 (24). The last dataset was collected from human fecal samples and represented a human gut microbiome (25). All datasets analyzed in this study are summarized in Table 1.

272 MAG ASSEMBLY PIPELINE

273 The pipeline developed here followed similar pipelines described by other authors (3, 26). All
274 sequence datasets were analyzed as follows. Trimmomatic (v0.36) (27) removed adapters and trimmed
275 low-quality bases from individual reads. Read leading and trailing quality scores were required to be
276 >3. The sliding window was set to 4 base pairs and filtered base pair windows with a mean score <15.
277 Quality controlled reads were assembled into contigs using MEGAHIT (v1.1.2; --presets meta-large)
278 (28). Due to RAM limitations, assembled contigs <3000 bp long were excluded from the analysis.
279 Redundant contigs were removed using CD-HIT (v4.6.8; cd-hi-est -c 0.99 -n 10) (29). Remaining
280 contigs with high intra-contig similarity based on intra-contig sequence alignments (Minimus2 -D
281 OVERLAP=100 MINID=95) were removed. The quality-controlled reads (i.e., after using
282 Trimmomatic) were mapped to the remaining contigs using Bowtie 2 (v2.3.3) (30) to generate a
283 coverage score for individual contigs.

284 Resultant contigs were iteratively clustered into MAGs using the unsupervised, clustering
285 algorithm Binsanity (v0.2.6) (26). Similar to Tully et al. (3), six initial clustering iterations were
286 performed with the parameter, *preference* (-p), set to -10 (iteration 1), -5 (iteration 2), -3 (iteration 3-6).
287 Between iterations, a refinement step (Binsanity-refine) was performed on the putative MAGs with
288 constant *preference* (-p) of -25. The refined putative MAGs were evaluated for contamination and
289 completeness using the software CheckM (v1.0.6) (10), which uses HMMER (v3.1) and Prodigal
290 (v2.6.3) (31). Contigs associated with putative MAGs meeting one of the following criteria: 1) had a
291 completeness > 90% and contamination < 10%, 2) had a completeness > 80% and contamination < 5%,
292 or 3) had a completeness > 50% and contamination < 5% were treated as high-quality. All other MAGs
293 were considered low-quality MAGs. MAGs defined as high-quality were not modified any further.
294 Contigs associated with the high-quality MAGs were not used in the subsequent reclustering and

295 refinement steps. The contigs associated with low-quality MAGs were pooled together and reclustered
296 during the next iteration of Binsanity clustering. After the sixth iteration, the remaining MAGs which
297 did not fall into one of the three categories underwent additional refinement using Binsanity-refine.
298 During this step, MAGs were iteratively refined with *preference* set to -10 (iteration 1), -3 (iteration 2),
299 and -1 (iteration 3). Between each refinement step, metrics of contamination and completeness were
300 evaluated using CheckM. Again, MAGs meeting the criteria of a high-quality category described above
301 were not further modified. The respective contigs associated with putative MAGs were not used in
302 proceeding refinement steps. After the last iteration of refinement, all MAGs were reevaluated for
303 completeness, contamination, and a taxonomic rank using CheckM. Completeness values for MAGs
304 with a minimum resolved taxonomic rank of phylum were summed together. Complete-genome
305 equivalents, g_{eq}, were defined as:

306 $$g_{eq} = \sum_{i=1}^{n} f_i \quad (1$$

307 where n is the total number of retrieved MAGs and f_i is the corresponding CheckM completeness (%).

308 Common practice when reporting MAGs involves discarding MAGs not meeting a criterion of
309 completeness and contamination. In the context of characterizing MAG response to sequencing effort a
310 MAG pipelines often fails at generating high-quality MAGs. One must discern whether this is a failure
311 of the bioinformatic pipeline versus insufficient sequencing effort (Fig 1). We argue that MAGs with
312 low completeness still provide information regarding the community level coverage when generating
313 collector's curves. More so, a community has finite number of unique marker genes, and an infinite
314 sequencing effort would result in sequencing all unique marker genes, regardless of whether or not
315 these genes cluster to one another.

316 SUBSAMPLING SEQUENCE READ DATASETS

317 The effect of decreased sequencing effort was simulated by subsampling the initial sequence read

318 datasets described above. Downloaded sequence read datasets were randomly sampled at set fractions
319 of 1%, 10%, 20%, 40%, 60%, 80%, 90%, 95%, and 100%. Each fraction was resampled, assembled,
320 and binned in triplicate. All triplicates were analyzed using the MAG assembly pipeline described
321 above.

322 MODELING MAG RESPONSE TO SEQUENCING EFFORT

323 Complete-genome equivalents as a function of sequencing effort was modeled for environmental
324 sequence datasets using the Gompertz equation, as reformulated by Zweitering et al. (19) for use with
325 microbial growth curves:

326 $$g_{eq}(A, \mu, \lambda, b) = A \times e^{-e^{\frac{\mu \times e}{A}(\lambda - b)+1}} \quad (2$$

327 where A, μ, and λ are fit coefficients and b is high-quality bases (Gbp).

328 DEFINING THE MICROBIAL METAGENOME AND SEQUENCING PROBABILITY

329 Here we draw on set theory to provide a theoretical grounding for our *in silico* simulations described
330 below. The expected number of sequences to sequence a fraction of an individual microbe's genome
331 can be modeled with probability theory by defining a community metagenome with set theory. Fig 8A-
332 E provides a cartoon illustrating the application of this set theory on a hypothetical microbial
333 population, G. G contains unique genomes (g) with finite abundances (n). The definition of microbial
334 species is somewhat contentious (32). Here we define g as the average genome (length and
335 composition) for all individuals genomes associated with the rank, species, based on an arbitrary
336 definition. As such, the species richness (s or unique g) of G depends on the definition of g. In the
337 example community (Fig 8A-E), s=6 and n=13. Thus, G is represented as (Fig 8A):

338 $$G = \{n_1 g_1, n_2 g_2 \dots n_s g_s | n \in \mathbb{N}\} \quad (3$$

339 where s is the species richness. When characterizing G via shotgun metagenomics, the ith genome, g_i,
340 can be sequenced at K unique sections given a characteristic read length, k, and average genome size, l,

341 in number of base pairs (Fig 8B). Thus, the number of physically-distinct *k*-sized reads, *K*, associated

342 with the ith genome, g_i, within *G* is equal to:

343 $$K_{g_i} = l(g_i) - k + 1 \quad (4$$

344 Note that equation 4 considers homologous DNA as unique fragments. From equation 4, the

345 metagenome, g_{MG}, for g_i is defined as the set of all physically-distinct *k*-sized reads (Fig 8C) or:

346 $$g_{MG,i} = \{g_{i,1,1+k}, g_{i,2,2+k} \cdots, g_{i,Kg_i,Kg_i+k}\} \quad (5$$

347 where the subscripts for g_i represent a given *k*-sized read spanning from an arbitrary starting base pair

348 to the arbitrary starting base pair plus *k*. By substituting $g_{MG,i}$ into all *g* for equation 3 (Fig 8D), the

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

350 $$G_{MG} = \{n_1 g_{MG,1}, n_2 g_{MG,2} \ldots n_s g_{MG,s} | n \in \mathbb{N}\} \quad (6$$

351 while the population of physically-distinct *k*-sized reads in the metagenome, G_{MG} (Fig 8E), is

352 represented as:

353 $$K_{MG} = \{g_{MG,1}, g_{MG,2} \cdots g_{MG,s}\} \quad (7$$

354 From equation 5, one can determine the cardinality, or the total number, of physically-distinct

355 *k*-sized reads in associated with G_{MG} (expressed as $|K_{MG}|$). When attempting to fully sequence G_{MG}

356 using shotgun metagenomics, we assume that sampling events (sequence reads) are independent and

357 are sampled with replacement. In fact, Illumina sequencing technology sequences reads in parallel via

358 the individual DNA fragments binding to individual clusters. Furthermore, the fragmented DNA cannot

359 be sequenced twice as the sequencing process is destructive (33). Nonetheless, the mass of DNA

360 extracted from a target environment represents a negligible fraction of the total DNA existing in that

361 environment. As the relative abundance of the *k*-sized reads in K_{MG} does not change when DNA is

362 extracted from an environment, sampling events can be treated as independent and thus, DNA sampling

363 reduces to sampling with replacement. If the proportion DNA mass extracted had a significant impact

364 on the remaining mass of DNA in the environment, then one would be more suited to sequence all the
365 DNA versus a smaller proportion of the DNA. The sequencer should have no impact on sampling
366 assuming no sequencing errors due to misreading or spatial sampling issues (i.e., clonal density issues).
367 These issues do exist, but for the sake of a first order approximation, these biases can be ignored.

368 The probability of sequencing all elements in G_{MG} reduces to a coupon collectors problem (34)
369 by making the above assumptions. Using the general functional form for calculating expected samples
370 for sampling all unique elements in a set (equation 13b in (8)), one can predict the number of sequences
371 necessary to sequence all elements in K_{MG}, such that the expected number of sequences, $\mathrm{E}(G_{MG})$, is:

372 $$\mathrm{E}(G_{MG}) = \int_0^{\infty}\left(1 - \prod_{j \in K_{MG}}(1 - e^{-p_j t})\right)dt \quad (8$$

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

376 $$p_j = \frac{n_i \times j \in K_{MG}}{|G_{MG}|} \quad (9$$

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

379 MODELING EXPECTED SEQUENCES

380 Equation 8 provides an estimate for the total number of sequences to sequence all K_{MG}. The
381 influence of increasing species richness (i.e., s in equation 3) on the expected number of sequences was
382 tested for four hypothetical communities. The first community had an even structure such that all
383 physically-distinct DNA fragments were equally distributed across all K_{MG}. In the second community,
384 90% of the physically-distinct DNA fragments were equally distributed in 50% of K_{MG}, and the remaining
385 10% of the physically-distinct DNA fragments were distributed equally across the remaining 50% of
386 K_{MG}. This community represented a community with relatively moderate species evenness. In the third

387 community, 90% of the physically-distinct DNA fragments were equally distributed across 10% of K_{MG},
388 and the remaining 10% of the physically-distinct DNA fragments were distributed equally across the
389 remaining 90% of K_{MG}. This community represented a community with relatively low species evenness.
390 The last community had 10 equally-sized groups. The abundance of the physically-distinct DNA
391 fragments in each group was based on the function form of a lognormal community (35) which has been
392 observed in microbial populations (e.g., (17, 36)), such that:

393 $$S(R) = S_0 e^{-a^2 R^2} \quad (10$$

394 where S_0 was treated as the maximum relative of abundance (S_0 = 1), a was the inverse width of the
395 distribution, R was treated as the positive octave range spanning 0 to 9, and $S(R)$ represented the
396 abundance for a given group. For the lognormal abundance distribution in Fig 3D, a was set to a value
397 of 0.2. Each hypothetical community started with $|K_{MG}| = 1 \times 10^2$. $|K_{MG}|$ incrementally increased at 10
398 equally-spaced, linear steps to a maximum of $|K_{MG}| = 1 \times 10^6$. As $|K_{MG}|$ increased, all community
399 structures remained constant. Graphical representation of rank abundance in Fig 3A was normalized by
400 a given $|K_{MG}|$ to reflect that populations retained the same structure even as population size varied. We
401 defined a normalized rank abundance r_n such that

402 $$r_n = \frac{r}{s} \quad (11$$

403 where r and s are untransformed rank abundance and richness, respectively. For each community, at
404 each step, the expected number of sequences was calculated using equation 8. The expected number of
405 sequences as a function of $|K_{MG}|$ were modeled with linear regressions.

406 Equation 8 gives the expected number of sequences required to sequence any sized community
407 to exhaustion. Numerical sequencing simulations were performed to determine the number of
408 sequences necessary to sequence a subset of all physically-distinct DNA (K_{MG}). These numerical
409 sequencing simulations were applied to four hypothetical community structures described above.

410 Numerical simulations were performed such that $|K_{MG}| = 3 \times 10^7,\ 4 \times 10^7, 5 \times 10^7, 7 \times 10^7, 9 \times$

411 10^7, and 1×10^8. During each of these simulations, the parameters read length (k) and average

412 genome size (l) were set to 100 and 1×10^6, respectively, for all g. Random elements from K_{MG} were

413 selected with replacement to simulate sequencing events. Numerical simulations were performed until

414 the percentage of $|K_{MG}|$ sequenced was 50%, 70%, 90%, 95%, 99%, or 100%. A weight distribution

415 was applied to elements in a given K_{MG}. The weight distribution biased sequencing to reflect the

416 relative abundances of the four hypothetical communities described above. The percentage of $|K_{MG}|$

417 sequenced was evaluated every 1×10^7 sequences. Numerical simulations were performed in triplicate

418 for all $|K_{MG}|$ and all target fractions of $|K_{MG}|$.

419 We explored the influence of community evenness on required sequencing effort by performing

420 sequencing simulations on 6 different lognormally-distributed communities. The numerical sequencing

421 simulations followed the simulations described above. The 6 lognormal communities were modeled

422 such that each community had S_0 =1, R=10, and $|K_{MG}| = 1 \times 10^7$. Values of a for the 6 lognormal

423 distributions were a=0, a=0.005, a=0.008, a=0.01, a=0.015, and a=0.02. Evenness was represented

424 using the Pielou evenness index (11), or the ratio of the Shannon diversity index (37) for an observed

425 community to an even community of equal richness. Shannon diversity was calculated in the context of

426 a metagenomes such that:

427 $$H_{MG} = \sum_{j \in K_{MG}} -p_j log(p_j) \quad (12$$

428 where p_j is the proportion that the jth k-sized read represents among all physically-distinct DNA sequences

429 in the metagenome. Thus, the Pielou evenness index (11) was calculated such that:

430 $$J = \frac{H_{MG}{}'}{H_{MG,max}} \quad (13$$

431 where J was the Pielou evenness index, $H_{MG}{}'$ was the metagenome Shannon diversity index, and

432 $H_{MG,max}$ represented the metagenome Shannon diversity index when all p_j were equal (a=0).

433 Lastly, numerical simulations were performed to determine the sequencing effort necessary to
434 achieve a target fraction for an individual metagenome (g_{MG}). Target fractions increased from 0.5 to 1
435 at 100 linearly-spaced intervals. The fraction of the metagenome community (G_{MG}) that g_{MG}
436 represented varied from 1% to 100% in 30 lognormally-spaced intervals. The target genome sizes
437 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.
438 The sequencing effort for a given combination of target fraction, genome size, and fraction of the
439 metagenome community was modeled using the gam function (mgcv R package; (38)). For modeling
440 purposes, target fraction was raised to the 6th power and both genome size and sequences were log-
441 transformed. The number of smooth dimensions for fraction of community, genome size, and target
442 fraction were heuristically varied till the resulting fit demonstrated residuals with a normal distribution.
443 Note that the objective here was not build a predictive model but simply a first order approximation for
444 simulations performed here. The generalized additive model was extrapolated for genome relative
445 abundances spanning 0.01 to 0.0001 at all sizes and target fractions. The extrapolation was necessary
446 due to the computational limitation of analyzing genome relative abundances <0.01.

447 RELATING MAG RESPONSE TO THE THEORETICAL SEQUENCING MODEL

448 Sequencing effort (in base pairs) used for predicting complete-genome equivalents for the sequence
449 read datasets maize soil, an estuarine sediment, the surface ocean, and the human gut, were related to a
450 genome relative abundance utilizing the generalized additive model presented in the previous section.
451 This was accomplished by setting a constant genome size and target fraction (0.5) and performing a
452 linear regression between genome relative abundance (a genome's fraction of a community) and log-
453 transformed base pairs. The linear regression was performed with genome sizes of 1 Mbp, 5 Mbp, and
454 20 Mbp. Complete-genome equivalents were predicted as a function of genome relative abundance
455 sequenced to a target fraction of 0.5.

Data Availability

All simulations and codes used for modeling sequencing effort are freely available on Github at: https://github.com/taylorroyalty/sequence_simulation_code. The GRASE GUI application is available at: http://shinyapps.io/adsteen/GRASE.

Acknowledgements

This research was supported by National Science Foundation grant OCE-1431598 and a C-DEBI subaward. This is C-DEBI contribution number (contribution number to be determined).

463 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. Lander ES, Waterman MS. 1988. Genomic mapping by end-characterized random clones: a mathematical analysis. Genomics 231–239.
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.
8. Rodriguez-R LM, Konstantinidis KT. 2014. Estimating coverage in metagenomic data sets and why it matters. ISME J 8:2349–2351.
9. Ghurye JS, Cepeda-Espinoza V, Pop M. 2016. Metagenomic assembly: Overview, challenges and applications. Yale J Biol Med 89:353–362.
10. 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 25:1043–55.

11. Pielou EC. 1966. The measurement of diversity in different types of biological collections. J Theor Biol 13:131–144.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

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

20. 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.

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

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

23. 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.

24. 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.

25. 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.

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

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

28. 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.

29. 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.

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

31. 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.

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

33. 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.

34. Flajolet P, Gardy D, Thimonier L. 1992. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. Discret Appl Math 39:207–229.

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

36. 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.

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

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

Tables

Table 1. Estimates of fit coefficients for the Gompertz equation (equation 2) for the complete-genome equivalents 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)	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 complete-genome equivalents experimentally observed to maximum complete-genome equivalents (A)

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	3.372×10^{8}	3.340×10^{10}	Caribbean Sea (5 mbsl)	(23)
Estuary Sediment	SRR5248164	Illumina HiSeq 2000	1.137×10^{8}	1.589×10^{10}	Sulfate Zone (8-10 cmbsf)	(4)
Maize Soil	SRR351473	Illumina HiSeq 2000	4.727×10^{8}	3.824×10^{10}	Surface Soil	(24)
Human Gut	SRR5127631	Illumina HiSeq 2000	5.095×10^{7}	4.847×10^{9}	--	(25)

* Combined forward and reverse paired-end reads.

Figures

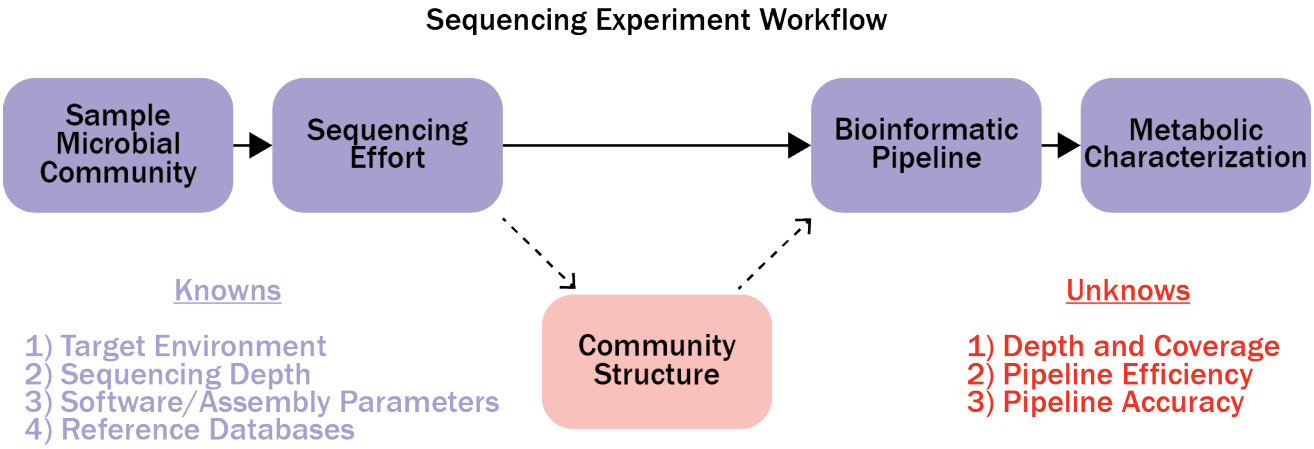


Fig 1. A flow diagram illustrating the workflow for sequencing experiments.

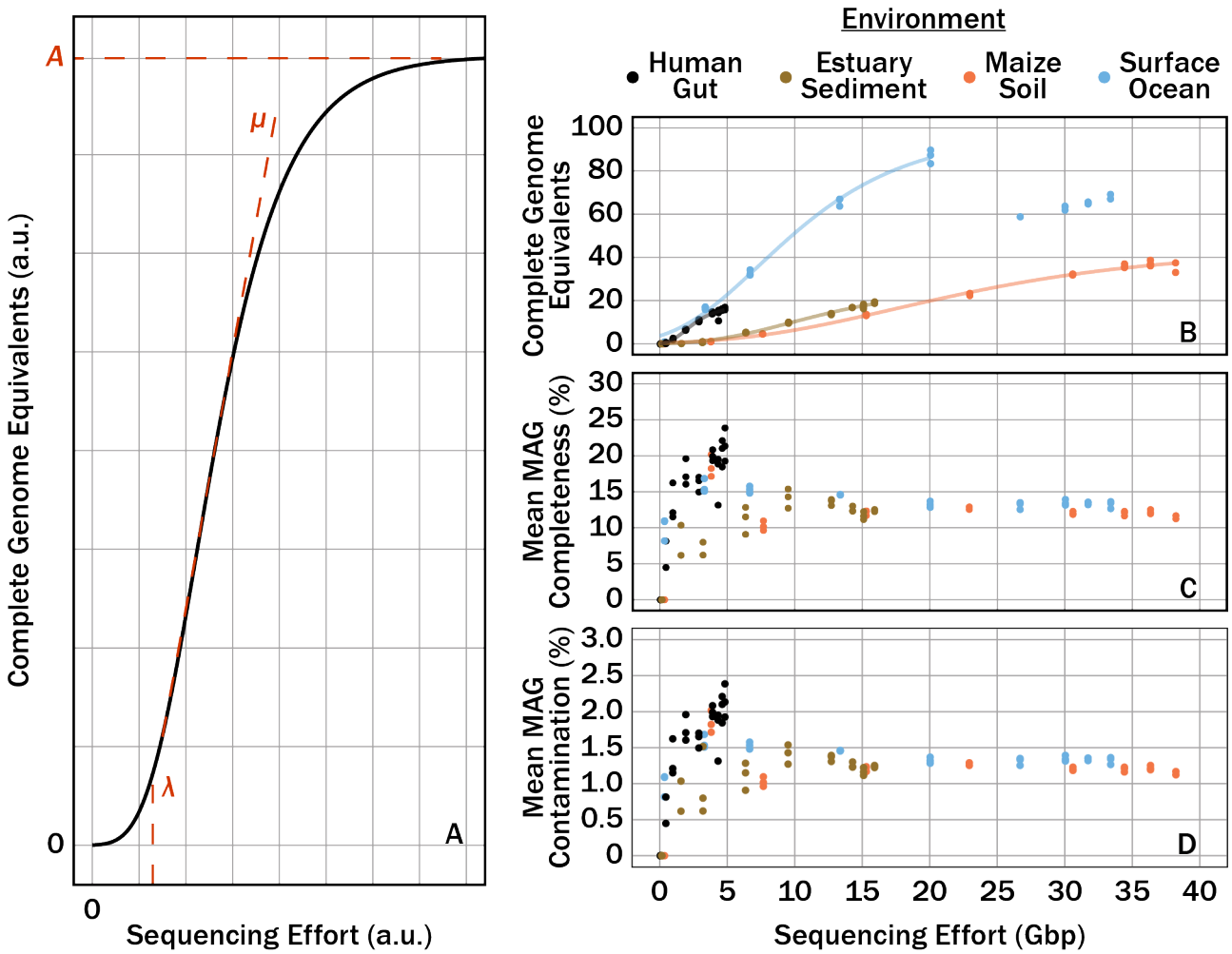


Fig 2. The influence that the parameters A, μ, and λ had on the Gompertz equation (A). The property of the Gompertz equation that each parameter influences is colored red. Complete-genome equivalents (B), mean MAG completeness (C), mean MAG contamination (D) as a function of simulated sequencing effort (Gbp) for sequence datasets of the human gut, maize soil, estuarian sediment, and surface ocean microbiomes, using the pipeline described in the methods section. Translucent lines in (B) correspond to nonlinear least squares fits of the Gompertz equation to the respective environmental dataset.

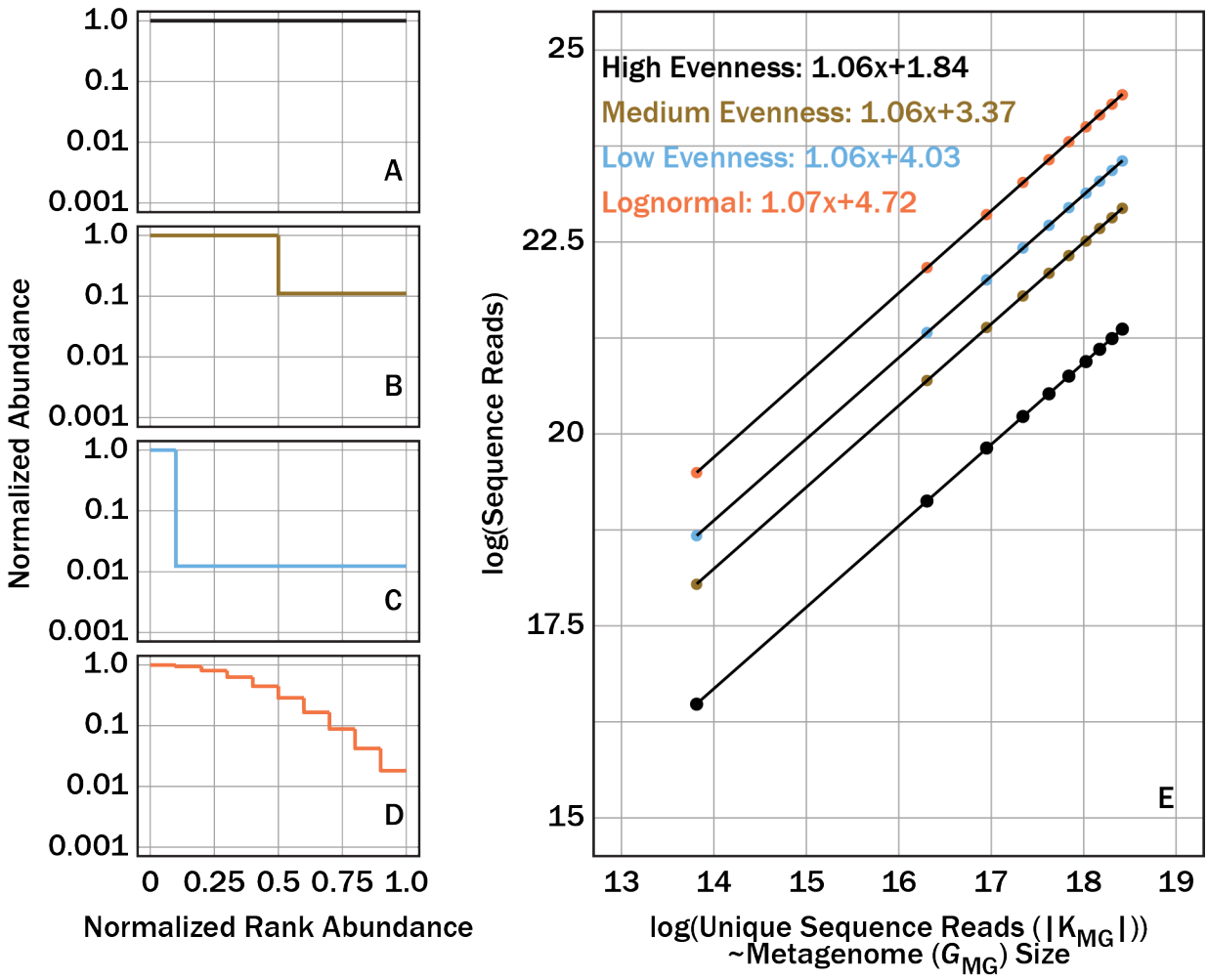


Fig 3. Average expected sequences required to fully sequence four different community structures, one with relatively high community evenness (A), relatively moderate community evenness (B), relatively low community evenness (C), and one with a lognormal community structure (D), were predicted using linear regressions (E) and the log of $|K_{MG}|$ from equation 7 as a predictor.

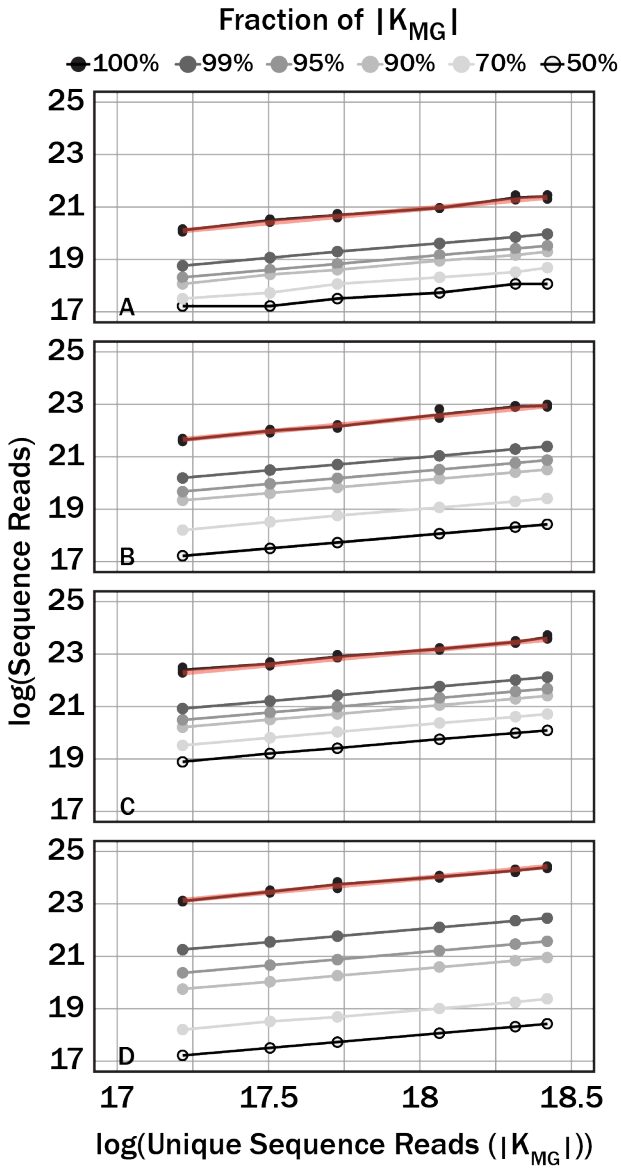


Fig 4. Sequences necessary to reach variable target sequencing depths (colors) for four different community structures, one with relatively high community evenness (A), relatively moderate community evenness (B), relatively low community evenness (C), and one with a lognormal community structure (D). Red translucent lines correspond with linear regression curves for the respective community in Fig 3E.

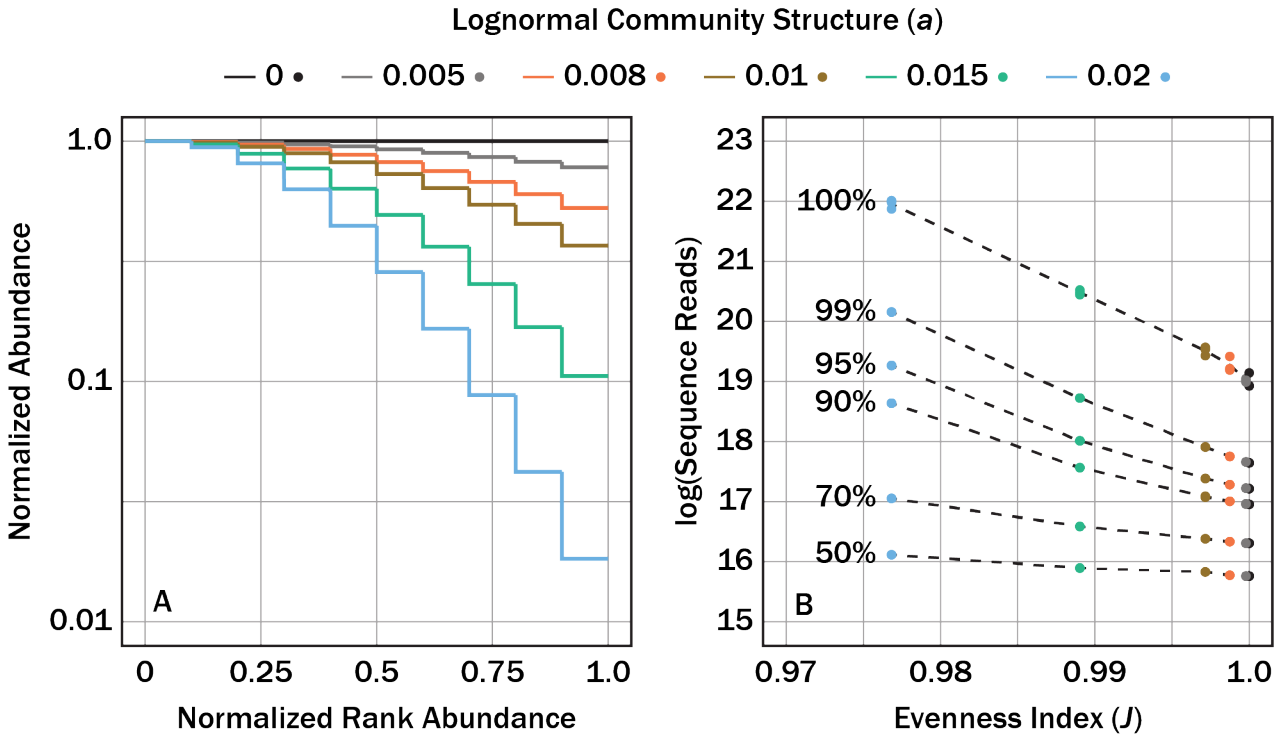


Fig 5. Numerical sequencing simulations applied to 6 hypothetical communities with different lognormal distributions that were defined by the parameter, *a*, from equation 10 (A). The number of sequences necessary to sequence a target fraction of $|K_{MG}|$ (dashed contours) as a function of the Pielou evenness index, *J*, for a given lognormal community structure (B).

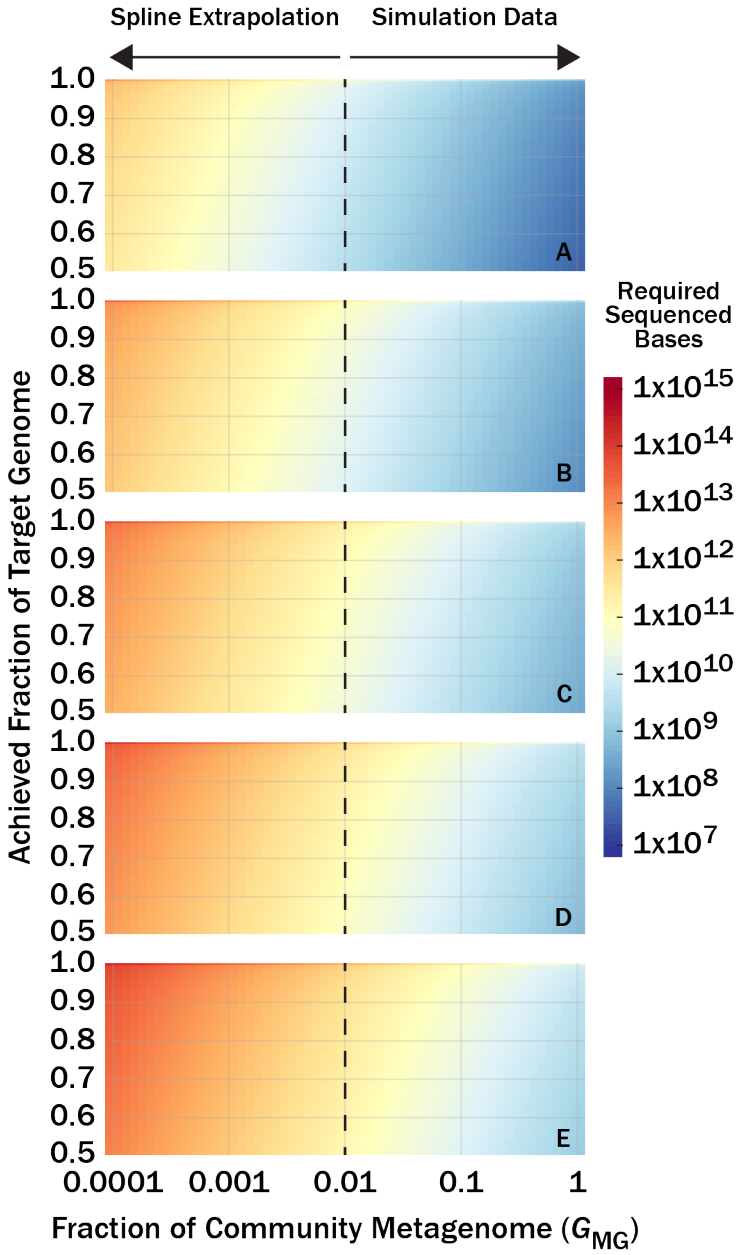


Fig 6. Numerical sequencing simulations show the number of bases (color bar) required to sequence a target fraction of a genome as a function of that genome's relative abundance in the community metagenome. Genomes evaluated were 0.5×10^6 (A), 2×10^6 (B), 5×10^6 (C), 10×10^6 (D), and 20×10^6 (E) base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

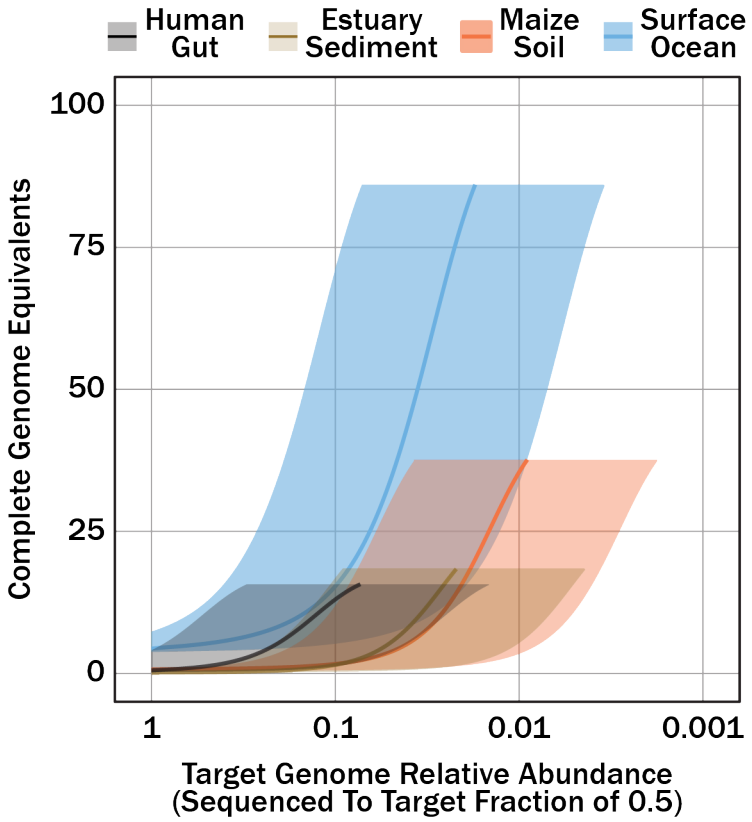


Fig 7. Complete-genome equivalents as a function of a target genome's relative abundance when sequenced to a target fraction (0.5) for human gut, maize soil, estuarian sediment, and surface ocean sequence datasets. Sequencing effort was converted to genome relative abundance sequenced to target fraction using the general additive model presented in Fig 6. 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 5 Mbp.

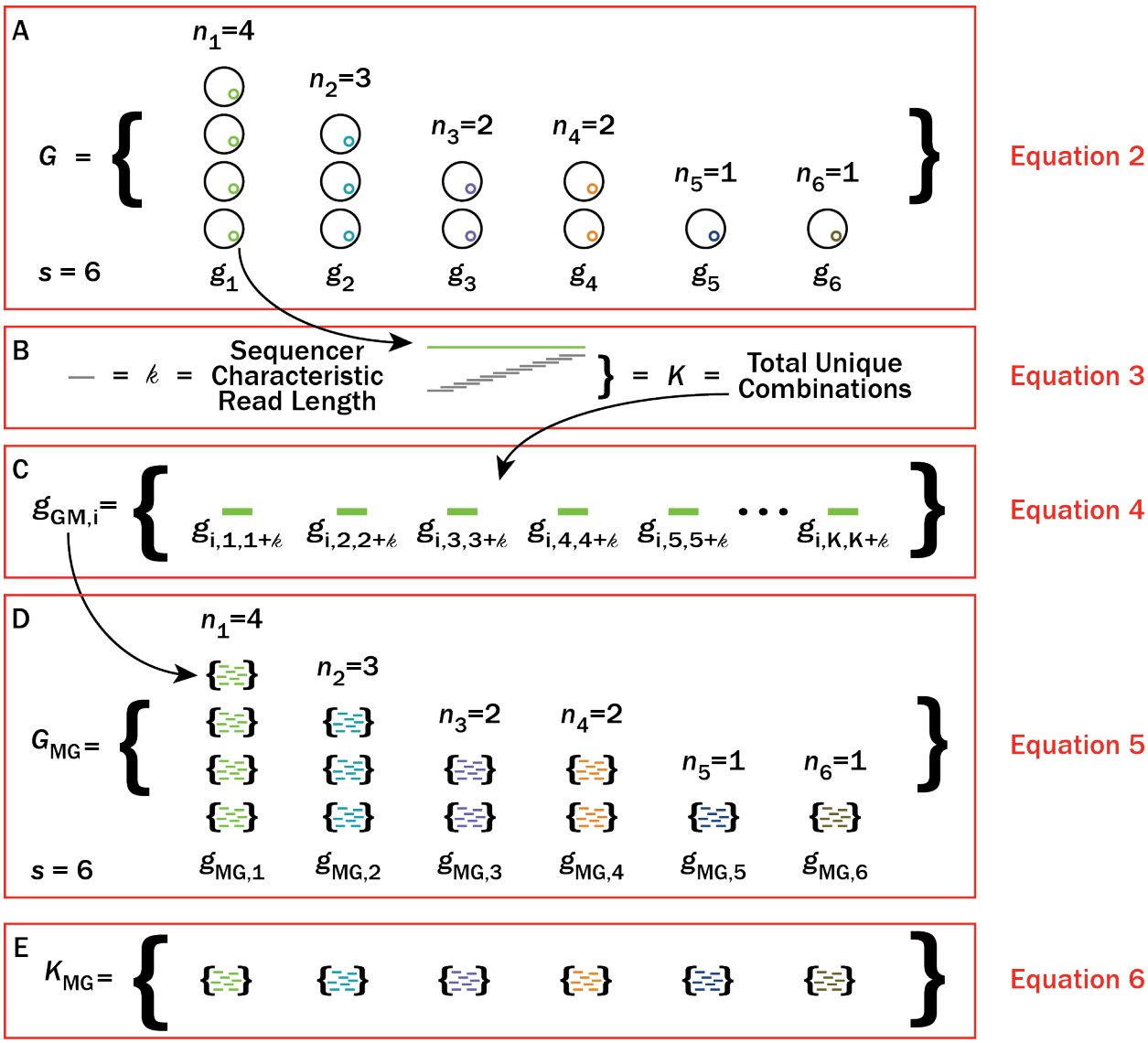


Fig 8. A cartoon illustrating an example microbial community (G), the metagenomes for genomes ($g_{MG,i}$) (as defined in equation 5) within G, and the overall metagenome for the given microbial community (G_{MG}). In this example, there are 6 genomes (s=6) and a total of 13 individual microbes. (A) Black circles represent individual microbes whose genomes are averaged together, g. The average genome, 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 (A) with metagenomes, g_{MG}, gives the metagenome of the microbial community, G_{MG}.

Sample Microbial Community

Sequencing Effort

Bioinformatic Pipeline

Metabolic Characterization

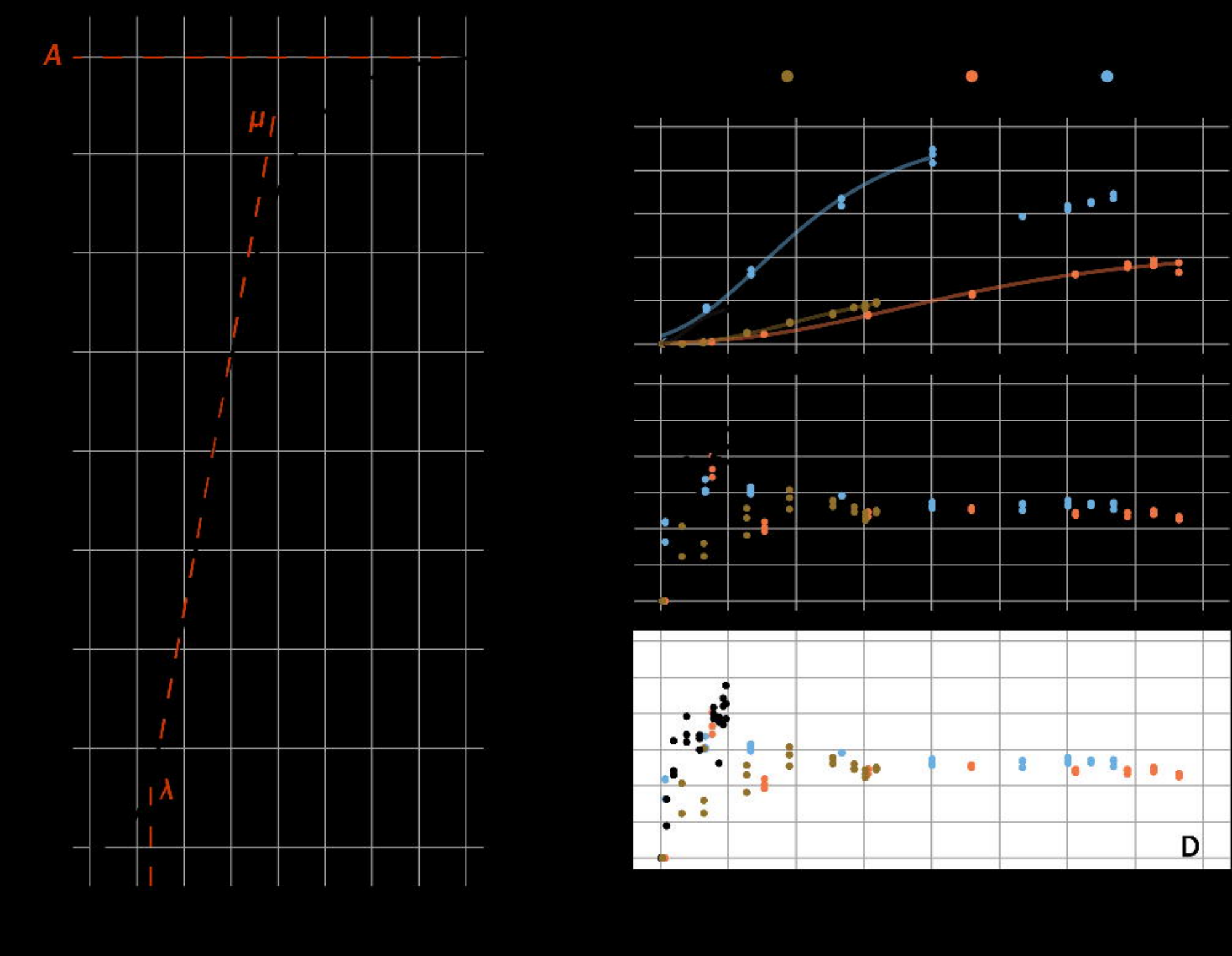
Knowns

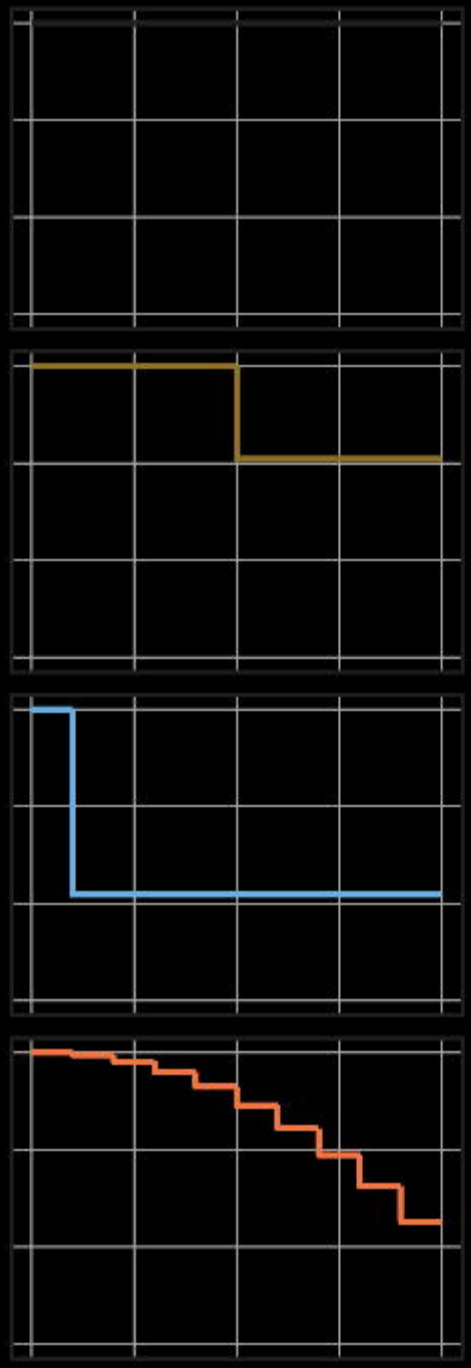
1) Target Environment
2) Sequencing Depth
3) Software/Assembly Parameters
4) Reference Databases

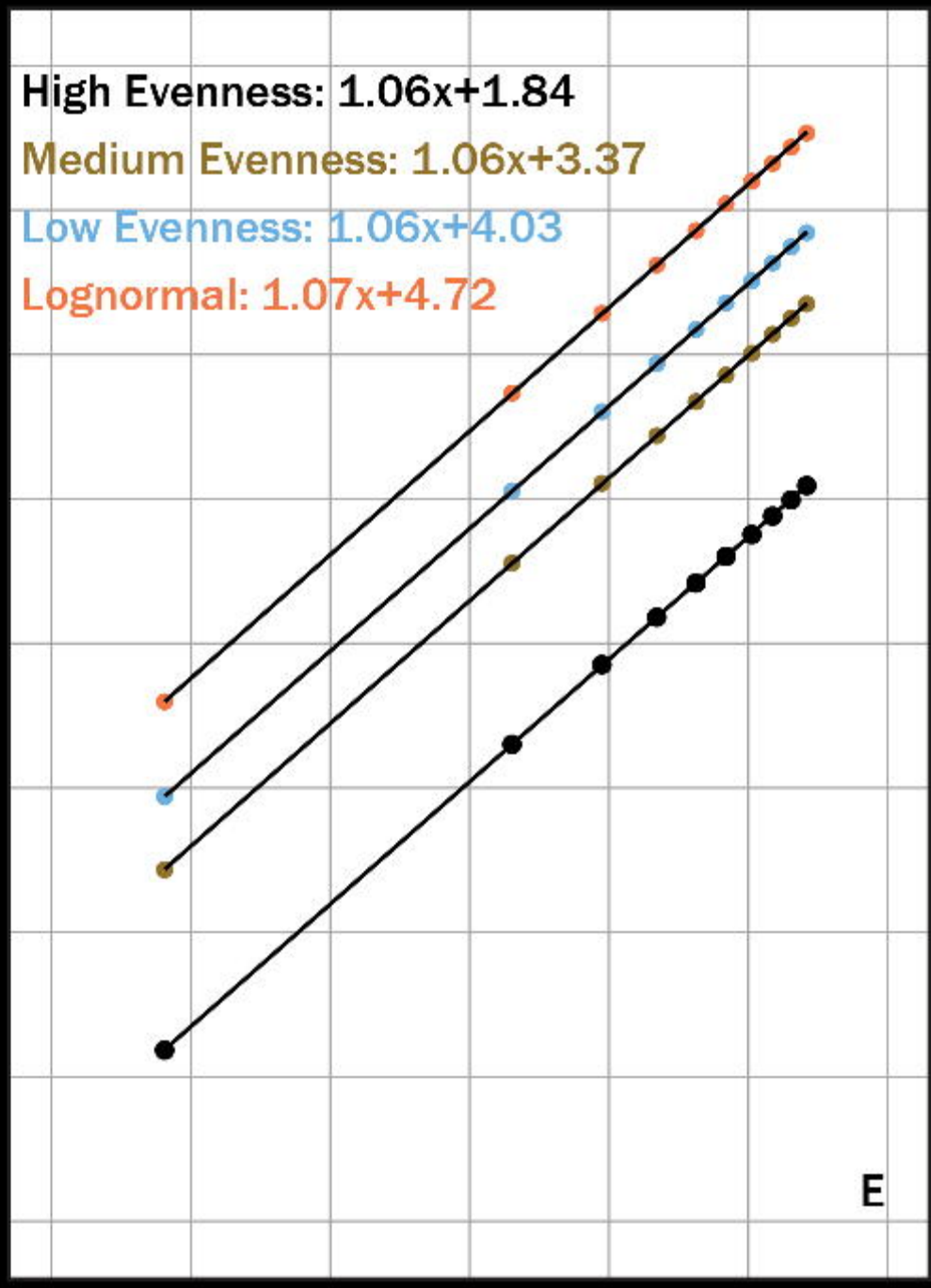
Community Structure

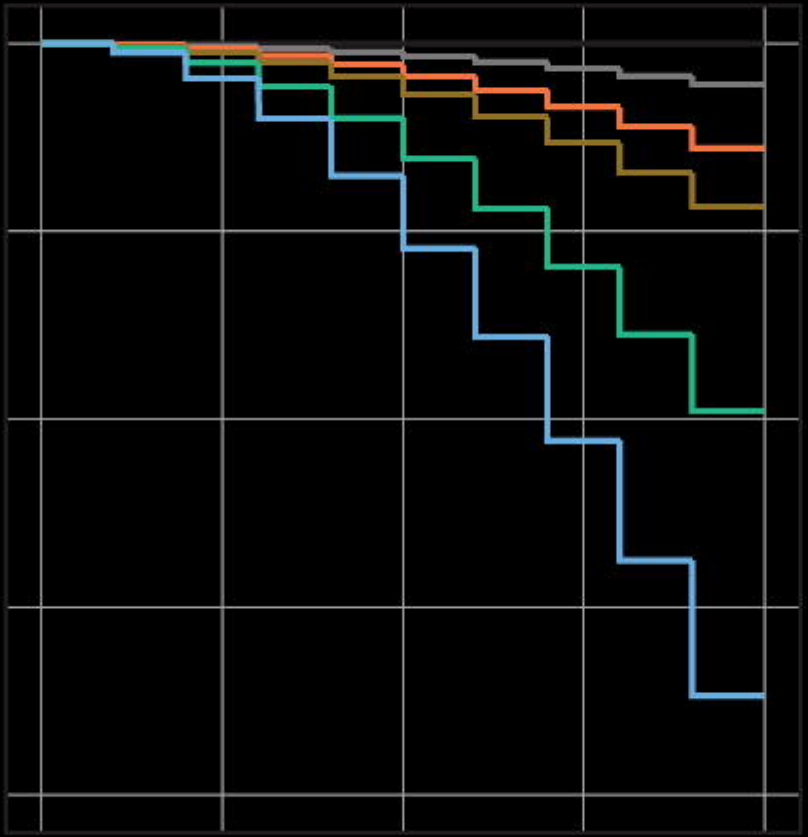
Unknows

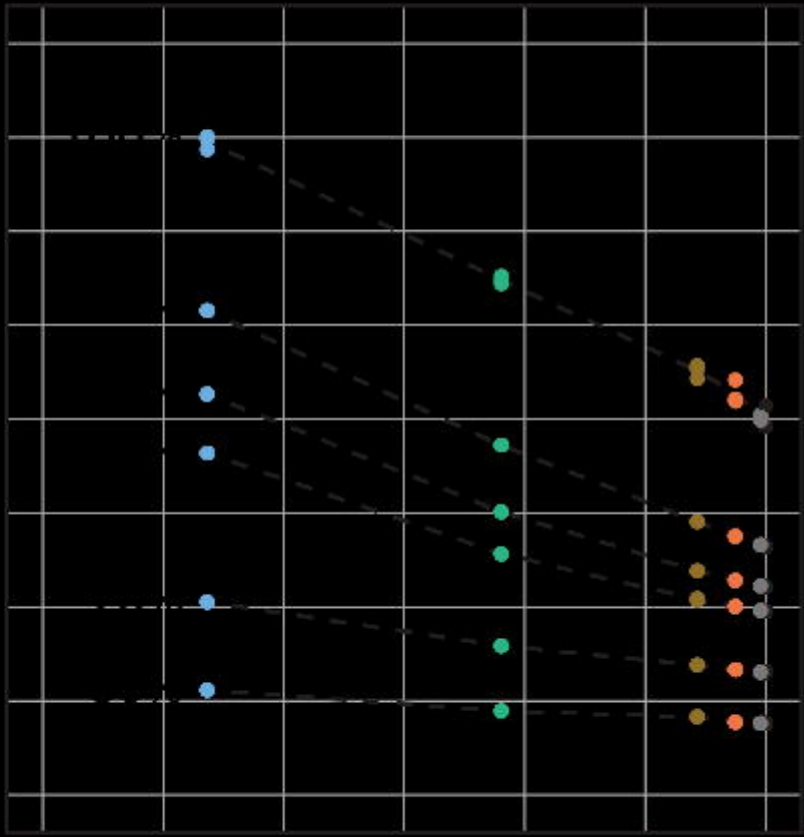
1) Depth and Coverage
2) Pipeline Efficiency
3) Pipeline Accuracy

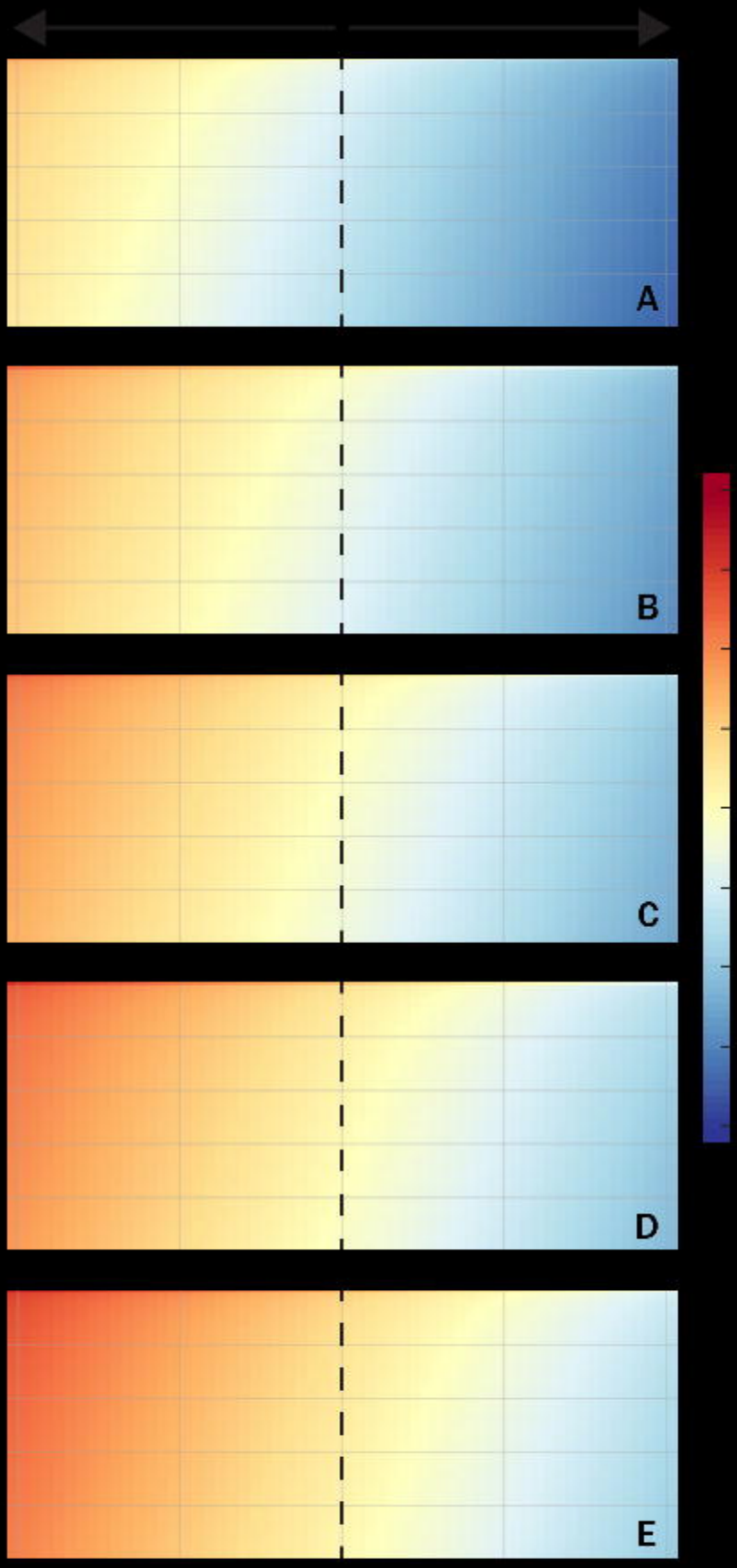

A
μ
λ
D

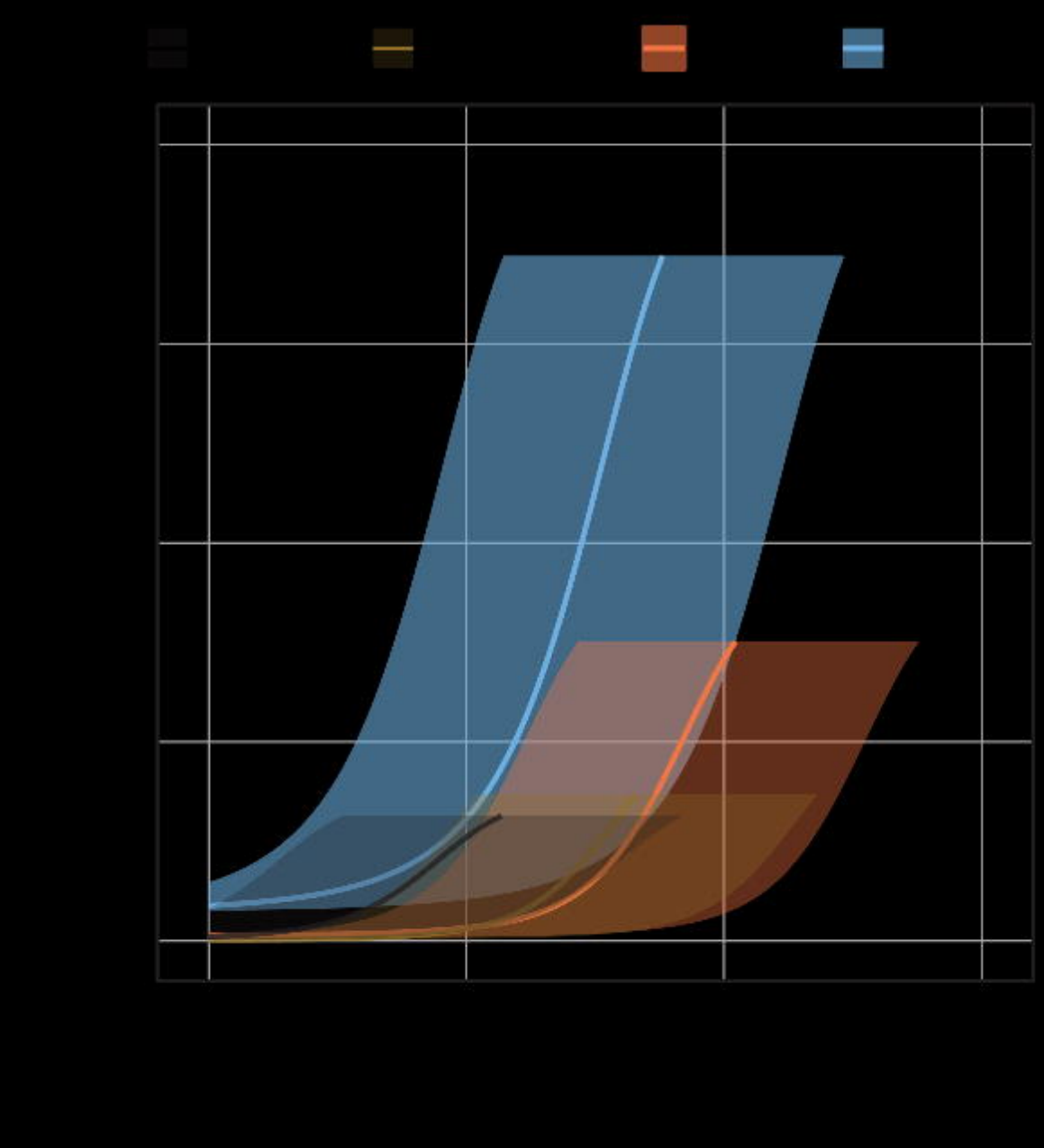


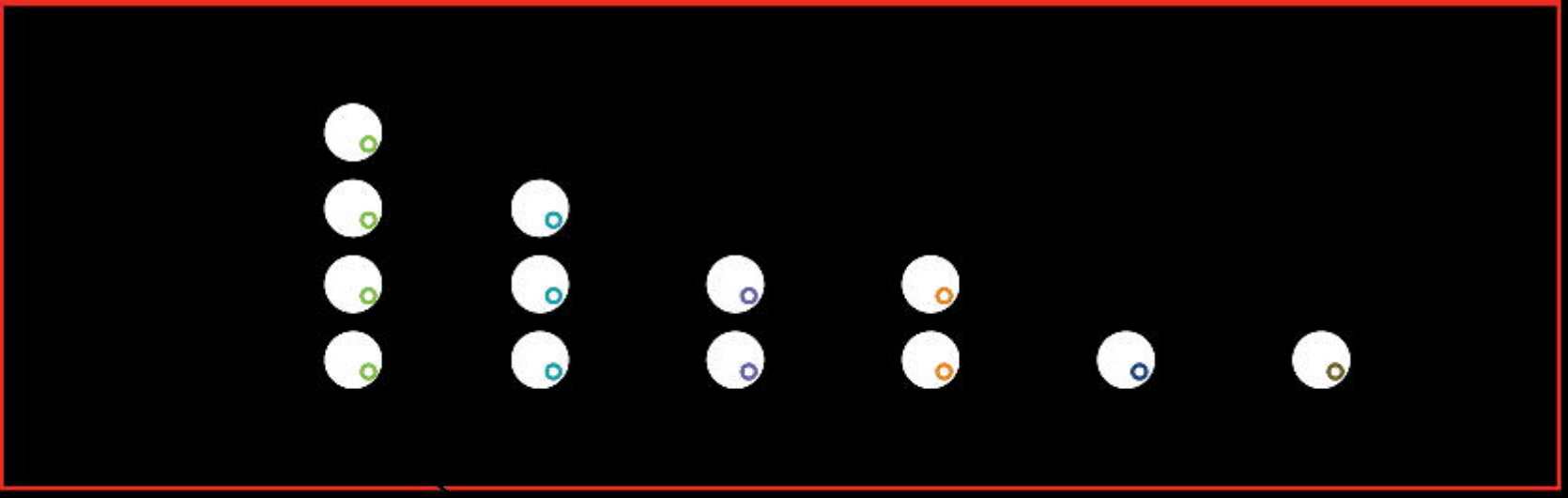

High Evenness: 1.06x+1.84
Medium Evenness: 1.06x+3.37
Low Evenness: 1.06x+4.03
Lognormal: 1.07x+4.72
E



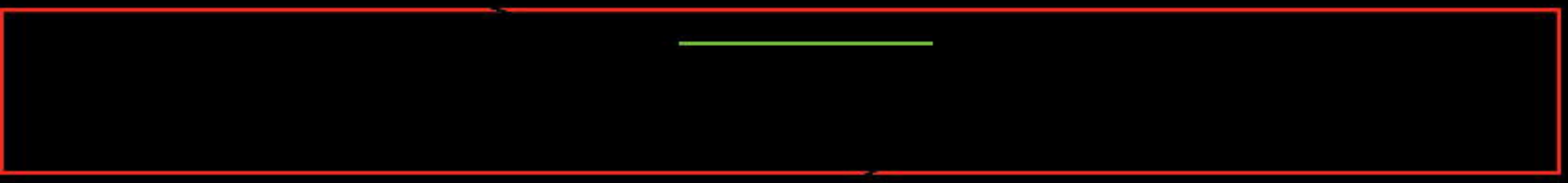



A
B
C
D
E



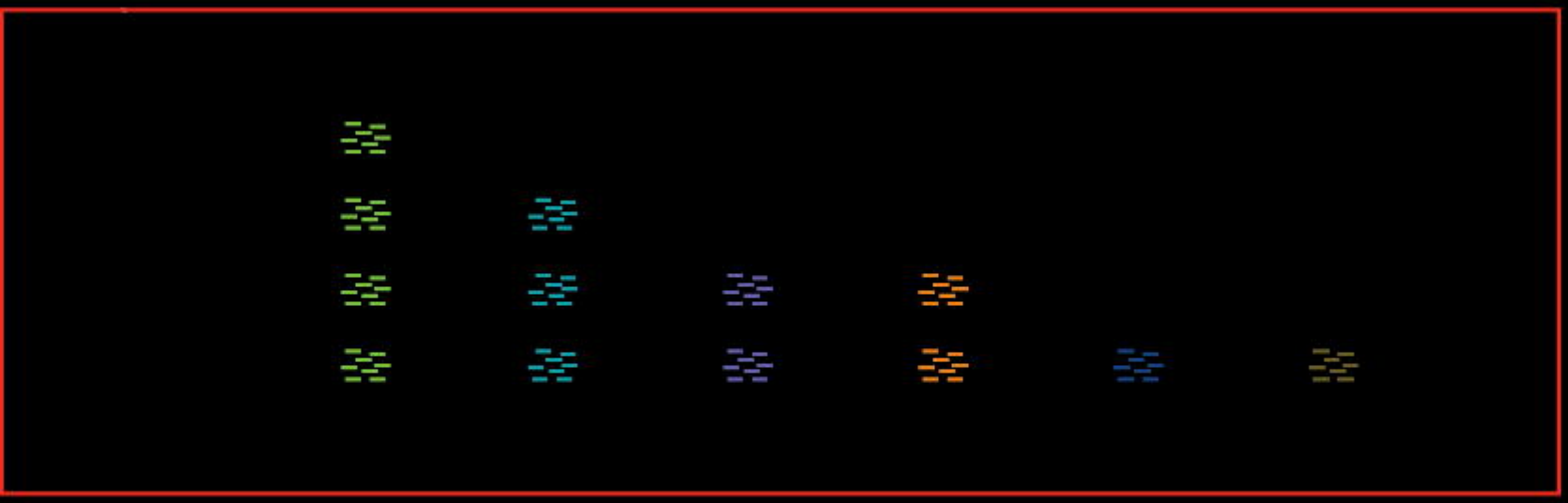


Equation 2



Equation 3

Equation 4



Equation 5

Equation 6