Quantitative modeling for microbial ecology and clinical trials

by

Scott Wilder Olesen

B.A., Williams College (2010)
M.A.St., University of Cambridge (2011)
M.Phil., University of Cambridge (2012)

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2016

Author .....
Department of Biological Engineering
19 August 2016

Certified by .....
Eric J. Alm
Professor
Thesis Supervisor

Accepted by .....
Forest White
Chair of Graduate Program, Department of Biological Engineering

Quantitative modeling for microbial ecology and clinical trials

by

Scott Wilder Olesen

Submitted to the Department of Biological Engineering
on 19 August 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Engineering

Abstract

Microbial ecology has benefited from the decreased cost and increased quality of next-generation DNA sequencing. In general, studies that use DNA sequencing are no longer limited by the sequencing itself but instead by the acquisition of the samples and by methods for analyzing and interpreting the resulting sequence data. In this thesis, I describe the results of three projects that address challenges to interpreting or acquiring sequence data. In the first project, I developed a method for analyzing the dynamics of the relative abundance of operational taxonomic units measured by next-generation amplicon sequencing in microbial ecology experiments without replication. In the second project, I and my co-author combined a taxonomic survey of a dimictic lake, an ecosystem-level biogeochemical model of microbial metabolisms in the lake, and the results of a single-cell genetic assay to infer the identity of taxonomically-diverse, putatively-syntrophic microbial consortia. In the third project, I and my co-author developed a model of differences in the efficacy that stool from different donors has when treating patients via fecal microbiota transplant. We use that model to compute statistical powers and to optimize clinical trial designs. Aside from contributing scientific conclusions about each system, these methods will also serve as a conceptual framework for future efforts to address challenges to the interpretation or acquisition of microbial ecology data.

Thesis Supervisor: Eric J. Alm
Title: Professor

Acknowledgments

Many people contributed to my PhD, including:

- my advisor Eric Alm, whom I thanks for for his mentorship, unswerving advocacy, and delightful collegiality,
- my thesis committee members Paul Blainey and Martin Polz,
- my scientific mentors Pierre Wiltzius, Sophia Gershman, Daniel Aalberts, Dieter Bingemann, Sarah Bolton, Ward Lopes, Paul Selvin, and David Wales,
- my collaborators and mentors inside the Alm Lab, especially Sarah Preheim, Ilana Brito, Thomas Gurry, Sarah Spencer, and Su Vora,
- my outside collaborators, especially Terry Hazen and his group,
- Doug Lauffenburger and Forest White for making Biological Engineering a great community to live and work in,
- my colleagues and friends in the Alm Lab,
- my colleagues and friends at MIT, especially the 2012 class of Biological Engineering graduate students, Jaime Goldstein, and the founding class of Writing Lab Fellows,
- my parents Mary and Neil and my brother Andrew.

My PhD was also aided by funding from Williams College's Dr. Herchel Smith Fellowship, MIT's Presidential Fellowship, the National Science Foundation's Graduate Research Fellowship, BP, and the Department of Energy.

Contents

Chapter 1

Introduction

Microbes, the oldest and most diverse type of life on Earth, are essential to human life and industry. Microbes catalyze key biogeochemical cycles, clean up toxic pollution, and can prevent or cure disease. Microbial ecology—the study of microbes' relationships with one another and their environment—has benefited tremendously from nucleic acid sequencing technology. RNA was first sequenced in the 1960s [1]. Microbiologists, notably Carl Woese, used sequences of 16S rRNA, present in all bacteria and archaea, to develop a more accurate "tree" of microbial life [2, 3]. DNA was first sequenced in the 1970s, and 16S rRNA gene sequences were directly amplified from the environment and sequenced for the first time in 1990 [4, 5]. High-throughput sequencing and sample multiplexing [6] have enabled microbial ecologists to collect large datasets at a decreasing cost of time and money.

Even as tools to process nucleic acid sequence data for microbial ecology mature, there remain challenges to interpreting sequence data and integrating them with other types of microbial ecology data. In this thesis, I present the results of three projects that aim, in part, to help interpret microbial sequence data. In Chapter 2, I discuss the Treatment Effect eXplorer for Microbial Ecology Experiments (`texmex`), an analytical tool designed to extract information about the dynamics of microbial taxa from microbial ecology experiments with few timepoints and few or no replicates. In Chapter 3, I discuss a novel biogeochemical model of a lake ecosystem, a novel network-analysis method for organizing microbial ecology sequence data, and

intepretative frameworks for linking those models and methods with survey data and a single-cell genetic assay. In Chapter 4, I discuss an "omics-free" model designed to guide the design and execution of clinical trials that use fecal microbiota transplants. In the rest of this chapter, I introduce and contextualize these contributions; in the final chapter, I discuss the limitations and potential extensions of theses studies.

1.1 Interpreting microbial ecology sequence data

A major challenge in analyzing and interpreting microbial ecology experiments that use sequencing is to determine which microbial taxa are meaningfully different in abundance. Most approaches to this question are statistical and focus, one-by-one, on the abundances of each taxon, analogous to methods that identify differentially expressed genes in transcriptomic data (e.g., DESeq [7]). The t-test and simple ordinations have been mainstays of this type of analysis for microbial ecology sequence data [8].

This approach presents difficulties when there are few or no biological replicates, compositional effects affect measurements of composition, or when it is important to integrate pre-test samples. I therefore developed `texmex`, which considers the distribution of abundances of microbial taxa *within* a sample to make a more coherent comparison of abundances across samples. I noted that microbial abudances within a sample follow the Poisson lognormal distribution, which had been previously studied in macroecology, and used that distribution to "normalize" the abundances of microbial taxa within samples before comparing abundances across samples. As described in Chapter 2, I applied this technique to short timeseries experiments measuring the response of ocean microbial communities to amendment with crude oil, verifying the presence of known oil degraders and suggesting that other organisms, implicated in other crude oil microcosm experiments, are also involved in crude oil degradation.

Microbial ecology surveys that use taxonomic marker gene sequence data confront a different problem: the relationship between phylogeny and function. Taxonomic marker gene sequencing (e.g., 16S rRNA amplicon sequencing) provides information

about microbial community structure but provides only indirect information about microbial community function [9]. Metagenomic sequencing provides more direct information about community function, but it is expensive and mostly discards information about the relationships between phylogeny and function.

In Chapter 3, I describe a set of tools used to identify potential consortia of cooperating organisms in a natural ecosystem. This study includes a taxonomic survey of the environment, an ecosystem-level model of microbial metabolisms, and a single-cell genetic assay that links phylogeny and function. For that study, I developed a network-analysis method (operational ecological units) for the survey data and an interpretative framework (inferred biomass) for the model result. These tools provided a conceptual link between all three types of data, which were together essential for the putative *in situ*, perturbation-free identification of microbial consortia.

1.2 Clinical trial design

Fecal microbiota transplantation (FMT) is a highly effective intervention for patients suffering from recurrent *Clostridium difficile*, a common nosocomial infection [10]. As the gut microbiome is assigned an increasingly important role in the development and maintenance of host gut health, immunity, and even psychology, the success of FMT for *C. difficile* has generated interest in using FMT to treat other conditions to which an "unhealthy" gut microbiota may contribute [11]. Although the exact mechanism by which FMT treats *C. difficile* is not well understood, a few clinical trials have already experimented with using FMT as a treatment for other conditions, and many more trials for a variety of diseases will probably begin in the coming years.

These early trials using FMT to treat conditions beyond *C. difficile* have had some confusing and disappointing results. Of two recent trials using FMT to treat inflammatory bowel disease, one failed [12] and the other produced an unexpected result: five of six stool donors in the trial appeared to produce results no better than treating the patients with placebo, while the sixth donor produced stool that appeared to have substantially greater efficacy [13].

If this sort of difference in efficacy between donors' stool is real, then what are the implications for clinical trial design? To answer this question, I designed a model of differences in stool efficacy and used the model to predict the statistical power of typical trial designs. These results, laid out in Chapter 4, show that, if the results from this recent inflammatory bowel disease trial are representative, then FMT's efficacy is only likely to be discovered using larger patient cohorts and response-adaptive allocation of donors' stool. I used this model to aid the design of a two-stage phase I clinical trial using FMT to treat a gastrointestinal condition, and I expect that these power calculations and adaptive allocation strategies will improve clinical trial design, improving the probability that patients will have access to a new therapy.

Bibliography

[1] F. Sanger, G. G. Brownlee, and B. G. Barrell. A two-dimensional fractionation procedure for radioactive nucleotides. *J Mol Biol*, 13(2):373–IN4, 1965.

[2] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci USA*, 74(11):5088–5090, 1977.

[3] N. R. Pace, J. Sapp, and N. Goldenfeld. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proc Natl Acad Sci USA*, 109(4):1011–1018, 2012.

[4] S. J. Giovannoni, T. B. Britschgi, C. L. Moyer, and K. G. Field. Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, 345:60–63, 1990.

[5] R. J. Case, Y. Boucher, I. Dahllöf, C. Holmström, W. F. Doolittle, and S. Kjelleberg. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol*, 73(1):278–288, 2007.

[6] M. Hamady, J. J. Walker, J. K. Harris, N. J. Gold, and R. Knight. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Meth*, 5(3):235–237, 2008.

[7] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):1–12, 2010.

[8] N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett, and C. Huttenhower. Metagenomic biomarker discovery and explanation. *Genome Biol*, 12(6):1–18, 2011.

[9] M. G. I. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko, and C. Huttenhower. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotech*, 31(9):814–821, 2013.

[10] Z. Kassam, C. H. Lee, Y. Yuan, and R. H. Hunt. Fecal microbiota transplantation for *Clostridium difficile* infection: Systematic review and meta-analysis. *Am J Gastroenterol*, 108(4):500–508, 2013.

[11] L. P. Smits, K. E. C. Bouter, W. M. de Vos, T. J. Borody, and M. Nieuwdorp. Therapeutic potential of fecal microbiota transplantation. *Gastroenterology*, 145(5):946–953, 2013.

[12] N. G. Rossen, S. Fuentes, M. J. van der Spek, J. G. Tijssen, J. H. A. Hartman, A. Duflou, M. Löwenberg, G. R. van den Brink, E. M. H. Mathus-Vliegen, W. M. de Vos, E. G. Zoetendal, G. R. D'Haens, and C. Y. Ponsioen. Findings from a randomized controlled trial of fecal transplantation for patients with ulcerative colitis. *Gastroenterology*, 149(1):110–118, 2015.

[13] P. Moayyedi, M. G. Surette, P. T. Kim, J. Libertucci, M. Wolfe, C. Onischi, D. Armstrong, J. K. Marshall, Z. Kassam, W. Reinisch, and C. H. Lee. Fecal microbiota transplantation induces remission in patients with active ulcerative colitis in a randomized controlled trial. *Gastroenterology*, 149(1):102–109.e6, 2015.

Chapter 2

A novel analysis method for paired-sample microbial ecology experiments

The contents of this chapter were published as: Olesen SW, Vora S, Techtmann SM, Fortney JL, Bastidas-Oyanedel JR, Rodríguez J, *et al.* (2016) A Novel Analysis Method for Paired-Sample Microbial Ecology Experiments. *PLoS ONE* **11**(5): e0154804. doi:10.1371/journal.pone.0154804.

The figures, tables, and supplementary figures and tables are at the end of the chapter.

ABSTRACT

Many microbial ecology experiments use sequencing data to measure a community's response to an experimental treatment. In a common experimental design, two units, one control and one experimental, are sampled before and after the treatment is applied to the experimental unit. The four resulting samples contain information about the dynamics of organisms that respond to the treatment, but there are no analytical methods designed to extract exactly this type of information from this configuration of samples. Here we present an analytical method specifically designed to visualize and generate hypotheses about microbial community dynamics in experiments that have paired samples and few or no replicates. The method is based on the Poisson lognormal distribution, long studied in macroecology, which we found accurately models the abundance distribution of taxa counts from 16S rRNA surveys. To demonstrate the method's validity and potential, we analyzed an experiment that measured the effect of crude oil on ocean microbial communities in microcosm. Our method identified known oil degraders as well as two clades, *Maricurvus* and Rhodobacteraceae, that responded to amendment with oil but do not include known oil degraders. Our approach is sensitive to organisms that increased in abundance only in the experimental unit but less sensitive to organisms that increased in both control and experimental units, thus mitigating the role of "bottle effects".

INTRODUCTION

Paired-sample microbial ecology experiments

Many microbial ecology experiments use amplicon-based sequencing (e.g., 16S rRNA gene sequencing) to study the dynamics of a microbial community, whose members are typically grouped into operational taxonomic units (OTUs). A straightforward experimental design uses control units (i.e., subjects, microcosms, or animals not subjected to the treatment) and pretests (i.e., measurements of microbial community composition taken before the treatment is applied to the experimental units). This two-timepoint, paired-sample design is intended to identify changes in community composition that are specific to the experimental treatment, rather than all the changes that occur in the experimental unit, some of which might be irrelevant to the intentionally applied treatment. For example, aquatic microbial communities transplanted into microcosms often suffer dramatic changes in community composition on account of the transplantation. These extraneous community composition changes are called “bottle effects”. Over the course of the experiment, bottle effects are entangled with the effects caused by the experimental treatment [1]. A similar complication can occur in experiments that study animal-associated microbiota, where changes in community composition can be caused by, for example, interactions with experimenters or circulation of a microbe in the animal facility.

If an experiment is performed with many replicates, there are statistical methods that can make statements about whether the observed changes between the groups were meaningful with respect to a null hypothesis. In some cases, an experiment is very

preliminary or the samples are so precious that robust replication is not feasible. For example, we intended the test the effect of crude oil on an ocean microbial community and did not have enough of the sample material, which was collected on a special cruise, for many replicates. Without sufficient replications, it is impossible for us to make a statement about the statistical significance of the results of such an experiment, but we wanted to obtain as much information as possible from such difficult-to-acquire samples, with the caveat that any conclusion would need a separate, properly-powered experiment for verification.

Because of the complications in paired-sample experiments in general and our experimental setup in particular, we aimed to create a method to visualize and generate hypotheses about the dynamics of microbial communities in paired-sample experiments with few or no replicates. Most popular tools for analyzing the results of microbial ecology experiments have shortcomings or complications when applied to this particular experimental setup. We review some classes of these existing tools below.

For convenience, we refer to the four samples in a two-timepoint, two-unit experiment as "control-before" (control unit, pretest sample), "control-after" (control unit, post-treatment sample), "experimental-before", and "experimental-after".

Existing analytical techniques

Ordination techniques. Ordination techniques include principal component analysis (PCA), multidimensional scaling (MDS), redundancy analysis (RDA), and canonical

analysis of principal components (CAP) [2, 3]. Although very useful for analyzing and visualizing the relationships between many samples, ordination techniques are not easy to interpret in the context of a paired-sample microbial ecology experiment. In an experiment with no replicates, an ordination plot will only have four points, making it difficult to visually or analytically extract interesting information about the community's dynamics.

Clustering techniques. In the context of microbial ecology, clustering techniques like hierarchical clustering or *k*-means clustering use a dissimilarity metric to infer which samples in a data set group together [2]. Clustering techniques are often combined with ordination techniques and give similar insight. They therefore suffer similar drawbacks when analyzing the results of paired-sample experiments. In the case of a single paired-sample experiment with no replicates, there are only four samples, so a clustering technique can only produce a limited number of analytical outcomes.

Beta diversity and other tests. Statistical tests like ANOSIM [4] and PERMANOVA [5] are designed to evaluate whether some set of samples in a data set are more similar to one another than they are to other samples in the data set. In general, rigorous statistical inference and machine learning methods require many replicates, which might be impractical for exploratory studies. For example, La Rosa *et al*. [6] calculate that at least 25 samples are needed to determine if two groups of human microbiome samples have meaningfully different compositions at 90% power.

Indicator species techniques. Indicator species techniques like IndVal [7] are intended to identify species that especially informative with respect to the ecological community's composition or abiotic context [8]. Indicator species techniques treat each sample as an independent community rather than, as we consider, timepoints from the same community. For example, if an OTU is rare in the control-before and control-after but abundant in the experimental-before and experimental-after, an indicator species approach would identify that OTU as important to distinguishing the experimental and control *units*, which, although true, does not correctly reflect that that OTU is probably not affected by the experimental treatment, since it changed in neither the control or experimental unit.

OTU-by-OTU techniques. There are techniques that can identify OTUs whose dynamics merit further investigation, even if those OTUs are not abundant enough to appear in a community composition chart or if there are not enough replicates for statistical inference. For example, OTUs can be ordered according to their change in relative abundance between the pretest and the post-treatment sample. However, unintuitive signals, called "compositional effects", can arise when standard analytical techniques are used with compositional data sets like OTU count data [9, 10]. In an aquatic microcosm experiment, for example, a single organism might bloom in response to the experimental treatment, causing an organism with constant absolute abundance to decrease in relative abundance. The reverse is also possible: an organism with constant absolute abundance increases in relative abundance when other organisms decrease in absolute abundance.

Opportunity for a distribution-based technique

To incorporate controls and pretests, it is important to be able to meaningfully compare OTUs' abundances across samples. As discussed above, relative abundances are susceptible to compositional effects. In contrast, nonparametric analysis, which uses only the ranks of abundance-ordered OTUs, is robust to compositional effects but loses much of the quantitative information encoded in the relative abundances. For example, an arbitrarily large change in the abundance of one OTU can cause arbitrarily large changes in the relative abundances of all OTUs but will only change the ranks of each OTU by, at most, one. Using ranks presents a tradeoff between robustness (i.e., each rank changing by at most one) and loss of quantitative information (e.g., if the most abundant OTU doubles in abundance, no ranks change). Ranks are also challenging to use with OTU count data because many OTUs have the same number of counts.

If OTU abundances were distributed in a way that could be reliably modeled, a compromise between relative abundances and ranks would be possible. For example, if one organism blooms and all others remain at constant absolute abundance, the overall shape of the distribution of abundances would change very little. As in an analysis that uses relative abundances, the blooming OTU would be registered because that OTU would move up in the distribution. As in an analysis that uses ranks, the unchanging OTUs would remain at the same places in the distribution even though their relative abundances decreased.

Our method

In this paper, we present an analytical method designed to measure the dynamics of OTUs across two timepoints, to correct for bottle effects using control units, and to correct for unit-specific effects using pretests. The method is framed in terms of an abundance distribution from macroecology research, the Poisson lognormal distribution, that we found accurately models the abundance distribution of OTU count data. As a test of the method's validity, we show that, in the context of a bioreactor experiment, this method reports that OTUs in microbial communities derived from the same inoculum and subjected to strong but identical conditions have well-correlated responses.

We used our method to identify OTUs in a complex ocean water community, collected off the Egyptian coast, that respond to amendment with crude oil. Most of the sequences we identified classified as *Maricurvus*, *Pseudomonas*, *Alcanivorax*, *Methylophaga*, and Rhodobacteraceae. These clades include known oil degraders as well as organisms that other microbial ecology experiments have suggested may degrade oil. These results demonstrate that our method will be useful for visualizing the effect of the treatment of interest on all OTUs and for quantifying dynamic changes in abundance in a paired-sample microbial ecology with few or no replicates, although properly-powered follow-up experiments would be needed to verify any of these dynamics.

MATERIALS & METHODS

Poisson lognormal distribution

Theory. Noting that many microbial communities are structured by complex ecological processes, we searched for an ecologically-motivated probability distribution that accurately models the abundance distribution of OTUs in natural microbial communities. The truncated Poisson lognormal (TPL) distribution is an attractive candidate. When a value is drawn from the TPL distribution, a true abundance λ is first drawn from a lognormal distribution (with scale parameter μ and shape parameter σ). Then a random integer is drawn from a Poisson distribution with mean λ. If the integer is zero, a new λ is drawn and is used to draw a new integer.

Among the many models of species-abundance relationships (e.g., [11, 12] and references therein), there is evidence and theory suggesting that fractions of the total niche space allotted to each organisms are approximately lognormal-distributed [13], and the Poisson distribution is a straightforward model for converting a continuous value λ into a random number of discrete counts. The Poisson lognormal has been used to model abundance distributions for plants and animals [14-16] and has been used in at least one study [17] that simulated microbial abundances. In most of these applications, the distribution is truncated at zero counts, since, in most cases, it is impossible to distinguish if a species is absent or if it present but very rare; in both cases, that species would present zero counts.

In a microbial ecology context, the TPL framework asserts that the abundances of

microbial species in an environment are lognormal-distributed, that is, that the logarithms of those abundances have a Gaussian distribution. The framework also asserts that sequencing produces an integer number of reads for each species. The number of sequencing counts for a species with true abundance λ is drawn from a Poisson distribution with mean λ. The parameter μ is related to the mean of the abundances of microbial species in that environment (conditioned on the depth of sequencing). The parameter σ describes the variability of those abundances.

Metrics. Once fit to the OTU abundances in a sample, the TPL distribution provides transformations from raw OTU counts to two different values. As mentioned above, in the TPL framework, OTUs' true abundances λ are assumed to be lognormal-distributed with scale μ and shape σ. If the Poisson lognormal distribution is an accurate model for OTU abundance distributions, samples might have differing parameters μ and σ, but the underlying distribution of $(\log \lambda - \mu)/\sigma$ should be similar across samples. We expect this because if, roughly speaking, λ is lognormal distributed, then $\log \lambda$ is normally distributed, and $(\log \lambda - \mu)/\sigma$ accounts for the differences in means and shapes of the normal distribution. An OTU's number of reads r is the maximum likelihood estimator of its true abundance λ, so we define the *normalized reads* $z \equiv (\log r - \mu)/\sigma$, which estimates $(\log \lambda - \mu)/\sigma$.

Given this metric z that can be compared across samples, we looked for a sensible way to combine these values as a measure of dynamics. To quantify an OTU's dynamics between the two timepoints, we define Δz, the change in rescaled reads. This metric is

similar to the log fold change in relative abundance, an application of the more general log-ratio transformation commonly used with compositional data sets. To show this connection, consider two samples $i \in \{0 = \text{before}, 1 = \text{after}\}$ from one microcosm, either control or experimental, with the TPL fit parameters μ_i and σ_i. One OTU of interest has counts r_i in the two samples. In this case,

$$\Delta z \equiv \frac{\log r_1 - \mu_1}{\sigma_1} - \frac{\log r_0 - \mu_0}{\sigma_0} = \frac{\log r_1}{\sigma_1} - \frac{\log r_0}{\sigma_0} + \text{constant}.$$

For comparison, the log fold change in relative abundance is

$$\log\left(\frac{r_1/N_1}{r_0/N_0}\right) = \log r_1 - \log r_0 + \text{constant},$$

where N_i is the total number of reads in sample i. If $\sigma_0 \approx \sigma_1$, then

$$\Delta z \approx \frac{\text{log fold change}}{\sigma_0} + \text{constant},$$

that is, Δz and the log fold change are approximately linearly related.

Aside from rescaled reads, the TPL distribution can also be used to transform raw OTU counts to the value of the cumulative distribution function F, which has the common range [0, 1] across all samples. Having fit the parameters μ and σ, an OTU with r reads has $F(r) \equiv \sum_{r' \le r} f_{\text{TPL}}(r')$, where f_{TPL} is the probability distribution function for the TPL distribution. Conveniently, ΔF is defined for all r_0 and r_1, while Δz and the log fold change are poorly defined when $r_0 = 0$ or $r_1 = 0$.

Implementation. We implemented the TPL fitting, computation of z and F, and basic visualizations in an R package *texmexseq*, which we have posted at CRAN (cran.r-

project.org). Our package is based on the poilog package [18], to which we add convenience functions for interacting with OTU tables, fitting the distribution to multiple samples, and visualizing Δz and ΔF.

Bioreactor experiment

Experimental design. We performed a bioreactor experiment to verify that our analytical method would identify similar dynamics in identically-treated microcosms. Briefly, three identical anaerobic serum bottles were loaded with sterile anaerobic media, glucose, and anaerobic sludge from Al Mafraq wastewater treatment plant (Abu Dhabi Sewage Services Company, Al Dhafrah, Abu Dhabi, UAE). Glucose was the only carbon source. Each bioreactor was incubated at 35 °C for 48 hours, at which point the bioreactor's contents was spun down and the cell pellet resuspended in fresh media. This process was repeated 7 times. The timepoints analyzed in this study are the initial inoculum's community and the bioreactor's final community (i.e., timepoint 7).

Experimental protocol. Inocula were stored at 4 °C before starting experiments. Fermentations were carried out in 150 mL serum bottles with a working volume of 60 mL. Anaerobic serum bottles were loaded with media, described below, and sludge. The initial biomass concentration for all the three inocula was 10 g/L dry weight matter. Sterile media consisted of 5 g/L glucose (autoclaved separately from mineral media) and phosphate buffer (0.2 g/L $Na_2HPO_4 \cdot 2H_2O$ and 2.5 g/L KH_2PO_4) diluted on basal anaerobic media [19]. Media pH was adjusted to 5.5 with 1 M HCl.

After inoculation, each bottle was crimped using sterile rubber stoppers and flushed with pure N_2 for 2 minutes using sterile 0.45 μm pore gas filters. Bottles were incubated immediately after flushing. To re-suspend the inoculum, the broth was centrifuged in sterilized containers at 5,000 g for 5 minutes. The resulting pellets were re-suspended in 60 mL of fresh media. Bottles were again crimped and flushed. Inoculation and media replacement were all performed in a UV-sterilized laminar flow chamber.

DNA from the bioreactors was extracted with MO BIO Ultra Clean Soil DNA isolation kit according to the manufacturer's protocol. Paired-end Illumina sequencing libraries were constructed using a two-step 16S rRNA PCR amplicon approach described in Preheim *et al.* [20]. Libraries were multiplexed and sequenced on an Illumina MiSeq with paired-end 150 bp reads.

16S data processing. Only forward reads were used in the analysis. Primers were trimmed from the reads by searching for the best-matching position in the read's first 20 bases, allowing a maximum of 2 mismatches between the primer sequence and the read sequence. Reads that did not match the primer were discarded. Reads were demultiplexed by assigning reads to the best-matching barcode sequence, allowing no more than 1 mismatched base. Reads with no acceptable barcode match were discarded. Reads were trimmed to 120 bases. Shorter reads were discarded. Reads with an expected number of errors, as calculated by Edgar & Flyvbjerg [21], greater than 1.0 were discarded. The sequence data for these experiments are in Datasets S1 and S2.

Aquatic microcosm experiment

Inoculum collection. Water samples were collected on 12/13 October 2012 aboard MV *Fugro Navigator* in the West Nile Delta region of the Nile Deep Sea Fan from a station (29.571° E, 31.813° N) with a sea floor depth of 1230 m. This work was conducted in BP's West Nile Delta Concession as part of a larger survey of Eastern Mediterranean ocean microbiology described elsewhere [22]. No specific permits were required for collection of these samples. These field studies did not include the collection of any endangered or protected species. Temperature, salinity, depth, and dissolved oxygen were measured through the water column with a Valeport Midas+ CTD. Samples were collected with Niskin bottles from four depths selected in consideration of differences in temperature, salinity, and depth: within the thermocline (50 m), within an area of increased salinity in the water column (250 m), two-thirds of sea floor depth (824 m), and 20 m above the sea floor (1210 m). Water was removed from the Niskin bottles and stored in pre-cleaned amber glass bottles at 4 °C until the microcosms were set up. In this paper, we analyze the result of the experiment performed using water from 824 meters depth.

Microcosm design and sampling. We performed a microcosm experiment to evaluate the effect of crude oil on the microbial community in those water samples. 2 L of water from each depth was used for microcosms, 1 L each for a control microcosm and experimental microcosm. Microcosms were incubated at room temperature in amber glass bottles wrapped in tin foil. The experimental microcosms were treated with 100 ppm v/v crude oil. The oil, Norne Blend, was selected because we expected it would be similar in

composition to oil in natural reservoirs near the sampling site. When we sampled, there were no wells near the sampling site that were producing oil.

At 0 and 72 hours after the microcosms were prepared, 100 mL subsamples were extracted and immediately filtered through a 0.2 μm filter. DNA was extracted from the filters according to the standard protocol for the MoBio PowerWater DNA Isolation Kit. We amplified a subunit of the V4 region of the 16S rRNA gene following the procedure described in Preheim *et al.* [20]. Extracted DNA was amplified using custom barcoded primers and sequenced with paired-end 250 bp reads on an Illumina MiSeq.

16S data processing. Primers were trimmed from the forward and reverse reads by searching for the best-matching position in the read's first 20 bases, allowing a maximum of 2 mismatches between the primer sequence and the read sequence. Read pairs without matching forward and reverse primers were discarded. Reads were demulitplexed by assigning reads to the best-matching barcode sequence, allowing no more than 1 mismatched base. Reads with no acceptable barcode match were discarded. Reads were merged by (i) evaluating alignments that would produce merged reads of 263 ± 5 bases, (ii) selecting the alignment with the greatest number of matching bases, (iii) assigning consensus bases and quality scores according to Edgar & Flyvbjerg [21]. Merged reads with more than 2.0 expected errors were discarded. Taxonomic information was collected using the RDP classifier [23] and, in select cases, NCBI BLAST [24]. We searched for chimeras in the data by performing UCHIME with the Broad gold database [25], UCHIME with the RDP training database (version 9), and *de novo* UPARSE at 99%

identity [26]. The sequence data for these experiments are in Datasets S3 and S4 and at MG-RAST [27] under accession 4685010.3.

OTU calling. For most analysis of the ocean and sludge experiments, we use unique sequences as OTUs (i.e., these are 100% identity OTUs). For visual clarity, we used 99% *de novo* clustering with UPARSE for all samples in Figure 2. In Figure S1, we called OTUs using multiple methods. To call OTUs with RDP, we truncated every sequence's taxonomy at the first level that has less than 80% bootstrap confidence, and two sequences that have the same truncated taxonomy are placed in the same OTU. To call reference-based OTUs, we used the Greengenes reference database [28] (August 2013 97% OTUs) and global usearch [29] with minimum 97% identity.

Phylogenetic tree. The sequences shown on the tree are the 303 sequences most abundant in the four microcosm experiments. Sequences were aligned to the Greengenes core set [28] using Pynast [30]. One sequence (#229) that did not align to the core set was excluded. The tree was constructed with FastTree [31] and drawn with APE [32].

HMP samples

The Human Microbiome Project [33] samples were downloaded from the 16S rRNA trimmed data set (HM16STR). Reads from the stool sample (#700014956) and vaginal sample (#700016101) were trimmed to 200 bp. Only V3-V5 region reads were used.

RESULTS

Disparate dynamics in control and experimental microcosms

Figure 1 shows data from our paired-sample aquatic microcosm experiment. These data highlight some of the issues mentioned in the Introduction. One of the OTUs (OTU 3) increases to a high abundance in the experiment-after sample, potentially introducing a large compositional effect. Furthermore, that OTU's apparently dramatic dynamics in the experimental microcosm should be treated with skepticism because OTU 3 also increases in the control unit. In contrast, another OTU (OTU 63) was not detected in the two control samples or in the experiment-before sample, but it has a high abundance in the experiment-after sample, suggesting that little or none of its increase in the experimental microcosm should be attributed to bottle effects. How should we compare an OTU's abundances in the four samples? How should we correct results of the experimental unit with information from the control unit? We were motivated to develop a method based on the Poisson lognormal distribution to answer these questions.

Poisson lognormal distribution accurately models 16S abundances

We found that the truncated Poisson lognormal (TPL) distribution is an excellent fit for the abundance distributions of OTUs from multiple environments (Figure 2). To quantify the quality of the fit, we conducted an empirical test to see if the differences between the theoretical and observed abundances can be attributed to chance. For each sample shown in Figure 2, we fit the Poisson lognormal distribution to the sample's data, simulated 10 000 datasets (each with a number of OTUs equal to the number in the observed data) using the fit parameters, and compared the chi-square goodness-of-fit statistic in these

simulations to the goodness-of-fit in the observed data. In all cases, the differences between the theoretical and observed distributions are attributable to chance ($p = 0.76$, 0.91, 0.66 for the first three panels in Figure 2), although the attribution to chance was marginal in the sample obtained from a microcosm after treatment with oil ($p = 0.054$; "oiled ocean" in Figure 2). The TPL distribution's fit is similar when the OTUs were called with some other common OTU-calling methods (Figure S1).

Replicate units yield well-correlated results

To check that the TPL distribution can be used to quantify dynamics, we compared replicates from the bioreactor experiment, which subjects a microbial community to strong selective pressures. We expected that strong selective pressures would cause dramatic changes in microbial community composition but that these effects would be similar across replicates. Analytically, this means that we expect that Δz, a measure of an OTU's dynamics, should be similar across replicates. Visually, this means that, if the Δz values for all OTUs in two replicates are plotted against one another, they should fall along the $y = x$ diagonal. In fact, the replicated bioreactors show these sorts of well-correlated dynamics (Figure 3). Each plot provides an immediate summary of the relationship between the dynamics of all OTUs in the four samples in the experiment.

Identification of known and putative oil-degrading organisms

Having shown that metrics derived from the TPL distribution can be used to quantify dynamics, we analyzed the results of an aquatic microcosm experiment with one control microcosm, one experimental microcosm, and two timepoints (pretest and post-treatment

samples). In this experiment, ocean water was treated with crude oil to gain insight into the effects of a potential oil spill in this region. Previous work has shown that, in many aquatic environments, a few species (especially in the genera *Alcanivorax* and *Cycloclasticus*) multiply to make up the majority of microbes after crude oil is added [34, 35]. We aimed to identify OTUs in this ecosystem that respond to amendment with crude oil.

To identify OTUs with suggestive responses to oil, we selected criteria for Δz and ΔF that would be consistent with a response to oil and not growth due to bottle effects. Specifically, we considered OTUs whose Δz in the experimental unit was greater than or equal to its Δz in the control unit minus 0.5. Roughly speaking, this selected OTUs whose abundances increased more in the experimental unit relative to the control unit. We selected this cutoff criterion because it was compatible with our expectations about specific responses to oil and also included some OTUs with finite Δz values (Figure S2). Analogous to the fold-change cutoffs used to identify interesting spots on microarrays, this sort of cutoff criterion does not necessarily OTUs whose dynamics would be considered statistically significant in a well-powered experiment.

We also considered OTUs whose ΔF in the treatment unit was greater than 0.5 but whose ΔF in the control unit was less than 0.5. Roughly speaking, this selected for OTUs whose position in the TPL distribution moved *up* past about 50% of other OTUs in the experimental unit but whose position in the distribution moved *down* past about 50% of OTUs in the control unit. A plot of the ΔF values which highlights the OTUs that meet

the criteria are shown in Figure S3. The OTUs that satisfy either the Δz or ΔF criteria are shown in the context of a bacterial phylogeny in Figure 4. They group into five monophyletic clades. Detailed information about the OTU's taxonomic classification and dynamics are reported in Table 1.

The OTUs that appear to respond to oil are all members of γ-Proteobacteria (clades A through D) or α-Proteobacteria (clade E). Among the γ-Proteobacteria, many of these OTUs correspond to phylogenetic groups that contain known oil degraders: the two OTUs in clade C are both classified as *Alcanivorax*, the seven OTUs in clade B are classified as *Pseudomonas* or Pseudomonadaceae [36], and the one OTU in clade D is classified as *Methylophaga* [37, 38].

All but one of the eight OTUs in clade A are classified as *Maricurvus*. These seven OTUs align to NCBI entries for *Maricurvus nonylphenolicus* and *Aestuariicella hydrocarbonica*. The first species, *M. nonylphenolicus* is the *Maricurvus* type strain and degrades nonylphenol [39], while the second, *A. hydrocarbonica* has a 16S sequence highly similar to *Maricurvus* and degrades multiple aliphatic hydrocarbons [40].

The OTUs in clade E, which are members of α-Proteobacteria, are classified by RDP as Rhodobacteracea, and they align equally well to 16S sequences from *Phaeobacter*, *Roseobacter*, *Pelagimonas*, and *Sulfitobacter* spp. Although genus *Phaeobacter* has no known oil-degrading species, it may have increased in abundance in other experiments that amended ocean water with crude oil [41], and *Sulfitobacter* spp. were abundant in

oiled beach sands [35] and in a large microcosm simulating an oil spill in ocean water [42].

DISCUSSION

The truncated Poisson lognormal distribution and microbial ecology

As noted above, the TPL distribution has been used to model the abundance distribution of plants and animals, but to our knowledge this is the first report in which the TPL distribution is used to model microbial abundances collected in 16S surveys.

When we laid out the logic of the TPL distribution, we described the Poisson distribution as the link from the continuous-valued true abundance λ to the discrete number of sequencing reads. However, the number of 16S genes is not identical between organisms, and so our approach uses a single layer (a stochastic change from λ to the number of reads) to model a process that actually has two layers (a deterministic change from the organism's true abundance λ to the true abundance of its 16S gene, and from there to reads). The quality of the TPL distribution's fit to 16S abundance data suggests that variations in 16S copy number need not be separately included to explain the observed abundance distributions. This idea contrasts against the approach of Kembel *et al.* [17], who showed that the abundance distribution of organismal counts N, computed from an estimate of a taxon's 16S copy number C and its 16S sequence data counts $N \times C$, fits the lognormal better than does the abundance distribution of 16S counts $N \times C$. Our results suggest that the lognormal is simply a poor fit for discrete 16S count data. In the ecosystems we studied, compounding the lognormal with a random Poisson distribution is sufficient to make an excellent fit to 16S OTU abundance data.

Interpreting microbial dynamics

The information in Table 1 helps demonstrate that three quantifications of OTU dynamics—change in relative abundance, change in rescaled reads z, and change in cumulative distribution function value F—provide complementary information about the OTUs' dynamics in our experiment. In multiple cases, OTUs satisfied both the Δz and ΔF criteria. However, the ΔF criteria tend to identify organisms that are less abundant and have non-finite Δz values. Some OTUs underwent a small change in relative abundance but a large change in ΔF, indicating the abundance distribution for 16S sequences is so skewed that that small change in relative abundance is sufficient for it to advance dramatically relative to other OTUs. For example, OTU 63 in clade A experiences a small increase in relative abundance (+0.8%) in the experimental microcosm, but its ΔF (+0.996 of a possible 1.0) indicates that its small change in relative abundance made it more abundant than most of the other OTUs in the sample. In this case, a small change in relative abundance for a lowly-abundant OTU can equate to a large ΔF. Conversely, a large change in relative abundance for a highly-abundant OTU can equate to a small ΔF.

We speculate that the Poisson lognormal's fit to microbial community structure may have further relevance to making inferences about microbial community dynamics. The problem of limited replicates is not new, and difficulties in replication were probably more pronounced for microarray studies. One approach to limited replication in microarray studies was to infer the variance in each gene's expression level, potentially improving the power of gene-by-gene t-tests without requiring more replicates, using Bayesian hierarchical models [43]. Just as genes with higher expression level tend to

have higher variance, it may be that there are trends in the variance or dynamics of OTUs' relative abundances. It may be that the Poisson lognormal distribution could provide a lens for discovering those trends, which in turn might provide better inference for uncovering OTUs' dynamics if they are integrated in a productive way. It is important to again note that no method, no matter how sophisticated, can obviate the need for sufficient replication.

As mentioned in the Introduction, sequencing count data sets are compositional and are therefore subject to compositional effects. Without a quantification of overall or absolute bacterial abundance, metagenomic 16S datasets can provide evidence —but never proof— that a measured bacterial species changed in absolute abundance. However, if the Poisson lognormal distribution accurately models the distribution of OTUs' abundances, and if most OTUs do not change in absolute abundance, then the OTUs' positions with respect to the distribution may reflect their absolute abundances better than their relative abundances do. A well-powered experiment comparing absolute abundances and OTU's positions in the Poisson lognormal distribution could evaluate this possible relationship.

Possible identification of oil degraders

A simple microcosm experiment, coupled with our analytical technique, identified many OTUs that may have increased meaningfully in abundance in response to the amendment with crude oil. OTUs classified as Rhodobacteraceae had Δz values indicating a small bloom, suggesting that these organisms might be involved in oil degradation without being capable of degrading oil on their own. On the other hand, the large blooms of

OTUs classified as *Maricurvus*, along with the recent discovery of the closely-related oil degrader *Aestuariicella hydrocarbonica*, suggest that these organisms might be relevant to oil degradation in the Eastern Mediterranean.

Uses and limitations for this method

In our experiment, we had enough sample material for a few microcosms, and we could non-destructively subsample each microcosm through time. Ideally, we would have had access to much more sample material so we could run many replicate experiments and use standard statistical approaches to validate any observed bacterial community dynamics. In our constrained setup, we aimed to mitigate extraneous, confounding effects like time, bottle effects, or contact with a common set of microbes by incorporating information about the community composition dynamics from multiple timepoints. Experiments in which subsampling is less onerous than replications —as in, for example, experiments studying the microbiome of animals in a facility or experiments studying the effect of a treatment *in situ*— might benefit from quantitative correction of experimental results using the data obtained from pretests and control units.

We expect that this method might also be applicable to situations beyond the very constrained one that originally motivated us. If an experimental setup has many replicates, it may be that performing rigorous statistical analyses on the Poisson lognormal metrics, rather than just the relative abundances, yields more useful results. We also expect that the Poisson lognormal distribution might fit data from amplicon-based sequencing of other taxonomic marker genes, like eukaryotes' 18S rRNA gene.

ACKNOWLEDGEMENTS

This material is based on work supported by BP Exploration/MIT Energy Initiative under Grant No. 6926835, the National Science Foundation under Grant No. 0821391, and the National Science Foundation Graduate Research Fellowship under Grant No. 1122374. The authors are grateful to BP and its partners for support in the sampling effort.

Data Availability: All relevant data are within the paper and its Supporting Information files. The oil microcosm sequences are also available at MG-RAST under accession 4685010.3.

SUPPORTING DATASET LEGENDS

Dataset S1. Unique sequences from sludge bioreactor experiments. Trimmed, quality filtered, dereplicated sequences.

Dataset S2. Count data from sludge bioreactor experiments. Table showing number of times each sequence appeared in each sample (three replicates, two timepoints each).

Dataset S3. Unique sequences from oil microcosm experiments. Trimmed, merged, quality filtered, dereplicated sequences.

Dataset S4. Count data from oil microcosm experiments. Table showing number of times each sequence appeared in each sample (water from four depths, control and experimental "oil" microcosms, two timepoints each).

REFERENCES

1. Hammes F, Vital M, Egli T. Critical Evaluation of the Volumetric "Bottle Effect" on Microbial Batch Growth. Appl Environ Microbiol. 2010;76(4):1278-81. doi: 10.1128/aem.01914-09.

2. Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol. 2007;62(2):142-60. doi: 10.1111/j.1574-6941.2007.00375.x.

3. Anderson MJ, Willis TJ. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. Ecology. 2003;84(2):511-25. doi: 10.1890/0012-9658(2003)084[0511:CAOPCA]2.0.CO;2.

4. Clarke KR. Non-parametric multivariate analyses of changes in community structure. Austral Ecol. 1993;18(1):117-43. doi: 10.1111/j.1442-9993.1993.tb00438.x.

5. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001;26:32-46. doi: 10.1111/j.1442-9993.2001.01070.pp.x.

6. La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLoS ONE. 2012;7(12):e52078. doi: 10.1371/journal.pone.0052078.

7. Dufrene M, Legendre P. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol Monogr. 1997;67(3):345-66. doi: 10.2307/2963459.

8. De Cáceres M, Legendre P. Associations between species and groups of sites: indices and statistical inference. Ecology. 2009;90(12):3566-74. doi: 10.1890/08-1823.1.

9. Jackson DA. Compositional data in community ecology: The paradigm or peril of proportions? Ecology. 1997;78(3):929-40. doi: 10.1890/0012-9658(1997)078[0929:CDICET]2.0.CO;2.

10. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2011;8(9). doi: 10.1371/journal.pcbi.1002687.

11. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett. 2007;10(10):995-1015. doi: 10.1111/j.1461-0248.2007.01094.x.

12. Volkov I, Banavar JR, Hubbell SP, Maritan A. Patterns of relative species abundance in rainforests and coral reefs. Nature. 2007;450(7166):45-9. doi: 10.1038/nature06197.

13. Curtis TP, Sloan WT, Scannell JW. Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci USA. 2002;99(16):10494-9. doi: 10.1073/pnas.142680199.

14. Grundy PM. The expected frequencies in a sample of an animal population in which the abundances of species are log-normally distributed. Part I. Biometrika. 1951. doi: 10.2307/2332589.

15. Bulmer MG. On fitting the Poisson lognormal distribution to species-abundance data. Biometrics. 1974;30. doi: 10.2307/2529621.

16. Williamson M, J. Gaston K. The lognormal distribution is not an appropriate null hypothesis for the species–abundance distribution. J Anim Ecol. 2005;74(3):409-22. doi: 10.1111/j.1365-2656.2005.00936.x.

17. Kembel SW, Wu M, Eisen JA, Green JL. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Comput Biol. 2012;8(10):e1002743. doi: 10.1371/journal.pcbi.1002743.

18. Engen S, Lande R, Walla T, DeVries PJ. Analyzing spatial structure of communities using the two-dimensional poisson lognormal species abundance model. Am Nat. 2002;160(1):60-73. doi: 10.1086/340612.

19. Bastidas-Oyanedel J-R, Mohd-Zaki Z, Pratt S, Steyer J-P, Batstone DJ. Development of membrane inlet mass spectrometry for examination of fermentation processes. Talanta. 2010;83(2):482-92. doi: 10.1016/j.talanta.2010.09.034.

20. Preheim SP, Perrotta AR, Martin-Platero AM, Gupta A, Alm EJ. Distribution-based clustering: using ecology to refine the operational taxonomic unit. Appl Environ Microbiol. 2013;79(21):6593-603. doi: 10.1128/AEM.00342-13.

21. Edgar RC, Flyvbjerg H. Error filtering, pair assembly, and error correction for next-generation sequencing reads. Bioinformatics. 2015. doi: 10.1093/bioinformatics/btv401.

22. Techtmann SM, Fortney JL, Ayers KA, Joyner DC, Linley TD, Pfiffner SM, et al. The unique chemistry of Eastern Mediterranean water masses selects for distinct microbial communities by depth. PLoS ONE. 2015;10(3):e0120605. doi: 10.1371/journal.pone.0120605.

23. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261-7. doi: 10.1128/Aem.00062-07.

24. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24(16):1757-64. doi: 10.1093/bioinformatics/btn322.

25. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 2011;21(3):494-504. doi: 10.1101/gr.112730.110.

26. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Meth. 2013;10(10):996-8. doi: 10.1038/nmeth.2604.

27. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;9(1):1-8. doi: 10.1186/1471-2105-9-386.

28. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069-72. doi: 10.1128/aem.03006-05.

29. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1. doi: 10.1093/bioinformatics/btq461.

30. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 2010;26(2):266-7. doi: 10.1093/bioinformatics/btp636.

31. Price MN, Dehal PS, Arkin AP. FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5(3):e9490. doi: 10.1371/journal.pone.0009490.

32. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289-90. doi: 10.1093/bioinformatics/btg412.

33. Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012;486(7402):215-21. doi: 10.1038/nature11209.

34. Head IM, Jones DM, Röling WF. Marine microorganisms make a meal of oil. Nat Rev Microbiol. 2006;4(3):173-82. doi: 10.1038/nrmicro1348.

35. Kostka JE, Prakash O, Overholt WA, Green SJ, Freyer G, Canion A, et al. Hydrocarbon-degrading bacteria and the bacterial community response in Gulf of Mexico beach sands impacted by the Deepwater Horizon oil spill. Appl Environ Microbiol. 2011. doi: 10.1128/AEM.05402-11.

36. Palleroni NJ, Pieper DH, Moore ERB. Microbiology of Hydrocarbon-Degrading Pseudomonas. In: Timmis KN, editor. Handbook of Hydrocarbon and Lipid Microbiology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 1787-98.

37. Rivers AR, Sharma S, Tringe SG, Martin J, Joye SB, Moran MA. Transcriptional response of bathypelagic marine bacterioplankton to the Deepwater Horizon oil spill. ISME J. 2013;7(12):2315-29. doi: 10.1038/ismej.2013.129.

38. Mishamandani S, Gutierrez T, Aitken MD. DNA-based stable isotope probing coupled with cultivation methods implicates Methylophaga in hydrocarbon degradation. Front Microbiol. 2014;5. doi: 10.3389/fmicb.2014.00076.

39. Iwaki H, Takada K, Hasegawa Y. Maricurvus nonylphenolicus gen. nov., sp. nov., a nonylphenol-degrading bacterium isolated from seawater. FEMS Microbiol Lett. 2012;327(2):142-7. doi: 10.1111/j.1574-6968.2011.02471.x.

40. Lo N, Kim KH, Baek K, Jia B, Jeon CO. Aestuariicella hydrocarbonica gen. nov., sp. nov., an aliphatic hydrocarbon-degrading bacterium isolated from a sea tidal flat. Int J Syst Evol Microbiol. 2015;65(6):1935-40. doi: 10.1099/ijs.0.000199.

41. Al-Awadhi H, Dashti N, Khanafer M, Al-Mailem D, Ali N, Radwan S. Bias problems in culture-independent analysis of environmental bacterial communities: a representative study on hydrocarbonoclastic bacteria. SpringerPlus. 2013;2:369. doi: 10.1186/2193-1801-2-369.

42. Jung S, Park J, Kown O, Kang J-H, Shim W, Kim Y-O. Effects of crude oil on marine microbial communities in short term outdoor microcosms. J Microbiol. 2010;48(5):594-600. doi: 10.1007/s12275-010-0199-2.

43. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. Bioinformatics. 2001;17(6):509-19. doi: 10.1093/bioinformatics/17.6.509.

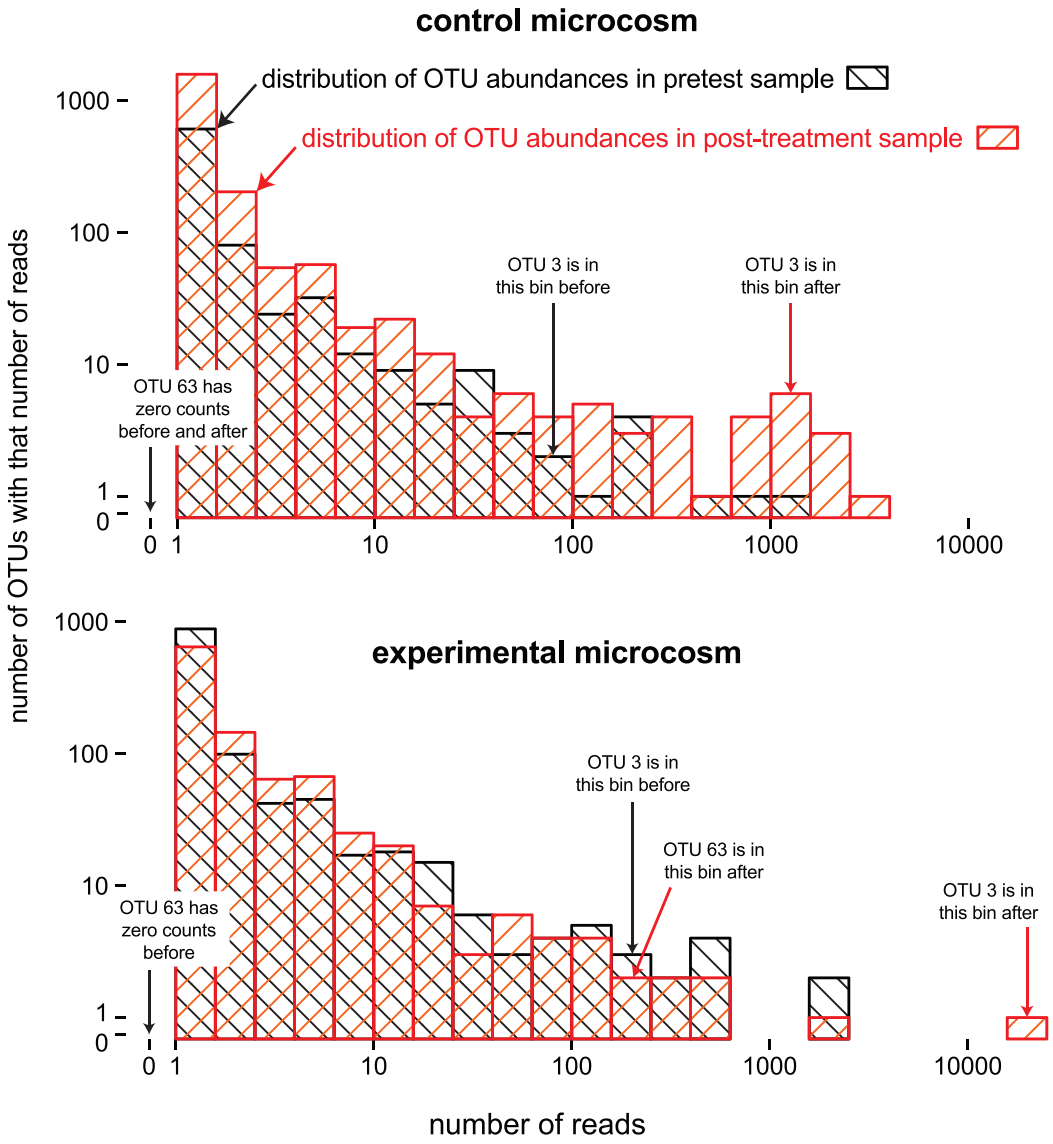


Figure 1. OTU dynamics and possible bottle effects in a paired-sample experiment. Four histograms are shown, one each for each sample in the experiment: control-before (above, black), control-after (above, red), experimental-before (below, black), experimental-after (below, red). Each histogram shows how many OTUs (logarithmic y-axes) have what number of associated reads (logarithmic x-axis). No bin is shown for OTUs with zero counts. The dynamics of two OTUs are shown: black arrows point to the abundance bin for OTUs in the "before" sample and red arrows point their abundance bins in the "after" samples.

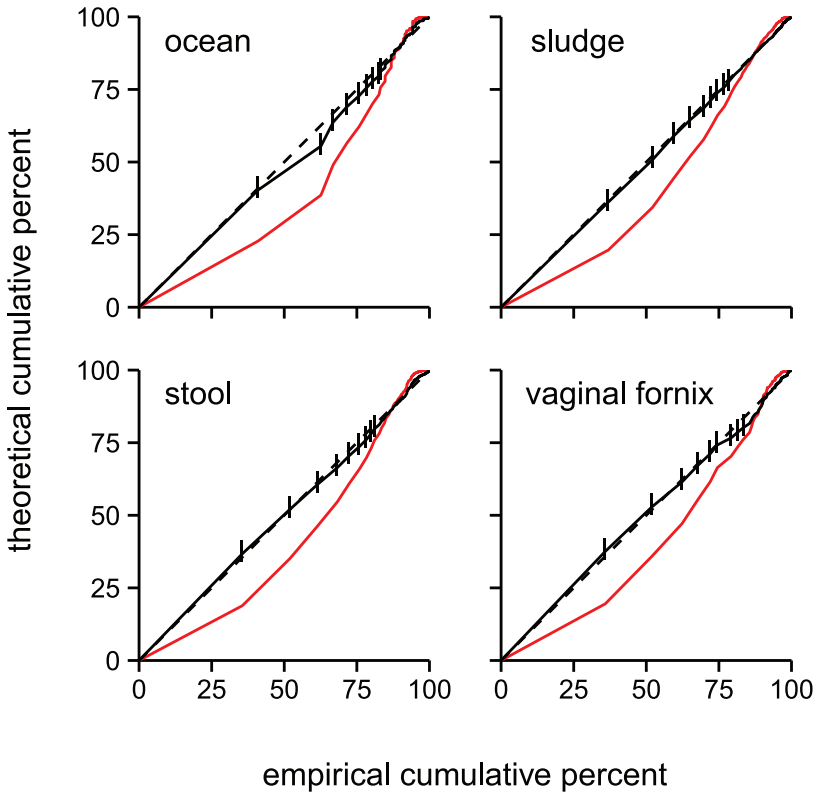


Figure 2. TPL distribution fits OTU abundance distributions in multiple ecosystems. Probability-probability plots comparing the empirical cumulative distribution function (horizontal axis) with the theoretical cumulative probability of a TPL distribution fit to each data set (vertical axis, black solid line). The first ten data points are marked with vertical dashes: the first dash (furthest lower left) represents the fraction of OTUs that have 1 read, the second dash represents the fraction of OTUs with 2 or fewer reads, and so forth. The dotted black line indicates a perfect fit of the TPL to the empirical distribution ($y = x$). The theoretical cumulative probability of a simple lognormal distribution (red line) is shown to emphasize the quality of the TPL fit. The ecosystems are ocean water from this study (top left), wastewater sludge from this study (top right), human stool (bottom left; Human Microbiome Project [HMP] sample), and human vagina (bottom right; HMP sample). 99% *de novo* OTUs are shown for all samples.

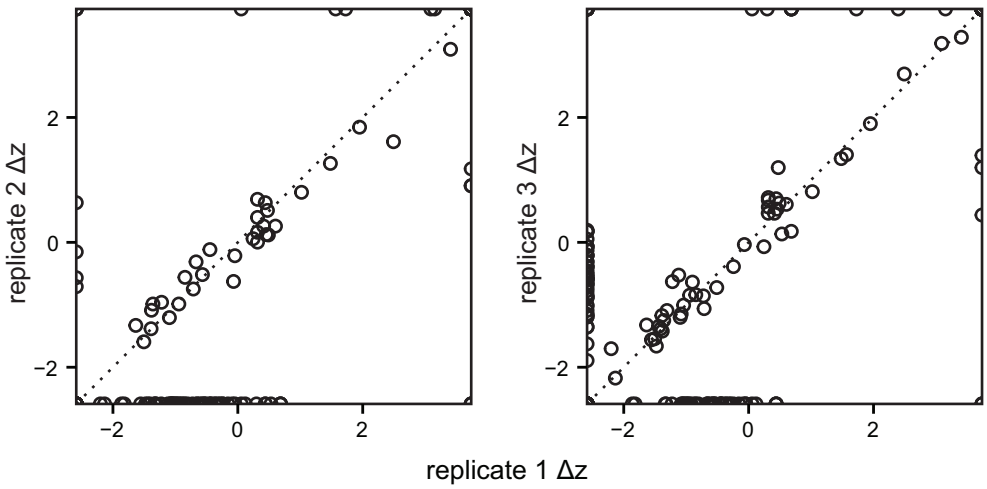


Figure 3. OTUs in replicate units have correlated dynamics. The dynamics of OTUs (circles) in three replicate bioreactors (replicate 1, x-axis; replicates 2 and 3, y-axes) inoculated with the same material and subjected to the same conditions. The dotted line ($y = x$) indicates a perfect correlation: an OTU on this line would have exactly the same Δz in both replicates, while deviations show differences in dynamics. For example, in the left plot, OTUs above the dotted line experienced a greater increase in abundance in replicate 2 than in replicate 1 (or, a smaller decrease in 2 than in 1), while OTUs below the line "grew more" in replicate 2 than in replicate 1 (or, "died less" in 2 than in 1). OTUs with infinite Δz are plotted on the plot's borders (e.g., the points in the lower-right corner of the first plot represent OTUs that have $\Delta z = +\infty$ in replicate 1 and $\Delta z = -\infty$ in replicate 2).

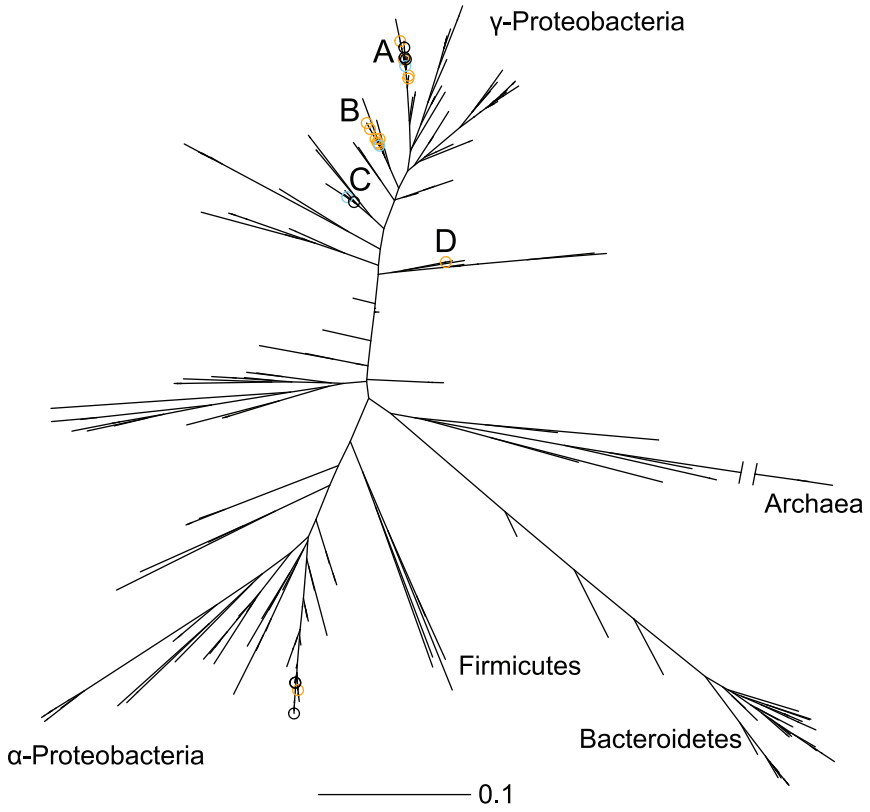
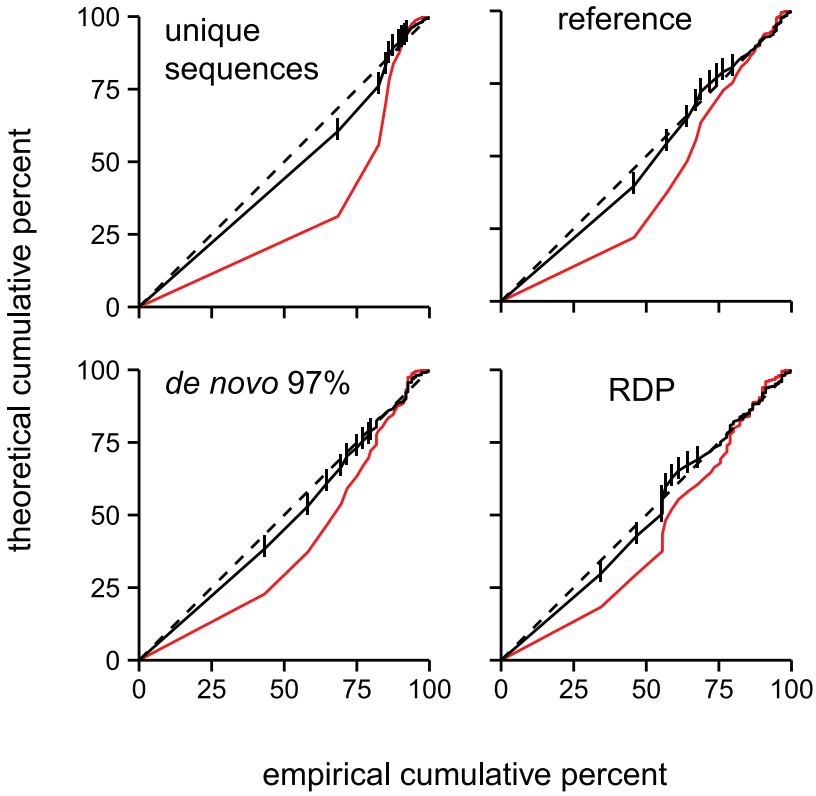


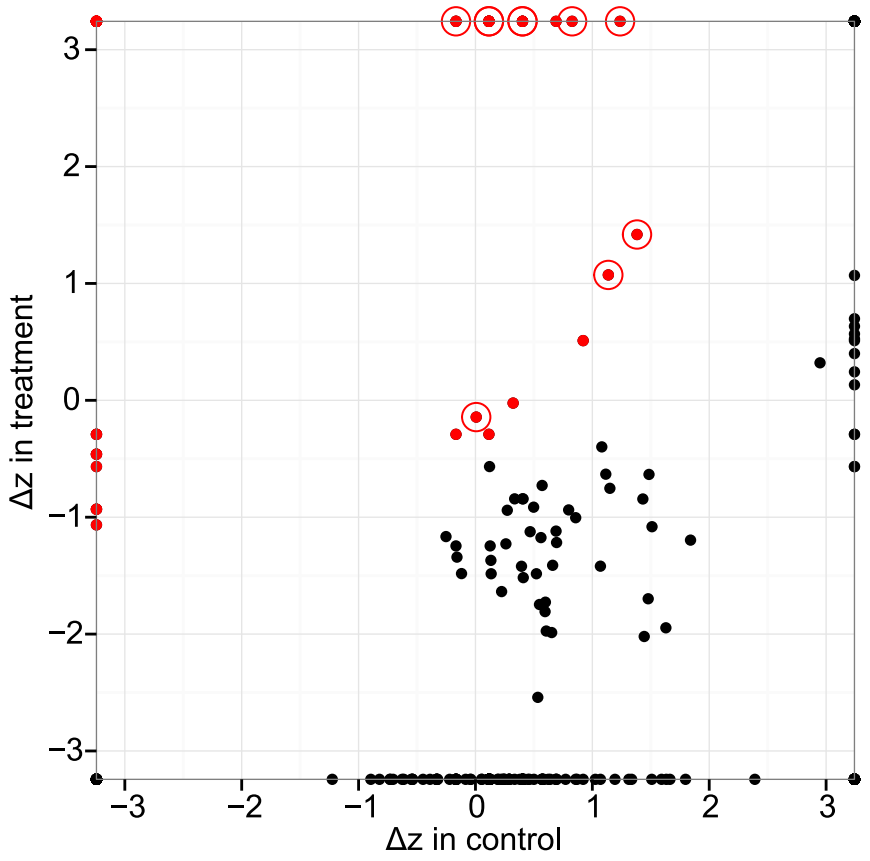
Figure 4. OTUs that respond to oil appear in five clades. On a phylogenetic tree built from the 16S sequences, organisms potentially responding to crude oil are marked with open circles. OTUs that satisfy the Δz criteria are marked with blue circles, OTUs that satisfy the ΔF criteria are marked with orange circles, and OTUs that satisfy both are marked with black circles. Information about the taxonomy and dynamics of these sequences are shown in Table 1. The five clades (A through E) are labeled, and select taxonomic groups are labeled to help orient the reader. The Archaea branch is truncated. Scale bar: substitutions per site.

Table 1. OTUs with dynamic behavior in response to amendment with oil (*on next page*). All OTUs that satisfied the Δz or ΔF criteria are listed. The first three columns show taxonomy. The most specific RDP taxonomic classification with at least 80% bootstrap support is shown. The next two columns indicate whether the OTU satisfied the Δz criteria, the ΔF criteria, or both. The next six columns show the changes in relative abundance (Δr.a.), rescaled reads z, and cumulative distribution function F in the control ("ct") and experimental ("ex") units. The value $\Delta z =$ n.a. is shown for OTUs that had zero counts at both timepoints in that microcosm; $\Delta z = \infty$ is shown for OTUs had zero counts before the treatment and more than zero counts after the treatment.

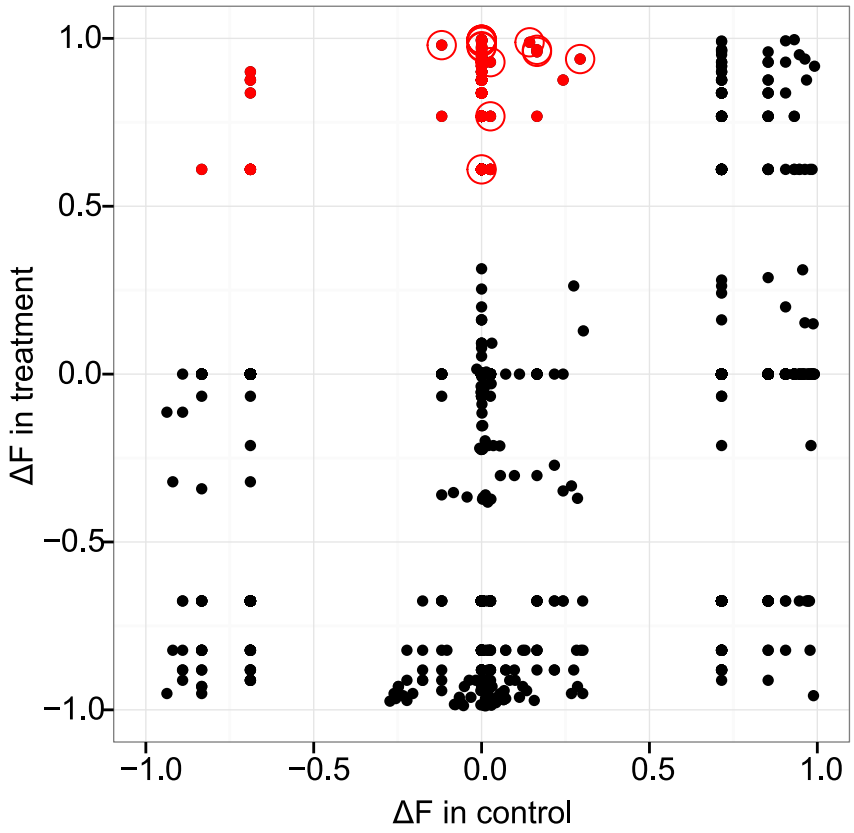
			criteria		Δr.a.		Δz		ΔF		
Clade	Classification	support	Δz	ΔF	ct	ex	ct	ex	ct	ex	OTU ID
A	*Maricurvus*	0.95	*		0.033	0.739	1.382	1.419	−0.0002	0.0041	3
	Maricurvus	0.87		*	0	0.008	n.d.	∞	0	0.9958	63
	γ-Proteobacteria	1		*	0	0.0033	n.d.	∞	0	0.9942	107
	Maricurvus	0.87		*	0	0.003	n.d.	∞	0	0.9939	111
	Maricurvus	0.96	*	*	0	0.0013	0.827	∞	0.1429	0.9883	119
	Maricurvus	0.97	*		−0.0002	0.0002	0.115	∞	0.0263	0.9292	256
	Maricurvus	0.9	*		−0.0001	0.0005	0.402	∞	0.1649	0.9659	262
	Maricurvus	0.94		*	0	0.0007	n.d.	∞	0	0.9767	291
B	*Pseudomonas*	1	*		0.0027	0.0629	1.137	1.073	0.014	0.0063	14
	Pseudomonas	0.99		*	0	0.0142	n.d.	∞	0	0.9961	42
	Pseudomonas	0.91		*	0	0.0112	n.d.	∞	0	0.996	53
	Pseudomonaceae	0.82		*	0	0.0045	n.d.	∞	0	0.995	88
	Pseudomonas	0.99		*	0	0.0025	n.d.	∞	0	0.9931	120
	Pseudomonas	0.91		*	0	0.0018	n.d.	∞	0	0.9913	167
	Pseudomonas	1		*	0	0.0016	n.d.	∞	0	0.9903	174
C	*Alcanivorax*	1	*	*	−0.0001	0.0004	0.402	∞	0.1649	0.9599	206
	Alcanivorax	1	*		−0.0007	−0.0001	0.007	−0.143	−0.0143	0.0148	270
D	*Methylophaga*	1		*	0	0.0006	n.d.	∞	0	0.9739	210
E	Rhodobacteraceae	1	*	*	−0.0003	0.0008	−0.167	∞	−0.1188	0.9799	104
	Rhodobacteraceae	1	*	*	0.0003	0.0003	1.237	∞	0.2935	0.9384	105
	Rhodobacteraceae	1	*	*	−0.0002	0.0001	0.115	∞	0.0263	0.7678	226
	Rhodobacteraceae	1		*	0	0	n.d.	∞	0	0.6098	288



S1 Fig. TPL fits OTUs called by different methods. Probability-probability plots comparing the empirical cumulative distribution function (x-axis) with the theoretical cumulative probability of a TPL distribution fit to the distribution of OTUs computed using different OTU-calling methods (y-axis, black line). This is same ocean sample as in Figures 1 and 2. The first ten data points are marked with vertical dashes: the first dash (furthest lower left) represents fraction of OTUs with 1 read, the second dash represents the fraction of OTUs with 2 or fewer reads, and so forth. The dotted black line indicates a perfect fit ($y = x$). The theoretical cumulative probability of a simple lognormal distribution fit to each OTU distribution (red) is shown to emphasize the quality of the TPL fit. The methods are unique sequences (i.e., 100% identity OTUs; top left), 97% reference-based OTUs from Greengenes (top right), *de novo* 97% OTUs (bottom left), and genus-level OTUs computed with RDP (bottom right). The empirical goodness-of-fit test described in the main text yields $p = 0.35, 0.40, 0.44, 0.41$ for these data.



S2 Fig. OTU dynamics measured by Δz. Each OTU present in the microcosm experiment described in the main text is shown. OTUs that meet the Δz criterion described in the text (Δz in treatment $>$ Δz in control -0.5) are in red. OTUs that meet the criterion and are among the 304 most abundant sequences in the four microcosm experiments (i.e., those shown as blue dots in Figure 4) are circled. OTUs with an undefined Δz value in either microcosm are not shown, while OTUs with infinite Δz values are shown at the border of the figure (e.g., an OTU with $\Delta z = +\infty$ in the control unit and $-\infty$ in the experimental unit would be shown in the lower-left corner).



S3 Fig. OTU dynamics measured by ΔF. Each OTU present in the microcosm experiment described in the main text is shown. OTUs that meet the ΔF criteria described in the text (ΔF in treatment > 0.5; ΔF in control < 0.5) are in red. OTUs that meet those criteria and are among the 304 most abundant sequences in the four microcosm experiments (i.e., those shown as red dots in Figure 4) are circled.

Chapter 3

Surveys, simulation, and single-cell assays relate function and phylogeny in a lake ecosystem

The contents of this chapter were accepted as: Sarah P. Preheim*, Scott W. Olesen* (these authors contributed equally), Sarah J. Spencer, Arne Materna, Charuleka Varadharajan, Matthew Blackburn, Jonathan Friedman, Jorge Rodríguez, Harold Hemond and Eric J. Alm. (2016) *Nature Microbiology*.

The figures are at the end of the chapter. The Supplementary Information is in Appendix A.

Summary paragraph:

Much remains unknown about what drives microbial community structure and diversity. Highly structured environments might offer clues. For example, it may be possible to identify metabolically similar species as groups of organisms that correlate spatially with the geochemical processes they carry out. Here, we use a 16S ribosomal RNA gene survey in a lake that has chemical gradients across its depth to identify groups of spatially correlated but phylogenetically diverse organisms. Some groups had distributions across depth that aligned with the distributions of metabolic processes predicted by a biogeochemical model, suggesting these groups performed biogeochemical functions. A single-cell genetic assay showed, however, that the groups associated with one biogeochemical process, sulfate reduction, contained only a few organisms that have the genes required to reduce sulfate. These results raise the possibility that some of these spatially-correlated groups are consortia of phylogenetically diverse and metabolically different microbes that cooperate to carry out geochemical functions.

Keywords: ecology / 16S / function / gradient / phylogeny / lake / co-occurrence

Introduction

Explaining the vast diversity of microbes found in many ecosystems [1,2] is a challenge for microbial ecology. Environments with chemical or other abiotic gradients like temperature have been a key resource for studying microbial ecology. For example, studies in Winogradsky columns [3], microbial mats [4], mine drainage sites [5], hydrothermal vents [6], and dimictic lakes [7] have provided insight about the relationships between environmental parameters, microbial diversity, and ecosystem functions. Microbial surveys with spatial scales comparable to those of the ecosystem gradients can identify groups of spatially-correlated organisms and relate the distribution of those organisms to the environmental gradients.

There are challenges to interpreting the relationship between organisms in spatially-correlated groups and environmental information. First, the relationship between an organism's spatial distribution and environmental parameters can be complicated. For example, a naïve expectation might be that sulfate-reducing organisms are abundant where sulfate concentrations are highest. In fact, the distribution of sulfate-reducing organisms also depends on the distribution of more favorable electron acceptors and the transport of sulfur compounds around the ecosystem. Even more subtly, bacterial populations may be capable of performing multiple metabolisms, and they can even be simply inactive. Thus, there is a need to develop techniques that provide quantitative expectations about factors that shape organismal distributions given observed environmental information.

A second challenge is that there are multiple experimental methods that can verify the relationships between function and phylogeny, but most of these methods are *in vitro* or perturb

the environment [8]. A method that relates phylogeny and function without perturbing the natural ecosystem would clarify the *in situ* functional relationships between organisms in a spatially-correlated group. Deep metagenomic sequencing along with differential genome binning techniques can produce draft genomes from complex communities [9] but is expensive and cannot target specific functions.

A third challenge to studying spatially-correlated organisms in ecosystems with gradients is relating the groups' diversities to their environmental functions, especially if these organisms are unrelated. Organisms in these groups could use similar resources, as it is know that many traits are widespread in the tree of life [10], or could have recently exchange genes through horizontal gene transfer [11]. Unrelated organisms with similar distributions could also be found together because they are part of multispecies, symbiotic associations [12]. The challenge lies in differentiating between these or other possibilities.

In this paper, we investigate spatially-correlated organisms in an ecosystem with gradients. First, we conducted a microbial survey of a dimictic lake. Second, we constructed a quantitative, dynamic biogeochemical model that shows how bacteria can drive the creation of chemical gradients. Third, we show that there are many groups of spatially-correlated organisms in this lake and relate those groups to the biogeochemical model. Finally, we used a single-cell assay to investigate the functional capabilities of the groups of spatially-correlated bacteria related to one modeled process, sulfate reduction. We show that, taken together, these results raise the possibility that these spatially-correlated groups are multispecies, symbiotic associations of microbes, that is, consortia [13].

Results

Community structure is influenced by geochemistry. We performed our study in Upper Mystic Lake, a dimictic, eutrophic freshwater lake outside Boston, MA because this lake is seasonally stratified and supports complex microbial communities that catalyze well-characterized biogeochemical cycles [14-19]. The seasonal stratification means that the deepest three-quarters of this lake is anoxic, supporting fewer predators that can complicate microbial distribution patterns, and that the deeper parts of the lake are relatively isolated from external inputs.

To characterize microbial diversity, we conducted amplicon-based bacterial surveys (16S rRNA gene library from DNA samples) along a vertical transect in the lake, collecting samples at approximately each meter of depth. We grouped 16S rRNA gene sequences into operational taxonomic units (OTUs) using the ecologically-informed distribution-based clustering algorithm, which merges sequences from related organisms that have similar spatial distributions [20]. We also measured major geochemical parameters (temperature, specific conductivity, dissolved oxygen, nitrate, iron and sulfate; Fig. S1-S3).

The 16S rRNA gene survey showed that biogeochemistry had a major influence on the bacterial community structure (Fig. 1). Transitions in community structure lined up with the lake's major geochemical features: the thermocline, oxycline, and nitrocline (Fig. 1b). Cyanobacteria were most abundant near the surface (Fig. 1a). Bacteroidetes, Actinobacteria, and Proteobacteria were abundant across depths. δ-Proteobacteria, which include most of the known sulfate-reducing bacteria, were abundant only below the nitrocline where more favorable terminal electron

acceptors were exhausted. The differences in geochemistry and bacterial community structure across depths suggest that organisms exploiting similar resources should have correlated distributions across depths.

Multiple groups of spatially-correlated OTUs in the lake. To identify groups of spatially-correlated organisms, we used hierarchical clustering to group the 536 most abundant OTUs into 49 groups based on the similarity of the OTUs' distributions across depths. We call these groups *operational ecological units* (OEUs) because they are groups of organisms that we expect have functional or ecological relationships (thus "ecological") but were defined in a purely statistical way (thus "operational" [21]).

Most OEUs contained OTUs from multiple phyla. OEUs ranged in size from 2 to 33 OTUs and contained 1 to 10 phyla. The number of phyla in each OEU (0.34 additional phyla per OTU beyond the first in the OEU; Fig. S4) was about as many as would be expected if phyla classifications had been randomly assigned to OTUs (0.35 +/– 0.01; 1000 permutations). To verify that the OEUs are robust to different bioinformatic methods, extraction methodologies, and sample years, we compared the results of the OEU analysis after varying each of these factors and found more OTUs together than would be expected by chance (Table S5).

A biogeochemical model reproduces chemical and biological structure and dynamics. Having characterized the lake's geochemistry and identified groups of spatially-correlated organisms, we set out to design a computational framework that predicts the function of bacteria in the lake. We found no existing dynamical model that treated all the major microbial metabolic

processes in a dimictic lake, so we modified and reinterpreted a model of chemical transport and microbial metabolism designed to simulate groundwater aquifers [22]. We chose to develop a model because the distribution of bacteria is the result of complex and interdependent biogeochemical cycles and hydrodynamic transport processes in the lake.

The model we developed simulates the major chemical species, redox cycles, and transport processes in the hypolimnion (Fig. S5; Tables S1-S3). We used previously published values [22] for many parameters (Table S4) and calibrated the model to match the chemical datasets (Figs. 2, S2-S3). In this lake, the water is typically well-mixed through the lake's depth in spring. During summer, warmer water sits on top of the cooler water at the bottom of the lake. Thermal resistance to mixing across this warm-cold plane (the thermocline, about 5 m deep at the time of sampling [Fig. 1b]) partially isolates water below the thermocline (the hypolimnion) from external and atmospheric influences. Heterotrophic microbes oxidize energy-rich carbon compounds as they diffuse or settle down into the hypolimnion, and the increasingly limited availability of terminal electron acceptors for these microbes leads to vertical chemical gradients. Reduced chemical species can be transported to oxidizing conditions closer to the lake's surface, fueling additional microbial activity. The model predicts the distribution of microbial metabolic processes and chemical species abundance in the lake from spring to autumn, when the thermocline breaks down and the lake mixes again.

We used the model to simulate the lake's biogeochemical dynamics for two datasets: a time series collected in 2013 and a single-time point survey collected in 2008. In both cases, the model predicted chemical dynamics (Figs. 2, S6) that were consistent with those expected from a

eutrophic, dimictic lake [19]. In 2013, we had a time series covering the five months before the bacterial survey, so we initialized the model using the observed chemical parameters from the first survey in the time series. The chemical dynamics predicted by the model (Fig. 2) accorded with our measured time series, and the predicted distribution of chemical species accorded with the final survey (Fig. S2).

Because we had no initial data for the 2008 survey, we initialized the lake in a homogenous composition, as would be expected from an idealized dimictic lake that perfectly mixed throughout its entire depth in the spring. In this case, the model predicted the emergence of chemical gradients from the initially homogenous composition (Fig. S6), and the predicted distribution of chemical species accorded with the single-timepoint survey (Fig. S3).

To relate the output of the model with our biological data, we reinterpreted the modeled rates as predictions of microbial distribution. Implicit biomass models, like the one we developed, predict the rates of processes catalyzed by all microbes performing that process and assume that the biomass of the microbial community equilibrates quickly to the changing chemical environment [22]. They also assume that "everything is everywhere" and are not constrained by ecological processes like dispersion. We therefore expected that the relative rate of a modeled process should be proportional across depths to the biomass of microbes performing that process. This interpretative framework, which we call "inferred biomass", reinterprets implicit biomass models as hypotheses about microbial community structure. Consistent with these assumptions, our model largely reproduced the distribution of key organisms known to perform the corresponding metabolisms in 2013 (Fig. 3) and 2008 (Fig. S7).

The model captures the major patterns in the lake's chemical dynamics, but there are discrepancies between the model and observation. For example, in the final 2013 survey, oxygen concentrations reach undetectable levels at about 5 meters, increase until about 8 meters, then decrease again until reaching the detection limit at 14 meters. In contrast, the model predicts that oxygen concentrations would decrease monotonically with depth.

Many groups of spatially-correlated organisms have spatial distributions that correspond to modeled processes. 63% of the OEUs from the 2013 dataset have a spatial profile that is similar (distance less than 0.25) to one or more of the biogeochemical processes simulated in the model (Fig. S8), and some of these OEU-process pairings are supported by the previously reported ecosystem functions of one or more of the OTUs in the OEU (Table S6). This spatial alignment between OEUs and modeled processes suggests that the growth of organisms in those OEUs is dependent on the energy provided by those processes. Because these OEUs are made up of OTUs that are spatially correlated, taxonomically diverse, and spatially aligned with modeled biogeochemical processes, it may be that these OEUs are consortia of organisms in syntrophic relationships.

Not all taxa corresponding to a modeled process have the same metabolism. There are other explanations for the properties of the spatially-correlated groups. Aside from consortia, they may also be groups of functionally redundant bacteria or simply groups of organisms subject to some pressure or process that only coincidentally led to spatial alignment with one of the modeled biogeochemical processes. To distinguish these explanations, we assayed the genetic capability

of the OTUs in some OEUs to carry out the biogeochemical process with which they spatially align. For example, if all OTUs in an OEU have the genetic capability to perform some process, then that OEU might represent functionally redundant organisms. Conversely, if none of those OTUs have the genetic capability, then that OEU probably has little to do with biogeochemistry. If only some OTUs in an OEU can perform some process, then that OEU might represent a consortium of syntrophic organisms.

We investigated one process, sulfate reduction, in greater detail because the spatial distribution of this process had one of the best matches between the model and observation and because there was a well-studied genetic marker for this function. The three OEUs with spatial distributions that best matched this process (Fig. 5) contained 14 OTUs that were in high abundance in both this survey and the positive control for the gene fusion assay described below. Among these 14 OTUs, 6 are classified as δ-Proteobacteria, the class that contains most of the bacteria known to reduce sulfate, and one of the δ-Proteobacteria OTUs corresponds to a known sulfate-reducing organism (Table S6). Among the other OTUs, 5 are classified as Bacteroidetes, which contains no known sulfate reducers and are instead regarded as specialists in the degradation of high molecular weight organic matter [23]. Because terminal oxidation processes of organic carbon under anaerobic conditions are rarely catalyzed by a single organism, we suspected that the sulfate reducers among the δ-Proteobacteria might be in a syntrophic relationship with the Bacteroidetes organisms, which provide low-molecular weight dissolved organic carbon to sulfate reducers. Intriguingly, the reference OTU for sulfate reduction (a clone similar to *Desulfatirhabdium butyrativorans*; Table S6) and an OTU classified as Bacteroidetes (with 93% identity to the 16S rRNA of the sugar-fermenting psychrophile *Prolixibacter bellariivorans* [24])

appeared together in the same OEU in both the 2008 and 2013 datasets, suggesting that, if some OEUs do represent consortia of syntrophic organisms, some of those association might persist across years in this ecosystem.

To probe the genetic capability of OTUs to perform sulfate reduction, we targeted a gene, dissimilatory sulfite reductase gene (*dsrB*), whose product is a key enzyme in sulfate reduction [25]. Specifically, we used a single-cell gene-fusion technique [26] that amplifies the 16S rRNA gene only in organisms whose genomes contain *dsrB*. The technique traps cells in polyacrylamide beads during DNA extraction, isolates the extraction products in oil droplets, and amplifies a concatenation of the *dsrB* and 16S sequences using within-droplet PCR. As a control, we performed a non-specific fusion assay to verify that a wide range of taxonomic marker sequences can be amplified with this method (Fig. 5).

The single-cell assay amplified 16S-*dsrB* amplicons whose 16S rRNA gene sequences corresponded to a small number of OTUs. Only 4 OTUs that appear in any of the OEUs were amplified by this technique, and the OEUs that contain them were identified as putative sulfate-reducing groups (Fig. 5). These results imply that the genomes of the organisms corresponding to these 4 OTUs contain *dsrB* and that the genomes of the rest of the organisms in these OEUs do not. In the first of the putative sulfate-reducing OEUs, the most abundant OTUs and two other OTUs appeared to have the genetic capacity to reduce sulfate. In the second OEU, only one OTU (about one-third as abundant as the most abundant OTU in the OEU) appeared to have this capacity. In the third OEU, no OTUs appeared to be capable of reducing sulfate.

These results raise the possibility that those two OEUs represent consortia of syntrophic organisms cooperating to carry out sulfate reduction. We therefore checked if any other OEUs contained organisms with known mutualistic associations. We identified two cases where OTUs within an OEU are likely part of a consortium. In the first case, one OEU contained an OTU that was 97.7% identical to the ammonia-oxidizing bacterium *Nitrosospira briensis* [27] and an OTU 99% identical to a nitrite-oxidizing enrichment culture clone *Candidatus* Nitrotoga arctica [28]. Nitrification is a two-step process, typically carried out by different organisms [29,30], so it is likely that these two organisms interact to carry out nitrification. In the second case, one OEU that aligns with the modeled distribution of a methane oxidation metabolism contained an OTU 94.8% identical to *Methylobacter tundripaludum* (a methane oxidizer) and an OTU 98% identical to a strain of *Methylotenera versatilis* (a non-methane-oxidizer methylotroph). A study using stable isotope-labeling concluded that these organisms cooperate during methane oxidation [31].

Discussion

Our approach combined field observations, quantitative modeling, and a single-cell genetic approach to relate taxonomic diversity in survey data to ecosystem-level functions. Our results suggest that there are previously unknown consortia in the lake whose members work together to carry out major environmental processes for at least some part of the year. Previous research has studied functions of organisms containing the same functional genes or able to incorporate the same labeled compounds. In contrast, our observational approach addressed multiple processes at the same time and did not require perturbing the environment. However, because we investigated a process, sulfate reduction, that showed a strong match between the model and observations, the results we observed should not necessarily be treated as representative of what would result from studying any of the processes.

Our analysis began by defining and studying OEUs, which are a type of ecological network. Previous studies have asserted ecological network associations between organisms in surveys based on co-occurrence patterns (i.e., mutual presence or absence) in space or time [32-35]. OEUs are the result of a different measure of "co-occurrence", since we require that OTUs co-occur at similar abundances to be placed into the same OEU. Although presence-absence patterns derived from sequencing data can be affected by technical issues like sequencing depth [36], we demonstrated that our grouping of OTUs was robust to technical issues of sample preparation method or OEU calling algorithms and provides richer interpretations of survey data.

Like other studies that investigate ecological interactions between microbial taxa, however, our results provide hypotheses about such interactions but do not prove that they exist. The inferred

ecological association between members of OEUs could be verified by experimental techniques like stable-isotope probing [37], FISH-NanoSIMS [38], or MAR-Fish [39]. Because these relationships may change as the conditions in the lake change throughout the season, the identification of many pairs of OTUs in the same OEU in both 2008 and 2013 is somewhat surprising. Thus, the co-occurring pairs of organisms in the same OEU in 2008 and 2013 are strong candidates to target for further investigation.

Instead of perturbative experimental techniques, we used a biogeochemical model to predict the spatial distribution of microbial metabolisms and hypothesized the function of OEUs whose spatial distributions have the same position and shape as the ones predicted by the model. Implicit biomass models in other ecosystems could be reinterpreted as inferred biomass models, which would allow survey data to be used as validation for the spatial distribution of microbial activities predicted by models or, conversely, predicted activities to be used to generate hypotheses about the function of microbes identified in surveys. The results are limited, however, because the link between OEUs and modeled biogeochemical processes was only inferential.

Despite the model's utility, there are discrepancies between its predictions and the observed chemical and biological data. We attribute these discrepancies to multiple causes. First, our model only simulates the underlying biogeochemical processes and does not account for other ecological and physiological factors that determine organismal distributions. For example, the reference OTU for methane oxidation is the most abundant methanotroph in the oxic region, but its distribution does not match the predicted distribution of methane oxidation, suggesting that

the organism performs other metabolisms or that methane oxidation is not well-described by the model. Second, our model is relatively simple and does not simulate many processes that are known to be part of lake's biogeochemistry, including complex carbon substrate utilization profiles. We aimed to create a model that captured the broad spatial patterns and temporal dynamics with the minimum number of processes possible, thus balancing the model's completeness with its simplicity. For example, the observed oxygen minimum at the thermocline is unusual but not rare for dimictic lakes. There are many competing explanations for this type of minimum (e.g., changes in temperature, predator abundance, or horizontal mixing [19]) that are all beyond the scope of the model, and so it fails to predict that minimum. Third, the model was designed to model the lake's general seasonal dynamics rather than its behavior in the specific season when we conducted the survey. For the purposes of this study, understanding the lake's general organizing processes was more important than understanding the dynamics in a particular year. Although these discrepancies limit the interpretability of hypotheses about OTUs' function that are generated by the biogeochemical model, we showed that additional bioinformatic and experimental evidence can together provide a more complete picture.

Our discovery of well-correlated organisms that probably cooperate to carry out biogeochemical functions raises exciting ecological questions. If some of the OEUs do represent consortia of syntrophic organisms, are the organisms that compose them physically associated because they inhabit similar particulate matter? What roles do microbes play within these consortia? Are these interspecific associations constant over time, or do they "reset" when the lake mixes in the winter? We expect that our combined framework of surveys, modeling, and single-cell genetics will be useful for *in situ*, non-perturbative identification of potential ecological interactions in

other microbial ecosystems, painting a picture of a microbial world filled with complex, interlocking relationships.

Methods

Sample collection (2012-2013). Water samples were collected at Upper Mystic Lake (Medford, MA), from one location in the middle of the lake (~42 26.155N, 71 08. 961W) where the total water depth is 23 meters. Water samples were collected in 2012 (October 2) and 2013 (March 26, May 10, June 17, July17, and August 15). Water samples were collected at approximately one- to two-meter intervals through 25 meters of plastic Tygon tubing using a peristaltic pump. Two volumes of water at each depth were pumped through the tubing before 50 mL was filtered through an in-line Swinnex filter holders onto sterile 0.22 μm filters (Millipore, Billerica, MA) and the filtrate collected in a 50 mL conical tube. Filters and filtrate were placed on dry ice immediately and transported back to the laboratory where they were placed at −80 °C until processing. To determine the influence of contamination from the tubing, sampling method, and carryover from the previous depth, we collected blanks by pumping 2 L of sterile water through the tubing before and after sampling. Blanks were distinct from other samples. One mL of both filtered and unfiltered water was put into a 1.5 mL microcentrifuge tube with 43 μL of concentrated HCl and placed in the dark during transport back to the lab for ferrous (Fe^{2+}) and total iron analysis, respectively. Samples were stored at −20 °C until iron was measured.

Sample collection (2008). The methods for collecting from Upper Mystic Lake on August 13, 2008 are described elsewhere [20]. Water was collected from Upper Mystic Lake (same location) on August 13, 2008 using a peristaltic pump and plastic Tygon tubing. Tubing was lowered to a point ~1 m from the bottom, running the pump in reverse to prevent water from entering the tubing until the appropriate depth was reached. Water from depth was allowed to flow through the tubing for 5 minutes before 14 mL were collected into a 15 mL sterile falcon tube and

immediately placed on dry ice. The first sample was taken from 22 meters depth and subsequent samples were taken every meter until 3 meters depth, then at 1.5 meters depth and at the surface. Samples were transported on dry ice and stored at −80 °C until processing about one year later.

Water conditions and chemistry. A Hydrolab minisonde (Hach Hydromet, Loveland, CO) was attached to the end of the tubing to record dissolved oxygen, temperature, and specific conductance during deployment. Nitrate, sulfate and chloride were measured by ion chromatography at the University of New Hampshire Water Quality Analysis Laboratory. Iron was measured by a modified ferrozine protocol [16,40]. Values for other chemical species used in the model but not directly measured were manually interpolated from previous measurements.

DNA extraction. DNA from half of the 2012 samples and all the 2013 samples was extracted with PowerWater DNA extraction kit (MoBio, USA). The manufacture's protocol was followed, except for the addition of proteinase K and alternative lysing protocol at 65 °C (MoBio Laboratories, Inc., Carlsbad, CA). Briefly, filters were sterilely transferred into the PowerWater Bead Tube and 20 µl proteinase K was added before incubating at 65 °C for 10 minutes. Tubes were vortexed on a horizontal MoBio vortex adapter. Proteins and inhibitors were removed with PW2 and PW3 before adding supernatant to Spin filter for column purification. After two washing steps, DNA was eluted with PW6 and used in PCR analyses. Samples from 2008 and the other half of the 2012 samples were extracted with the Qiagen DNeasy Blood and Tissue Kit (Qiagen, USA), as previously described [20,41]. Of the 2012 samples, 9 were prepared in duplicate, one replicate per extraction method.

Illumina library construct design. The Illumina library was created with a two-step protocol in order to add cluster binding and sequencing primer sites to the construct in the second round of PCR amplification. First step PCR amplification primers (PE16S_V4_U515_F, 5′–ACACG ACGCT CTTCC GATCT YRYRG TGCCA GCMGC CGCGG TAA–3′ and PE16S_V4_E786_R, 5′–CGGCA TTCCT GCTGA ACCGC TCTTC CGATC TGGAC TACHV GGGTW TCTAA T–3′) contain primers U515F and E786R targeting the V4 region of the 16S rRNA gene, as described previously [20,42]. Additionally, a complexity region in the forward primer (5′–YRYR–3′) was added to aid image-processing software used during Illumina next-generation sequencing. The second-step primers incorporate the Illumina adapter sequences and a 9-bp barcode for library recognition (PE-III-PCR-F, 5′-AATGA TACGG CGACC ACCGA GATCT ACACT CTTTC CCTAC ACGAC GCTCT TCCGA TCT–3′; PEIII-PCR-001-096, 5′–CAAGC AGAAG ACGGC ATACG AGATN NNNNN NNNCG GTCTC GGCAT TCCTG CTGAA CCGCT CTTCC GATCT–3′, where N indicates sample barcode position). Libraries from 2008 were created with the primer skipping protocol, as previously described [41].

Illumina library preparation and sequencing. Real-time PCR was done to ensure uniform amplification and avoid over-cycling. Both real-time and first-step PCRs were done similarly to the manufacture's protocol for Phusion polymerase (New England BioLabs, Ipswich, MA). Samples were cycled with the following conditions: denaturation at 98 °C for 30 sec annealing at 52 °C for 30 sec and extension at 72 °C for 30 sec for 40 cycles. Samples were divided into four 25-µl technical replicate reactions during both first- and second-step cycling reactions and cleaned using Agencourt AMPure XP-PCR purification (Beckman Coulter, Brea, CA). Paired-end sequencing was performed at Massachusetts Institute of Technology BioMicro Center

(BMC) on an Illumina MiSeq with 250 bases for each the forwards and reverse reads and 8 base indexing read. Non-standard Illumina indexing primers were used to initiate sequencing from just after the sequencing primer binding site for the barcode sequence (anti-reverse BMC index primer; 5′– AGATC GGAAG AGCGG TTCAG CAGGA ATGCC GAGAC CG –3′). To improve base-calling efficiency, 25% phiX control was added to the sample during sequencing.

Raw data processing and OTU calling. Raw sequence data was demultiplexed and quality filtered using custom scripts (github.com/almlab/SmileTrain). Overlapping paired end reads were merged and sequences were pre-clustered with USEARCH [43]. Sequences were aligned to a subset of the Silva alignment [44] with mothur [45], and OTUs were called with distribution-based clustering [20] with default parameters. Reads were trimmed to 102 bases before OTU calling, which was sufficient to capture the differences between key populations while still ensuring high quality base calls. Sequences were checked for chimeras using UCHIME [46] and sequences not aligned to the Silva reference database were discarded. Sequences from 2008 were processed as previously described [20]. Sequencing of the 2008 library was not long enough for paired end reads to overlap, so only the forward read was used. Sequences were trimmed to 76 nucleotides before OTU calling.

Community analysis. OTUs were classified with RDP [47]. The phylum level assignments referenced in various analyses throughout the paper are from RDP. The dendrogram of Jensen-Shannon divergences was produced using Ward's method (R version 3.2.2). The phylogenetic tree of OTU sequences, aligned to the Silva database as mentioned above, was calculated using

PhyML [48] with GTR model (estimated as best model), 0 invariant sites, 6 rate categories (estimated).

Calling operational ecological units (OEUs). To call OEUs, OTUs were initially filtered by abundance in all samples. OTUs making up less than 0.25% of counts across all samples were excluded from the OEU analysis, leaving 536 OTUs. Next, counts from technical replicate samples at the same depth were pooled, and OTUs that were abundant in the no-template negative control samples (i.e., more than 10% of that OTU's reads mapped to the negatives) were excluded. The sequence counts for each OTU were converted to relative abundance (i.e., the number of counts corresponding to each OTU in a sample was divided by the sum of all counts in that sample). Every OTU was then converted to a profile (i.e., the relative abundance of each OTU in a sample was divided by the sum of that OTU's relative abundances in all samples). The square of the Euclidean distance between OTU profiles was used as the dissimilarity metric in a hierarchical clustering analysis (Ward's method). The cluster dendrogram was cut to produce 50 candidate OEUs. OEUs were trimmed for quality as follows: if an OEU had at least one OTU with a mean Pearson correlation with the other OTUs in the cluster of less than 0.75, the OTU with the lowest mean correlation was removed from the cluster, and the filtering was repeated. OEUs with fewer than two member OTUs were excluded from further analysis. Every member OTU had a mean correlation of at least 0.75 with the other OTUs in the OEU. Of the original 536 OTUs and 50 candidate OEUs, 491 OTUs (92% of initial OTUs) in 49 OEUs (98% of initial OEUs) remained after quality filtering. Scripts for OEU calling are available at github.com/almlab/oeu.

Quantifying OEU reproducibility related to OEU number parameter. The OEU-calling algorithm requires a parameter: the number of initial OEUs at which to cut the cluster dendrogram. Figure S9 shows a comparison of the results produced by this OEU-calling algorithm when the dendrogram is cut to produce different numbers of candidate OEUs. In the main analysis, 50 OEUs were used because:

1. The choice of the number of initial OEUs represents a trade-off between two types of errors. A Type I error (i.e., the incorrect assertion that two OTUs are ecologically related) occurs more often with a smaller number of OEUs, while a Type II error (i.e., the incorrect assertion that two OTUs are unrelated) occurs more often with a large number of OEUs. The analyses presented here depend on OEUs correctly identifying true ecological associations, so avoiding Type I errors is more important, so a relatively large number of OEUs is appropriate.
2. Increasing the number of initial OEUs increases the number of OTUs included in the final analysis (since fewer OTUs are excluded by the final quality-control step) and decreases the within-OEU variance, so a relatively large number of initial OEUs is appropriate.
3. Increasing the number of initial OEUs decreases the size of each OEU: it causes them to contain fewer OTUs. To keep a high enough number of OTUs per OEU to generate hypotheses to test in the single-cell assay, a number of OEUs much greater than 50 would have been inappropriate.

Quantifying OEU reproducibility across time points. To compare the OEU composition across years, OEU compositions in the 2013 data, presented above, was compared to OEU compositions for corresponding 2008 data. To call OEUs on the 2008 dataset,

- OTUs that were abundant in the sample nearest the lake bottom (those with more than 5% of their reads from 23 meters depth) were excluded,
- OTUs that were abundant in the negative samples (those with more than 10% of their reads in the two blank samples) were excluded,
- low-abundance OTUs (those with less than 0.25% of all reads) were excluded, and
- the same OEU-calling methodology presented above was used.

Quantifying OEU reproducibility across DNA extraction and sequencing methodologies. To call OEUs on the duplicate 2012 samples prepared using two DNA extraction methodologies, the same exclusion criteria as for the 2008 data were used (except that OTUs with less than 0.1% of all reads were considered low-abundance). Sequences in the 76 bp dataset were matched to the 102 bp dataset by searching for exact sequence matches of the shorter sequence within the longer dataset. If multiple 76 bp OTUs matched the same 102 bp OTU, only the most abundant OTU was kept as the corresponding OTU. Not all OTUs were represented in both datasets, so some OTUs did not have a corresponding OTU in the other dataset.

Quantifying OEU reproducibility across OEU calling methodologies. To compare the effects of different OEU-calling algorithms, OEU calling was performed as described for the main 2013 dataset except replacing the Euclidean distance (i.e., L2 norm) with the L1 distance (i.e., Bray-Curtis).

Statistical methodology quantifying OEU reproducibility. To quantify the reproducibility of the OEUs between timepoints or between sample preparation methods, we computed the numbers of pairs of OTUs such that:

- both OTUs are present in both datasets (e.g., in both 2008 and 2013 datasets), and
- both OTUs were in the same OEU in both datasets (e.g., OTUs *A* and *B* were both in OEU *X* in the 2008 dataset and both in OEU *Y* in the 2013 dataset).

We compared this number of pairs against the number of pairs satisfying the same criteria that would arise at random, specifically, if we randomly shuffled the abundances of OTUs in each sample before computing the Euclidean distance between OTUs.

Reference OTU selection. Reference OTUs were selected by matching the Illumina OTUs to Sanger clone sequences. Only exact matches between the 77 bp Illumina OTUs and Sanger clones were considered. Three Illumina OTUs matched multiple Sanger clones with nucleotide distances between clones larger than 0.1 and resulted in OTU distributions that were the product of two distinct organismal signals. These were corrected by aligning Sanger clone sequences to identify discriminating bases 5′ of the Illumina OTU sequence end point. One or two differentiating bases were identified for each of the three cases and the length of sequence required to differentiate between the two sequences was determined. Once a unique sequence was identified to discriminate the different clones in the Illumina data, a count of the discriminating sequence across libraries was generated from the raw data expressed as a percent of total reads. This replaced the previously merged OTU for populations 16, 141 and 125 (OTU IDs).

To gain functional information for the most abundant OTUs, we generated a Sanger-sequenced clone library to provide more phylogenetic information for the shorter Illumina OTUs. To make the Sanger-sequenced clone library, 16S rRNA sequences were amplified from DNA extracted from the 6 meter and 21 meters samples with Phusion polymerase (New England Biolabs, Ipswich, MA) and 27F and 1492R primers [42]. PCR products were cloned into the pCR Blunt II plasmid with the Zero Blunt TOPO PCR cloning kit (Invitrogen, Carlsbad, CA) and sequenced in at least one direction with Sanger sequencing (Genewiz, South Plainfield, NJ). Longer Sanger sequences were assigned the functional capabilities of the best BLAST hit [49] to a type strain or genome sequence. To verify expected profiles for each process in the biogeochemical model, we selected a set of nine reference OTUs involved in the modeled biogeochemical processes. These reference OTUs were among the 100 most abundant OTUs and had sequences that matched longer 16S rRNA sequences from clone libraries sequenced with Sanger sequencing developed from the lake samples.

Functions, OTU IDs, full genome matches, and accessions for reference OTUs are shown in Table S6. The matching Sanger sequences for the reference OTUs corresponded to organisms with metabolisms characterized by genomic analysis or *in vitro* experiments. Reference OTUs had spatial distributions in the lake that were consistent with their purported metabolism.

Biogeochemical model. The biogeochemical model (almlab.mit.edu/mystic.html), inspired by Hunter *et al.* [22], was run with Matlab (version 8) and supporting Python scripts (version 2.7). Details on the mechanics, implementation, and parameter values are in the Supplementary

Information. Briefly, the water under the thermocline is modeled as 17 linked compartments, one per meter depth. Within each compartment, a minimal set of abstracted chemical species interconvert through a minimal set of modeled primary and secondary redox reactions (Fig. S5, Tables S1-S3). Primary oxidation rates follow a formulation informed by the relative favorability of electron acceptors. Secondary oxidation rates follow simple mass action rate forms. Chemical species are transported between adjacent compartments via bulk diffusion (for all species) and settling (for biomass and oxidized iron). The outside world is modeled by constant source terms: oxygen and biomass are added in the uppermost compartment (at the thermocline), while methane is added in the lowermost compartment (at the sediment). The resulting set of ordinary differential equations is solved numerically.

We intended to model the general distribution of chemical and biological species in the lake. Because the model is conceptual, it includes many simplifications compared to the aquifer model. First, transport is modeled compartment-by-compartment, using ordinary differential equations rather than partial differential equations. We greatly reduced the number of simulated chemical species (from 25 to 9). Many simulated chemical species consist of multiple chemical species found in nature (e.g., the modeled oxidized sulfur species includes hydrogen sulfide H_2S, bisulfide HS^-, and sulfide S^{2-}; there is only one modeled carbon species). Other chemical species found in nature are not treated in the model (e.g., elemental sulfur S^0 and all manganese compounds) because less is known about their importance to the lake's biogeochemistry. The primary redox reactions are borrowed almost exactly from the aquifer model (excepting some parameter changes and the removal of manganese as an electron acceptor). The secondary redox reactions are similar to those in the aquifer model (excepting some parameter changes, the

removal of some reactions, and the addition of iron oxidation on nitrate). Precipitation-dissolution, acid dissolution, and adsorption reactions relevant in the groundwater system were part of the original aquifer model but were not simulated here. Further details about these alterations of the original model are included in the Supplementary Information.

Comparing OEUs and biogeochemical processes. We asserted that certain OEUs are related to certain modeled biogeochemical processes by comparing the spatial distributions of OEUs and modeled processes and by manual bioinformatic inference. Average Euclidean distance to each process for all OTUs within an OEU was calculated (Fig. S8). OEUs containing the reference OTUs were chosen to represent each modeled process because existing literature about the reference OTU suggests that those OTUs perform that modeled process. To assign an OEU to a process, we further required that the imputed process be one of the two processes least distant from that OEU, as described above, except for methane oxidation on oxygen, which is not within the lowest two distances for that OEU (see Discussion).

Linking taxonomic marker sequences with a functional gene. Data for 16S rRNA gene fusion products with both selective (*dsrB*) and non-selective (barcode) sequences experiments was obtained from a previous analysis [26]. Briefly, seven mL of water from both the 2 meter and 21 meter samples on August 12, 2013 was added to 7 mL of 50% glycerol (25% final concentration) to preserve membrane integrity for single-cell techniques, immediately placed on dry ice and stored at −80 °C. 16S rRNA gene sequences were fused to a 20 base pair droplet barcode to control for effects of the protocol on limiting diversity. In a separate reaction, 16S rRNA gene sequences were fused to a portion of the diagnostic gene for dissimilatory sulfite reduction

(*dsrB*) to probe for functional information. Cells are trapped in 10 μm diameter polyacrylamide beads [50]. Poisson statistics predict that only 0.45% of beads will contain more than one cell. The DNA trapped within the beads is used as the template for PCR inside an emulsion [51]. The first set of primers for 16S rRNA gene amplification include U515F and 1492R, and the fusion reaction is nested within the 16S gene using E786R. The *dsrB* gene primers were adapted from Wagner *et al.* [52] and slightly modified to fit the needs of the molecular construct. Sequences and the results of a traditional *dsrB* survey are provided in the original publication. The *dsrB* gene is highly conserved across known sulfate reducers [53], but it is possible that there are variants of the gene that are prevalent in the lake that these primers did not amplify, in which case the following analysis would contain false negatives (i.e., OTUs that can reduce sulfate but did not produce 16S-*dsrB* amplicons). Comparison of the *dsrB*-16S rRNA gene fusion assay to a bulk *dsrB* gene survey in the original previous analysis demonstrate significant overlap [26], showing the fusion PCR assay targets a wide variety of reductive *dsrB* genes from the *δ-Proteobacteria dsrB* supercluster.

Nucleotide sequence accession numbers. All clone sequences were submitted to GenBank (accession no. KC192376 to KC192544). Illumina data were submitted to the Sequence Read Archive under study accession number PRJNA217938.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374 and by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC0008743.

References

1 Pace, N. R. A Molecular View of Microbial Diversity and the Biosphere. *Science* **276**, 734-740, doi:10.1126/science.276.5313.734 (1997).

2 Wilmes, P., Simmons, S. L., Denef, V. J. & Banfield, J. F. The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev* **33**, 109-132, doi:10.1111/j.1574-6976.2008.00144.x (2009).

3 Rundell, E. A. *et al.* 16S rRNA Gene Survey of Microbial Communities in Winogradsky Columns. *PLoS ONE* **9**, e104134, doi:10.1371/journal.pone.0104134 (2014).

4 Ward, D. M., Ferris, M. J., Nold, S. C. & Bateson, M. M. A Natural View of Microbial Biodiversity within Hot Spring Cyanobacterial Mat Communities. *Microbiol Mol Biol Rev* **62**, 1353-1370 (1998).

5 Bier, R. L., Voss, K. A. & Bernhardt, E. S. Bacterial community responses to a gradient of alkaline mountaintop mine drainage in Central Appalachian streams. *ISME J* **9**, 1378-1390, doi:10.1038/ismej.2014.222 (2015).

6 Schrenk, M. O., Kelley, D. S., Delaney, J. R. & Baross, J. A. Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. *Appl Environ Microbiol* **69**, 3580-3592, doi:10.1128/aem.69.6.3580-3592.2003 (2003).

7 Pinel-Alloul, B. & Ghadouani, A. in *The Spatial Distribution of Microbes in the Environment* (eds Rima B. Franklin & Aaron L. Mills) 203-310 (Springer Netherlands, 2007).

8 Neufeld, J. D., Wagner, M. & Murrell, J. C. Who eats what, where and when? Isotope-labelling experiments are coming of age. *ISME J* **1**, 103-110, doi:10.1038/ismej.2007.30 (2007).

9 Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533-538, doi:10.1038/nbt.2579 (2013).

10 Martiny, A. C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J* **7**, 830-838, doi:10.1038/ismej.2012.160 (2013).

11 Klein, M. *et al.* Multiple lateral transfers of dissimilatory sulfite reductase genes between major lineages of sulfate-reducing prokaryotes. *J Bacteriol* **183**, 6028-6035, doi:10.1128/Jb.183.20.6028-6035.2001 (2001).

12 Orphan, V. J. *et al.* Comparative analysis of methane-oxidizing archaea and sulfate-reducing bacteria in anoxic marine sediments. *Appl Environ Microbiol* **67**, 1922-1934, doi:10.1128/aem.67.4.1922-1934.2001 (2001).

13 Madigan, M. T., Martinko, J. M., Dunlap, P. V. & Clark, D. P. *Brock Biology of Microorganisms*. 12th edn, (Pearson/Benjamin Cummings, 2009).

14 Varadharajan, C. *Magnitude and spatio-temporal variability of methane emissions from a eutrophic freshwater lake* PhD thesis, Massachusetts Institute of Technology, (2009).

15 Varadharajan, C. & Hemond, H. F. Time-series analysis of high-resolution ebullition fluxes from a stratified, freshwater lake. *J Geophys Res Biogeosci* **117**, doi:10.1029/2011jg001866 (2012).

16 Senn, D. B. *Coupled arsenic, iron, and nitrogen cycling in arsenic-contaminated Upper Mystic Lake* PhD thesis, Massachusetts Institute of Technology, (2001).

17 Senn, D. B. & Hemond, H. F. Nitrate controls on iron and arsenic in an urban lake. *Science* **296**, 2373-2376, doi:10.1126/science.1072402 (2002).

18 Peterson, E. J. R. *Carbon and electron flow via methanogenesis, SO42-, NO3- and Fe3+ reduction in the anoxic hypolimnia of Upper Mystic Lake* Masters thesis, Massachusetts Institute of Technology, (2005).

19 Wetzel, R. G. *Limnology*. 3rd edn, (Academic Press, 2001).

20 Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A. & Alm, E. J. Distribution-Based Clustering: Using Ecology To Refine the Operational Taxonomic Unit. *Appl Environ Microbiol* **79**, 6593-6603, doi:10.1128/aem.00342-13 (2013).

21 Jax, K. Ecological Units: Definitions and Application. *Q Rev Biol* **81**, 237-258, doi:10.1086/506237 (2006).

22 Hunter, K. S., Wang, Y. F. & Van Cappellen, P. Kinetic modeling of microbially-driven redox chemistry of subsurface environments: coupling transport, microbial metabolism and geochemistry. *J Hydrol* **209**, 53-80, doi:10.1016/S0022-1694(98)00157-7 (1998).

23 Thomas, F., Hehemann, J.-H., Rebuffet, E., Czjzek, M. & Michel, G. Environmental and gut Bacteroidetes: the food connection. *Front Microbiol* **2**, doi:10.3389/fmicb.2011.00093 (2011).

24 Holmes, D. E., Nevin, K. P., Woodard, T. L., Peacock, A. D. & Lovley, D. R. Prolixibacter bellariivorans gen. nov., sp. nov., a sugar-fermenting, psychrotolerant anaerobe of the phylum Bacteroidetes, isolated from a marine-sediment fuel cell. *Int J Syst Evol Microbiol* **57**, 701-707, doi:10.1099/ijs.0.64296-0 (2007).

25 Leloup, J., Quillet, L., Oger, C., Boust, D. & Petit, F. Molecular quantification of sulfate-reducing microorganisms (carrying dsrAB genes) by competitive PCR in estuarine sediments. *FEMS Microbiol Ecol* **47**, 207-214, doi:10.1016/S0168-6496(03)00262-9 (2004).

26 Spencer, S. J. *et al.* Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *ISME J* **10**, 427-436, doi:10.1038/ismej.2015.124 (2016).

27 Purkhold, U., Wagner, M., Timmermann, G., Pommerening-Roser, A. & Koops, H. P. 16S rRNA and amoA-based phylogeny of 12 novel betaproteobacterial ammonia-oxidizing isolates: extension of the dataset and proposal of a new lineage within the nitrosomonads. *Int J Syst Evol Microbiol* **53**, 1485-1494, doi:10.1099/ijs.0.02638-0 (2003).

28 Alawi, M., Lipski, A., Sanders, T., Pfeiffer, E.-M. & Spieck, E. Cultivation of a novel cold-adapted nitrite oxidizing betaproteobacterium from the Siberian Arctic. *ISME J* **1**, 256-264, doi:10.1038/ismej.2007.34 (2007).

29 Canfield, D. E., Kristensen, E. & Thamdrup, B. in *Advances in Marine Biology* Vol. 48 (eds D. E. Canfield, E. Kristensen, & B. Thamdrup) 205-267 (Academic Press, 2005).

30 Costa, E., Perez, J. & Kreft, J. U. Why is metabolic labour divided in nitrification? *Trends Microbiol* **14**, 213-219, doi:10.1016/J.Tim.2006.03.006 (2006).

31 Beck, D. A. C. *et al.* A metagenomic insight into freshwater methane-utilizing communities and evidence for cooperation between the *Methylococcaceae* and the *Methylophilaceae*. *PeerJ* **1**, e23, doi:10.7717/peerj.23 (2013).

32 Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat Rev Micro* **10**, 538-550, doi:10.1038/nrmicro2832 (2012).

33 Eiler, A., Heinrich, F. & Bertilsson, S. Coherent dynamics and association networks among lake bacterioplankton taxa. *ISME J* **6**, 330-342, doi:10.1038/ismej.2011.113 (2012).

34 Gies, E. A., Konwar, K. M., Beatty, J. T. & Hallam, S. J. Illuminating Microbial Dark Matter in Meromictic Sakinaw Lake. *Appl Environ Microbiol* **80**, 6807-6818, doi:10.1128/aem.01774-14 (2014).

35 Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci USA* **108**, 4578-4585, doi:10.1073/pnas.1000081107 (2011).

36 Horner-Devine, M. C. *et al.* A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology* **88**, 1345-1353, doi:10.1890/06-0286 (2007).

37 Radajewski, S., Ineson, P., Parekh, N. R. & Murrell, J. C. Stable-isotope probing as a tool in microbial ecology. *Nature* **403**, 646-649, doi:10.1038/35001054 (2000).

38 Dekas, A. E. & Orphan, V. J. Identification of diazotrophic microorganisms in marine sediment via fluorescence in situ hybridization coupled to nanoscale secondary ion mass spectrometry (FISH-NanoSIMS). *Methods Enzymol* **486**, 281-305, doi:10.1016/S0076-6879(11)86012-X (2011).

39 Nielsen, J. L. & Nielsen, P. H. in *Handbook of Hydrocarbon and Lipid Microbiology* (ed Kenneth N. Timmis) (Springer-Verlag, Berlin, 2010).

40 Stookey, L. L. Ferrozine–a new spectrophotometric reagent for iron. *Anal Chem* **42**, 779-781, doi:10.1021/ac60289a016 (1970).

41 Blackburn, M. C. *Development of new tools and applications for high-throughput sequencing of microbiomes in environmental or clinical samples* Masters thesis, Massachusetts Institute of Technology, (2010).

42 Lane, D. J. in *Nucleic acid techniques in bacterial systematics* (eds E. Stackebrandt & M Goodfellow) 125-175 (John Wiley & Sons, 1991).

43 Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Meth* **10**, 996-998, doi:10.1038/nmeth.2604 (2013).

44 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590-D596, doi:10.1093/nar/gks1219 (2013).

45 Schloss, P. D. *et al.* Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl Environ Microbiol* **75**, 7537-7541, doi:10.1128/aem.01541-09 (2009).

46 Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-21200, doi: 10.1093/bioinformatics/btr381 (2011).

47 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267, doi:10.1128/aem.00062-07 (2007).

48 Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).

49 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

50 Tamminen, M. V. & Virta, M. P. J. Single gene-based distinction of individual microbial genomes from a mixed population of microbial cells. *Front Microbiol* **6**, doi:10.3389/fmicb.2015.00195 (2015).

51 Turchaninova, M. A. *et al.* Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* **43**, 2507-2515, doi:10.1002/Eji.201343453 (2013).

52 Wagner, M. *et al.* Functional marker genes for identification of sulfate-reducing prokaryotes. *Method Enzymol* **397**, 469-489, doi:10.1016/S0076-6879(05)97029-8 (2005).

53 Wagner, M., Roger, A. J., Flax, J. L., Brusseau, G. A. & Stahl, D. A. Phylogeny of dissimilatory sulfite reductases supports an early origin of sulfate respiration. *J Bacteriol* **180**, 2975-2982 (1998).

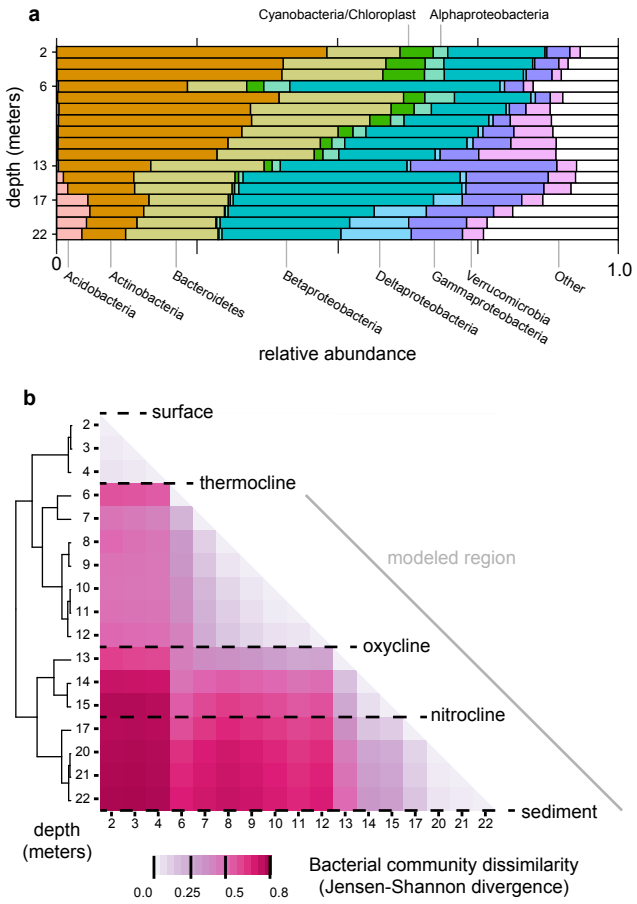


Figure 1. Bacterial survey of the lake identified communities that vary with depth. **a** The relative abundance of the seven most abundant phyla at four representative depths. Proteobacteria is divided into the classes α-, β-, γ-, and δ-Proteobacteria. (ε-Proteobacteria were not abundant [$< 0.5\%$ at every depth] and are not shown.) **b** Each square shows the dissimilarity between bacterial communities at two depths (e.g., the lower-left square shows the dissimilarity between the samples from the surface and from 22 meters depth). Major features (dotted lines) of the lake are noted: the thermocline, where the temperature gradient is steepest; the oxycline, where dissolved oxygen falls to 0.3 mg/L; the nitrocline, where nitrate concentration falls below detection; and the sediment at the lake bottom. The biogeochemical model treats the region below the thermocline (gray line).

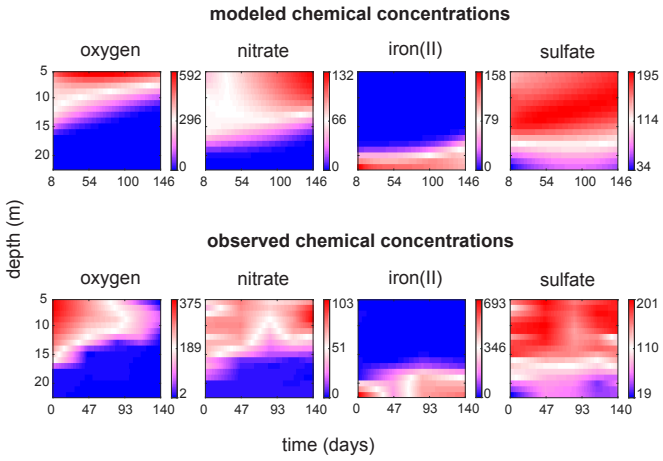


Figure 2. The model creates a dynamic picture of chemical changes that occur in the lake through the lake's depth (vertical axis) across time (horizontal). The model predicts changes in chemical species (top row; colorbar scales are μM) which are consistent with the observed chemical dynamics within the lake in 2013 (bottom row; colorbar scales are μM; interpolated from five timepoints for oxygen, nitrate and sulfate and interpolated from four time points for iron). The model was initiated from the observed conditions in March 2013. Only a subset of chemical species included in the model are shown.

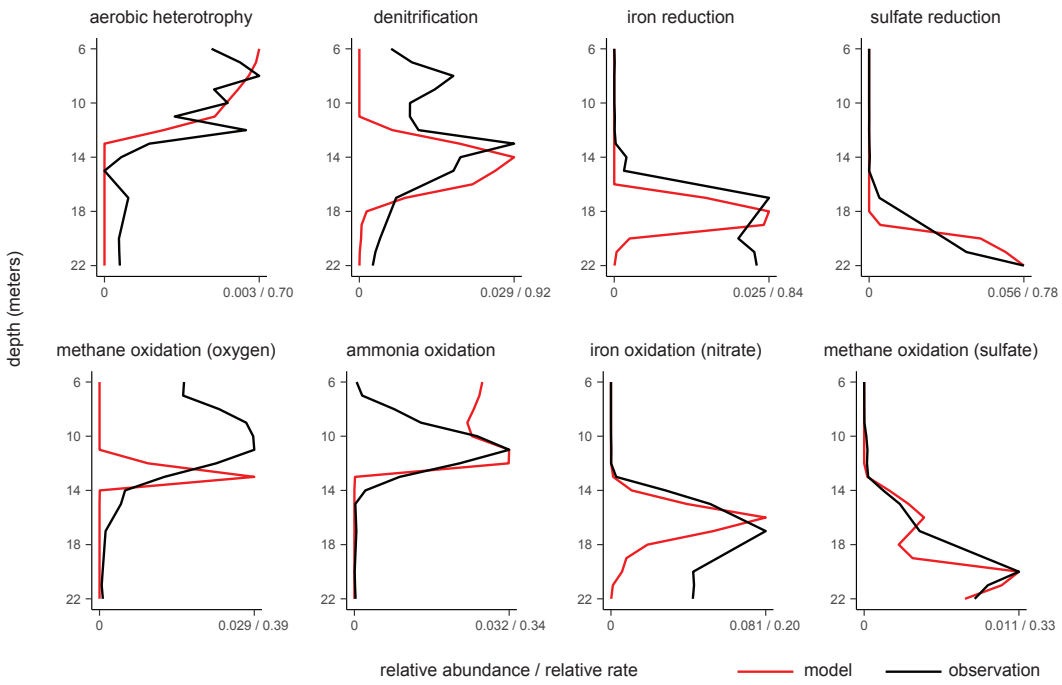


Figure 3. Distribution of key populations (black lines, relative abundance) from 2013 and their correspondence with modeled processes (red lines, relative rate). Even though the two sets of lines represent entirely different quantities (relative abundance of an organism vs. relative prevalence of a metabolic process) their peaks and sometimes their spreads roughly correspond, suggesting that the distribution of these organisms is largely determined by the relative favorability of the modeled metabolic processes within the lake.

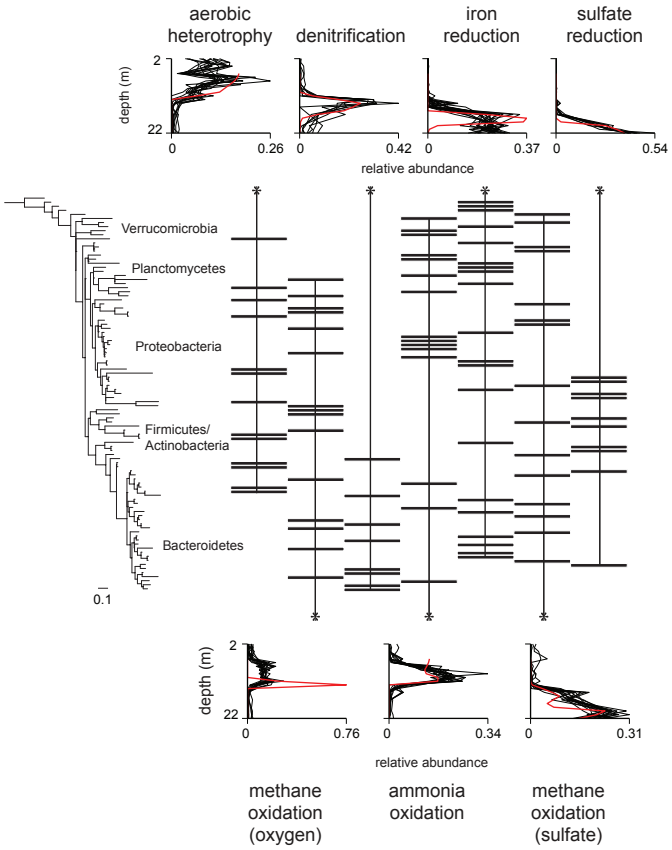


Figure 4. Operational ecological units (OEUs) are comprised of phylogenetically diverse OTUs that largely align with modeled processes. The phylogenetic tree (left; scale bar is substitutions per site) shows the relationship between OTUs' representative 16S rRNA gene sequences in OEUs containing key populations. Every OTU (on the rows) in the tree is a member of one OEU (on the columns). Each bar indicates that the OTU in that row belongs to the OEU represented by that column. Each inset shows the distributions (black lines) with depth for OTUs in that OEU as well as the distribution (red line) of a biogeochemical process (inset labels) predicted by the model. The insets above and below the main figure correspond to the adjacent OEU columns marked by asterisks. Modeled processes are only shown for the modeled region, i.e., below 5 meters depth. OEUs were matched to the modeled processes as described in the Methods.

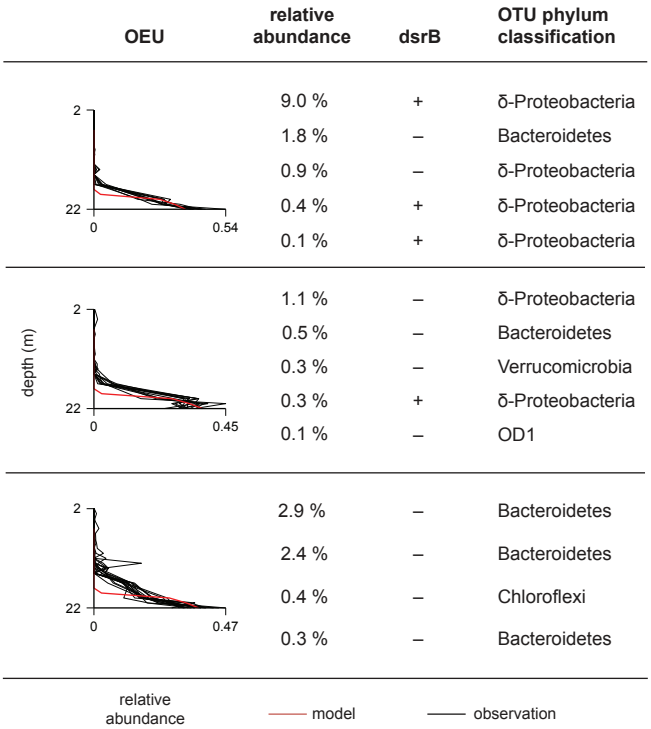


Figure 5. OEUs corresponding to sulfate reduction do not have metabolically identical OTUs. The three OEUs with profiles corresponding to sulfate reduction in the model are shown (the black lines are OTUs within each OEU; the red lines are the relative rate of sulfate reduction predicted by the model). The rows in each OEU correspond to each of that OEU's member OTUs. Of these three OEUs, the single-cell assay determined that two OEUs contained member OTUs that carry the diagnostic enzyme for sulfate reduction. The relative abundance in the second column is percent of the control non-specific 16S-barcode fusion library corresponding to each OTU sequence. The third column indicates whether a fusion product was identified in the *dsrB*-16S gene fusion library (+/−). The fourth column indicates phylum level classification for each OTU, even if the OTU could be classified at a lower taxonomic rank.

Chapter 4

Designing fecal microbiota transplant trials that account for differences in donor stool efficacy

The contents of this chapter were submitted with Scott W. Olesen*, Thomas Gurry* (these authors contributed equally), and Eric J. Alm as authors.

The supplementary information is in Appendix B. The supplementary tables are at the end of the chapter.

Abstract

Fecal microbiota transplantation (FMT) is a highly effective intervention for patients suffering from recurrent *Clostridium difficile*, a common hospital-acquired infection. FMT's success as a therapy for *C. difficile* has inspired interest in performing clinical trials that experiment with FMT as a therapy for treating conditions like inflammatory bowel disease, obesity, diabetes, and Parkinson's disease. Results from clinical trials that use FMT to treat inflammatory bowel disease suggest that, for at least one condition beyond *C. difficile*, most FMT donors produce stool that is not efficacious. The optimal strategies for identifying and using efficacious donors have not been investigated. We therefore formulated an optimal Bayesian response-adaptive donor selection strategy and a computationally-tractable myopic heuristic. This algorithm computes the probability that a donor is efficacious by updating prior expectations about the efficacy of FMT, the placebo rate, and the fraction of donors that are efficacious. In simulations designed to mimic a recent FMT clinical trial, for which traditional power calculations predict ∼100% statistical power, we found that accounting for differences in donor efficacy reduced the predicted statistical power to ∼9%. For these simulations, using the Bayesian allocation strategy more than quadrupled the statistical power to ∼39%. We use the results of similar simulations to make recommendations about the number of patients, number of donors, and choice of clinical endpoint that clinical trials should use to optimize their ability to detect if FMT is effective for treating a condition.

Author Summary

Many clinical trials test the ability of a drug to a treat a disease by comparing that drug against a placebo. In fecal microbiota transplant (FMT) trials, the "drug" is stool taken from a donor. For some diseases, however, the outcome of FMT seems to depend strongly on the choice of donor, suggesting that different donors may be producing different "drugs". Standard clinical trials are not designed to account for this possible multiplicity of drugs. We translated data from a FMT clinical trial into a model of donor stool efficacy and used this model to evaluate how different trial designs can affect a trial's ability to measure FMT's effectiveness. We found that, if only some donors produce efficacious drugs, standard clinical trial designs are likely to conclude that FMT is ineffective, thus denying patients an efficacious therapy. We also show that donor allocation strategies that adaptively use the best-performing donors can dramatically improve a trial's performance. Because FMT is not well enough understood for researchers to *a priori* identify efficacious stool, our approach

makes no assumptions about the biology of the condition being treated.

Introduction

Fecal microbiota transplant (FMT), the transfer of stool from a healthy person into an ill person's gut, is a highly effective treatment for recurrent *Clostridium difficile* infections, which kill 30,000 Americans a year. Despite FMT's efficacy and increasingly widespread use, the biological mechanism by which FMT cures the infection is not fully understood [1–7]. FMT's success in treating *C. difficile* has generated interest in experiments to use FMT to treat other conditions related to the gut and the gut-associated microbiota [8, 9]. However, emerging evidence suggests using FMT for these other diseases will be more challenging. Notably, in a recent study by Moayyedi *et al.* [10] that used FMT to treat ulcerative colitis, patients appeared to respond to stool from only one of the six stool donors. Stool from all other donors was no more efficacious than placebo. These results suggest that the effectiveness of FMT can depend strongly on the choice of stool donor.

Ideally, information collected before or during a clinical trial could be used to identify which donors are efficacious. Predictions about a donor's efficacy could be repeatedly updated depending on that donor's performance, the performance of other donors, and prior expectations about donor efficacy and heterogeneity. There is, however, not enough clinical information about this variability or biological information about the mechanism by which FMT treats disease to create or validate a model for any particular indication. We therefore present a general model of differences in donor stool efficacy that is not specific for any clinical indication or biomarker. Using this model, we formulate an optimal, adaptive, Bayesian algorithm for allocating donors. We also formulate a computationally-tractable myopic Bayesian allocation heuristic.

Using simulations of clinical trials, we show that differences in donor efficacy and the trial's strategy for allocating donors can have substantial impacts on the trial's statistical power. We compare the performance of non-adaptive approaches to matching patients and donors against a variation of a previously-studied adaptive algorithm (a variation of the "play the winner" strategy [11–13]) and our own Bayesian heuristic. We find that, in many cases, traditional non-adaptive donor allocation strategies are likely to falsely conclude that FMT in inefficacious. Adaptive strategies, however, can substantially increase a trial's ability to detect if FMT is efficacious.

Methods

Model of differences in donor stool efficacy

The results of the trial reported in in Moayyedi *et al.* [10] raise the possibility that only some donors produce stool that is efficacious for treating ulcerative colitis. We developed a simple model of differences in the efficacy of stool produced by different donors (Fig. 1). The model we present assumes:

- Each donor produces efficacious stool or inefficacious stool. Stool from inefficacious donors is inert and has no effect beyond placebo.
- All patients receive one course of treatment, each from one donor. Multiple patients can receive stool from the same donor, but each patient receives stool from only one donor.
- Patient responses are dichotomous: a patient either responds to treatment (i.e., reaches a positive clinical endpoint) or not.
- Patients are exchangeable.

These assumptions mean that the model has only three parameters:

1. the placebo rate p_{placebo},
2. the probability p_{eff} that a patient will respond to stool from an efficacious donor, and
3. the frequency f_{eff} of efficacious donors among the general donor population.

These parameters can be easily related to results from clinical trials. A conversion between the clinical trial results from Moayyedi *et al.* [10] and the parameters in the model are shown in Table 1.

Further details about the model, its extensibility, and techniques for drawing random variates of its parameters are in S1 Appendix.

Donor allocation strategies

In simulations of clinical trials, four different strategies for choosing which donor to use with which patient were evaluated. The first two strategies, block allocation and random allocation, are non-adaptive, that is, the allocation decisions about which donor will be used

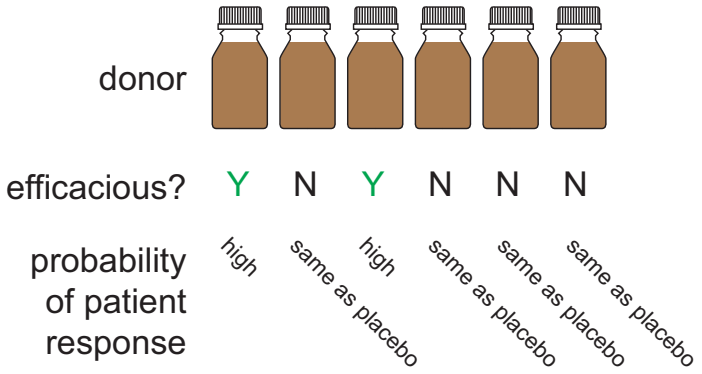


Figure 1. The model of differences in donor efficacy. In the model, donors are efficacious or not. Patients respond to FMT from an efficacious stool donor with probability p_{eff}. An FMT from an inefficacious stool donor is considered identical to a placebo, i.e., patients respond with probability p_{placebo}. The fraction of donors in the general donor population that are efficacious is f_{eff}.

Relevant trial result	Parameter	Estimated value
2 of 37 patients in placebo arm achieved remission	p_{placebo}	$2/37 = 0.054$
1 of 6 donors appeared efficacious	f_{eff}	$1/6 = 0.17$
7 of 18 patients allocated to efficacious donor achieved remission	p_{eff}	$7/18 = 0.39$

Table 1. Estimates of model parameters from the data from an ulcerative colitis clinical trial. These clinical data are drawn from the results of the trial reported in Moayyedi *et al.* [10], which used FMT to treat ulcerative colitis.

for which patient can be made before the trial begins. The other two strategies, a randomized urn-based strategy and a myopic Bayesian strategy, are response-adaptive strategies, that is, the allocation about which donor to use for a patient will depend on outcomes of previous patients (Fig. 2).

Block allocation In a block allocation, patients are evenly allocated to donors. For example, if there are 30 patients and 6 donors, the first 5 patients are treated with the first donor, the second 5 patients are treated with the second donor, etc. (In clinical practice, patients would be randomized within the blocks.)

Random allocation In a random allocation, patients are allocated to donors at random. (On average, random allocations are similar to block allocations, so these two types of non-adaptive simulation yield similar results.)

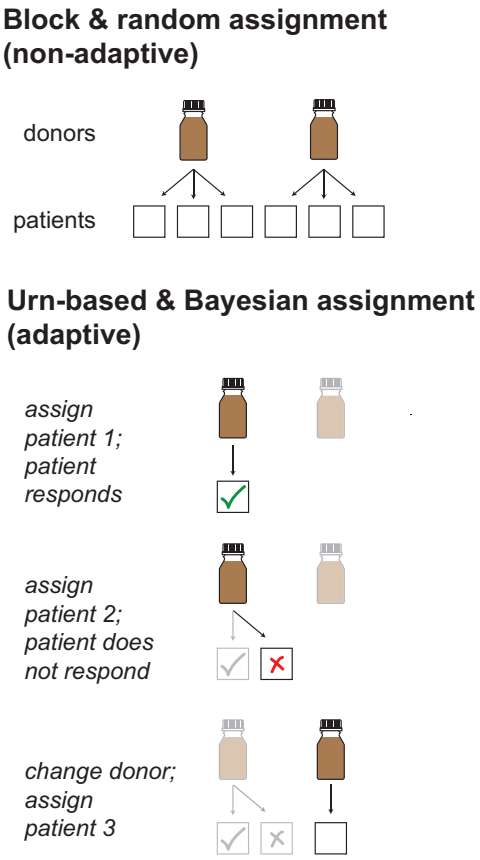


Figure 2. Adaptive donor allocations change depending on the trial's progress. In a non-adaptive allocation, the patients and donors can be matched before any patient is treated. In an adaptive allocation, the outcome of previous patients' treatments can affect the allocation of future patients. For example, if a donor is used for a patient who responds to treatment, that donor might be used again. If the patient does not respond, a different donor might be used.

Urn-based allocation Our urn-based allocation strategy is a variation on the "play-the-winner" strategies designed and studied as an ethical [14] and statistically-rigorous way to decide how to allocate patients to a treatment arm when a trial includes more than one treatment arm [15–17]. In this study, we used the generalized Pólya's urn [18] with parameters $w = 1$, $\alpha = 3$, $\beta = 0$, and without replacing the drawn ball.

Adaptive allocation strategies, like this urn-based approach, are designed to reduce the probability of a Type II error (i.e., to avoid concluding that FMT is ineffective when it actually is effective) without increasing the probability of a Type I error (i.e., without increasing the chance that a trial will conclude that FMT is effective when it is actually ineffective). If FMT is effective for some donors, this urn-based allocation should allocate more patients

to efficacious donors. If FMT is ineffective (i.e., there are no efficacious donors), then an adaptive strategy should not increase the risk of a Type I error, since any and all allocation strategies will allocate only inefficacious donors.

Myopic Bayesian strategy The urn-based approach is randomized, which is desirable in clinical trials because it can reduce certain kinds of bias [11]. However, random approaches are not guaranteed to make the best choices about donor allocation. We therefore designed a myopic Bayesian deterministic donor allocation strategy that uses the results collected during a simulated trial to make the best choice about how to allocate the next patient, similar to other "bandit" problems [19]. Unless noted, the myopic Bayesian strategy was initialized with a uniform prior on the model parameters.

Computing the posterior predictive probabilities in the myopic Bayesian allocation strategy requires a computationally-intensive numerical integration. The value of the integral was computed using Monte Carlo integration with Suave (SUbregion-Adaptive VEgas), an importance sampling method combined with a globally adaptive subdivision strategy. Sampling for this integral was performed with Sobol pseudo-random numbers. The integrator was implemented in C++ using the Cuba package [20]. Wrappers for the integration routine were implemented in Python 3 and simulations were then parallelized to run on multiple cores to optimize computational run time [21].

Details about the optimal Bayesian strategy and the derivation of the posterior predictive probabilities are in S1 Appendix.

Simulated clinical trials

The expected fraction of patients allocated to efficacious donors and the the statistical power of clinical trials using different donor allocation strategies were estimated using simulated clinical trials. In each simulation, the three model parameters (p_{placebo}, p_{eff}, f_{eff}) and the number of patients in the trial were fixed. For each combination of the trial parameters, 10,000 lists of 6 donors each were randomly generated. Donors were designated as efficacious or not efficacious by random chance according to the frequency of efficacious donors f_{eff}. The same donor lists were used for simulations for each of the allocation strategies.

In one set of simulations, the number of donors was varied among 1, 3, 5, 10, 15, and 30 donors. For those simulations, lists of 30 donors were generated for each parameter set and trial iteration. The lists were truncated for the simulations using less than 30 donors.

For each allocation strategy and donor list, a trial was simulated. In each simulation,

a patient allocated to an efficacious donor responds to the treatment with probability p_{eff}. Patients allocated to inefficacious donors or to the placebo arm respond with probability p_{placebo}. For adaptive allocations, the outcomes from all the previous patients treatment were determined before the donor for the next patient was selected. An equal number of patients was allocated to the treatment and placebo arms.

Clinically-relevant parameter values Simulations were performed for all combinations of parameter values selected to reflect clinically-relevant possibilities:

- The placebo rate p_{placebo} is either 0.05 (a low placebo rate consistent with stringent, objective outcomes; e.g., endoscopic Mayo score [10, 22]) or 0.25 (a high placebo rate consistent with self-reported, subjective outcomes [23, 24]).
- The efficacy p_{eff} of efficacious donors is either 0.4 (similar to the value in Table 1) or 0.95 (efficacy of FMT to treat *C. difficile* infection).
- The frequency f_{eff} of efficacious donors is either 0.15 (similar to the value in Table 1) or 0.9 (reflective of the fact that almost any well-screened donor produces stool that can successfully treat *C. difficile* infection).
- The number of patients in each of the treatment and control arms is 15, 30, or 60, corresponding to a range of patient numbers typical for Phase I and small Phase II clinical trials.

Of these combinations, the set of values most similar to the one in Table 1 is $p_{\text{placebo}} = 0.05$, $p_{\text{eff}} = 0.4$, $f_{\text{eff}} = 0.15$, $N_{\text{patients}} = 30$.

Computing statistical power After determining the outcome of all the patients in the trial, the p-value of a one-sided Fisher's exact test (asserting that the response rate in the treatment arm was greater) was calculated. The proportion of simulations that produced $p < 0.05$ was the estimate of the statistical power for that allocation strategy under those trial parameters. Confidence intervals were calculated using the method of Clopper and Pearson [25]. Values are rounded to two or three significant digits.

Results

Trials using adaptive strategies allocate more patients to efficacious donors

The purpose of adaptive donor allocation strategies is to identify and use efficacious donors. We therefore expected that simulated trials that use adaptive strategies would allocate more patients to efficacious donors (compared to simulated trials that used the block or random donor allocation strategies).

For every parameter set simulated, the average fraction of patients allocated to efficacious donors was greater in the adaptive strategies (urn-based and myopic Bayesian) than in the non-adaptive strategies (block and random; S1 Table). The two non-adaptive allocation strategies performed almost identically: for each parameter set, their results differed by less than 1 percentage point. The two adaptive strategies performed similarly: for half of the parameter sets, their results differed by less than 2 percentage points. In the remaining parameter sets, their results varied by between 2 and 9 percentage points.

When efficacious donors are common ($f_{\text{eff}} = 0.9$), the adaptive and non-adaptive strategies performed similarly (Fig. 3). In other cases, the performance of the two strategies differed substantially. For example, for the parameterization most similar to the one in Table 1, the random strategy allocated 15% of patients to efficacious donors while the myopic Bayesian strategy allocated 41% of patients to efficacious donors.

Trials using adaptive allocation have higher statistical power

Because trials that used the adaptive donor allocation strategies allocated more patients to efficacious donors than the trials that used the non-adaptive strategies, we expected that trials using adaptive strategies would have greater statistical power.

The adaptive strategies consistently yielded higher statistical powers than the non-adaptive strategies (Table 2). When efficacious donors are rare, the performance gap is larger. For example, for the parameterization most similar to the one in Table 1, a trial that uses random allocation is expected to have 9% power, while the myopic Bayesian strategy can deliver 39% power. The gap in performance is smallest when selecting a donor at random is likely to yield an efficacious donor: among the trials with $f_{\text{eff}} = 0.9$, the adaptive and non-adaptive statistical powers differed by less than 6 percentage points.

Low statistical powers when f_{eff} is small are likely due to the fact that if *all* the available donors are not efficacious, then no allocation strategy should make a trial achieve significance.

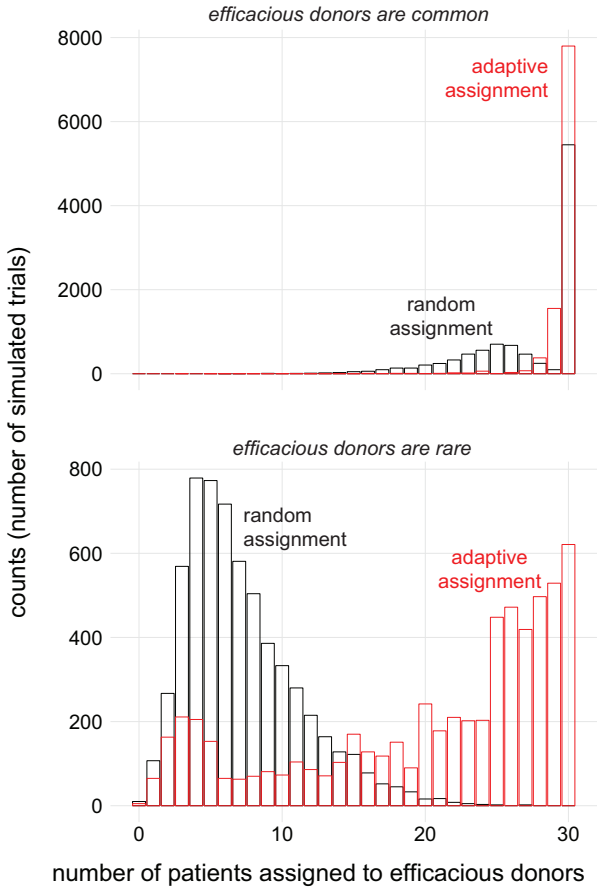


Figure 3. Adaptive strategy allocates more patients to efficacious donors. (top) For the parameterization most similar to the one in Table 1 (top), the adaptive strategy (red) allocated more patients to efficacious donors than random allocation (black) did. When efficacious donors are common (bottom; same parameters as top but with $f_{\text{eff}} = 0.9$), the two strategies allocate similar numbers of patients to efficacious donors. For visual clarity, only trials in which there was at least one efficacious donor are shown.

For example, if only 15% of donors are efficacious ($f_{\text{eff}} = 0.15$) and there are only 6 donors (the number used in these simulations), then we expect that 38% of trials will have no good donors (using the binomial distribution function). We therefore separately analyzed the simulated trials in which no donors were efficacious and the simulated trials in which at least one donor was efficacious (S2 Table). When no donors are efficacious, trials with adaptive or non-adaptive strategies have ∼0% power. Among trials with at least one efficacious donor, the difference in statistical power between adaptive and non-adaptive strategies is greater than the difference computed using the results of all trials.

Conversely, the power computed in traditional calculations that do not account for

Trial parameters				FMT power (%)				
p_{placebo}	p_{eff}	N_{patients}	f_{eff}	block	random	urn	Bayesian	Naive power (%)
0.05	0.4	15	0.15	4.57	5.04	17.2	19.1	68.1
		30		8.44	8.89	33.7	39.4	93.8
		60		23	24.1	53.5	58	100
		15	0.9	59.1	59.7	63.4	64.3	67.8
		30		87.7	87.4	92.1	93.2	94.2
		60		99.2	99.1	99.8	99.9	100
	0.95	15	0.15	19.9	21.3	60.3	60.7	100
		30		33.9	34.2	61.4	61.6	100
		60		55.9	53.2	63.2	63.1	100
		15	0.9	99.8	99.5	100	100	100
		30		100	100	100	100	100
		60		100	100	100	100	100
0.25	0.4	15	0.15	2.8	2.88	3.43	3.77	11.4
		30		4.44	4.55	7.16	7.27	25.5
		60		6.55	6.15	13.6	13.9	47.3
		15	0.9	10.5	10.3	10.5	10.7	11.7
		30		21.6	21.3	23	23	25
		60		40.7	40.7	44.5	44.8	47.2
	0.95	15	0.15	9.04	10.3	38.6	47.4	99.4
		30		20	20.3	60	60.1	100
		60		32.6	32.7	63.4	62.7	100
		15	0.9	94.3	93.9	98.2	99	99.6
		30		99.6	99.4	100	100	100
		60		100	100	100	100	100

Table 2. Adaptive strategies yield clinical trials with higher statistical power. "FMT power" is the power computed by simulating the results of trials that would occur if the frequency of efficacious donors if f_{eff}. "Naive power" is the power computed in the situation in which all donors are efficacious (i.e., $f_{\text{eff}} = 1.0$). All 95% confidence intervals on these values are within 1% of the reported value and are not shown.

differences in donor efficacy (i.e., that assume that all donors are efficacious, or equivalently $f_{\text{eff}} = 1.0$) is, in many cases, substantially higher than the power computed when accounting for differences in donor efficacy (Table 2, column "Naive powers"). For example, for the parameter set most similar to the one in Table 1, the naive calculation predicts 94% power, but the calculation that accounts for differences in donor efficacy predicts only 9% power for non-adaptive allocation strategies. The differences between the powers computed by the naive method and our approach is largest when f_{eff} is small.

Performance of adaptive strategies depend on their parameterization

In these simulations, we varied the actual values of p_{placebo} and p_{eff} but we always initialized the adaptive algorithms the same ways. To determine the sensitivity of the adaptive allocation algorithms' performance to their initialization, we simulated trials in which the actual model parameters were fixed but the algorithms' initializations varied (S3 Table and S4 Table). The myopic Bayesian algorithm's performance was mostly robust to the parameterization of its prior distribution except when the prior was strong and inaccurate. Accurate priors, weak priors, and uniform priors provide comparable performance. In contrast, the urn algorithm delivered widely varying powers, from 15% to 40%, depending on its parameterization.

Increasing the number of available donors benefits the adaptive Bayesian strategy

Increasing the number of available donors increases the probability that at least one of them will be efficacious. We therefore determined, for each donor allocation strategy, the number of available donors that optimized the trial's expected power. Simulations showed that increasing the size of the donor "pool" almost always increased the power of trials using the myopic Bayesian donor allocation but, depending on the parameter set, could increase or decrease the power of trials using other allocation strategies (S5 Table).

Table 3 shows how donor selection strategy and model parameter values affects the optimal number of donors. Notably, when efficacious donors are uncommon ($f_{\text{eff}} = 0.15$) and only moderately efficacious ($p_{\text{eff}} = 0.4$), the non-adaptive strategies perform optimally when only one donor is used. In other words, when using non-adaptive strategies in this parameter regime, it is wiser to take the 15% chance of picking a single efficacious donor than it is to distribute patients across many donors, allotting around 15% of them to efficacious donors. In contrast, the myopic Bayesian donor allocation almost always benefits from a larger donor pool.

Trial parameters			Optimal N_{donors}			
f_{eff}	p_{eff}	p_{placebo}	random	block	urn	Bayesian
0.15	0.4	0.05	1	1	10	10–30
0.15	0.4	0.25	1	1	5	10
0.15	0.95	0.05	3–15	3–5	15–30	15–30
0.15	0.95	0.25	3	3	10	10–30
0.9	0.4	0.05	1–30	3–30	3–30	3–30
0.9	0.4	0.25	1–30	1–30	1–30	5–30
0.9	0.95	0.05	3–30	3–30	3–30	3–30
0.9	0.95	0.25	3–30	3–30	3–30	3–30

Table 3. The optimal number of donors varies by donor selection strategy and model parameter values. For each parameter value, trials were simulated using 1, 3, 5, 10, 15, and 30 donors. (The number of patients was fixed at 30.) The number of donors that optimized the expected power was identified. If multiple numbers of donors yielded powers within 0.05 of the optimal value, all those numbers are reported as a range.

Discussion

Model limitations

We chose to use a simple model for a simple use case because, in the absence of data about the treatment histories of hundreds of patients using dozens of donors, we do not believe that more complicated models will be more useful to aiding trial design. The model's greatest weakness is that it cannot be validated, but it is exactly the model's purpose to improve the probability of collecting the kind of information that could validate or invalidate it. In light of the dearth of data, we developed a simple model, and it could be that the simplifications we made limit the model's validity. For example, the model assumes that each donor produces efficacious stool or inefficacious stool (when in fact there is probably day-to-day and donor-to-donor variation in stool efficacy) and that all patients receive one course of treatment (while, say, patients who do not respond to a first treatment might be treated with stool from a different donor).

In our simulations, we assumed that the outcome from all the previous patients treatments are known before the donor for the next patient is selected. In reality, patients in an FMT trial overlap. The urn-based method can still be used for overlapping patients [18], but the myopic Bayesian method would require some modification. For example, clinicians could choose to consult the myopic Bayesian's rankings of donors intermittently, or patients could be allocated in proportion to the predicted probabilities of successful treatments.

Differences in donor efficacy should be accounted for in trial design

Our results entail recommendations to clinicians. First, the powers we computed here are, in many cases, well below the powers computed assuming that all donors are efficacious. We therefore encourage researchers to consult our predictions about statistical power when deciding on the size of their trials.

Second, a high placebo rate can substantially decrease the statistical power of an FMT trial. We therefore encourage researchers to use the most stringent outcome measurement possible (e.g., an endoscopic Mayo score for inflammatory bowel disease).

Third, adaptive donor allocation strategies consistently delivered higher statistical power than traditional, non-adaptive approaches. We therefore recommend that researchers use such an adaptive strategy. The urn-based strategy has the advantages that similar response-adaptive strategies may be familiar to clinicians, it is randomized, and it is simple to implement. However, an urn-based strategy needs to be carefully parameterized: a badly-parameterized urn-based strategy performs similarly to random allocation. The adaptive Bayesian donor allocation algorithm performs well even when using the "default" settings (a uniform prior) but is complex and deterministic. To fully leverage this strategy, a clinician would need to consult the algorithm's output after every patient outcome and follow the algorithm's deterministic instructions, which might introduce bias.

Fourth, adaptive algorithms benefit from having access to a "bank" of 10 or more donors. Researchers hoping to achieve the full benefits of adaptive donor selection must be prepared to change donors multiple times during the trial.

Finally, researchers reporting about FMT trials should include information about the donors, notably how many donors were used and what proportion of patients alloted to each donor responded to treatment. This information will help future researchers account for differences in donor efficacy.

Future research may identify mechanistic explanations of FMT's efficacy

The adaptive allocation strategies we described here have a narrow aim: to increase the number of successful patient outcomes in a trial. In theory, an adaptive trial design is capable of more. For example, if it were hypothesized that FMT succeeded or failed because of the presence or absence of some particular microbial species in the donor's stool, then an adaptive trial design could recommend donor choices that aim to identify that critical species.

We did not pursue a hypothesis-centric approach because we believe it is premature. Even the mechanism by which FMT treats *C. difficile*, the most well-studied case, remains unclear. We expect that strong hypotheses about mechanism will come from retroactive comparison of efficacious vs. inefficacious stool *after* clinical trials have definitively show that FMT is effective for treating some disease. Our study aims to do exactly this. Until then, we hope that our results about adaptive donor allocation help more patients benefit from FMT and will help clinicians identify those conditions that FMT can treat.

Acknowledgments

This material is based on work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374 and by the Center for Microbiome Informatics and Therapeutics at MIT.

Supporting Information

S1 Appendix

Model design and Bayesian algorithm. Details about the mathematical model and the adaptive Bayesian donor allocation algorithm are in Appendix B of this thesis.

S1 Table

Fraction of patients allocated to efficacious donors. The same simulated trials were analyzed to create Table 2 and this table. N_{p} is the number of patients.

Trial parameters				Fraction assigned to efficacious donors (mean +/- std. dev.)							
				block		random		urn		Bayesian	
p_{plac}	p_{eff}	f_{eff}	N_{p}	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
0.05	0.4	0.15	15	0.15	0.15	0.15	0.17	0.29	0.3	0.32	0.33
0.05	0.4	0.15	30	0.15	0.14	0.15	0.16	0.36	0.35	0.41	0.38
0.05	0.4	0.15	60	0.15	0.14	0.15	0.15	0.44	0.39	0.49	0.42
0.05	0.4	0.9	15	0.9	0.12	0.9	0.14	0.95	0.075	0.96	0.068
0.05	0.4	0.9	30	0.9	0.12	0.9	0.13	0.97	0.051	0.98	0.041
0.05	0.4	0.9	60	0.9	0.12	0.9	0.13	0.98	0.03	0.99	0.022
0.05	0.95	0.15	15	0.15	0.15	0.15	0.17	0.44	0.36	0.51	0.42
0.05	0.95	0.15	30	0.15	0.15	0.15	0.16	0.52	0.42	0.56	0.45
0.05	0.95	0.15	60	0.15	0.15	0.15	0.15	0.57	0.45	0.6	0.47
0.05	0.95	0.9	15	0.9	0.13	0.9	0.14	0.97	0.047	0.99	0.028
0.05	0.95	0.9	30	0.9	0.12	0.9	0.13	0.98	0.027	1	0.013
0.05	0.95	0.9	60	0.9	0.12	0.9	0.13	0.99	0.014	1	0.0069
0.25	0.4	0.15	15	0.15	0.15	0.15	0.17	0.2	0.25	0.21	0.29
0.25	0.4	0.15	30	0.15	0.15	0.15	0.16	0.24	0.3	0.24	0.34
0.25	0.4	0.15	60	0.15	0.15	0.15	0.15	0.29	0.34	0.28	0.38
0.25	0.4	0.9	15	0.9	0.13	0.9	0.14	0.93	0.14	0.94	0.15
0.25	0.4	0.9	30	0.9	0.12	0.9	0.13	0.94	0.13	0.95	0.14
0.25	0.4	0.9	60	0.9	0.12	0.9	0.13	0.96	0.11	0.96	0.14
0.25	0.95	0.15	15	0.15	0.15	0.15	0.17	0.36	0.34	0.45	0.42
0.25	0.95	0.15	30	0.15	0.15	0.15	0.16	0.45	0.38	0.52	0.44
0.25	0.95	0.15	60	0.15	0.15	0.15	0.15	0.53	0.42	0.57	0.46
0.25	0.95	0.9	15	0.9	0.13	0.9	0.14	0.96	0.076	0.99	0.056
0.25	0.95	0.9	30	0.9	0.12	0.9	0.13	0.98	0.048	0.99	0.029
0.25	0.95	0.9	60	0.9	0.12	0.9	0.13	0.99	0.029	1	0.016

S2 Table

Power of simulated clinical trials, conditioned on presence of efficacious donors. A subset of the data used to create Table 2 (only those simulations with $N_{\text{patients}} = 30$ and $p_{\text{placebo}} = 0.05$) was analyzed by separately estimating the power for the trials in which no donors were efficacious ("no efficacious donors") and in which at least one donor was efficacious ("some efficacious donors"). "All simulations" shows the same data as in Table 2. For $f_{\text{eff}} = 0.9$, none of the 10,000 simulations had a donor pool with no efficacious donors. All 95% confidence intervals are within 1% of the reported value and are not shown.

Parameters		Simulation conditions		Power (%)			
p_{eff}	f_{eff}	Donor pool quality	$N_{\text{simulations}}$	random	block	urn	Bayesian
0.4	0.15	all trials	10000	8.9	8.4	34	39
0.4	0.15	no efficacious donors	3794	0.37	0.29	0.4	0.16
0.4	0.15	some efficacious donors	6206	14	13	54	63
0.4	0.9	some efficacious donors	10000	87	88	92	93
0.95	0.15	all trials	10000	34	34	61	62
0.95	0.15	no efficacious donors	3851	0.62	0.44	0.34	0.31
0.95	0.15	some efficacious donors	6149	55	55	100	100
0.95	0.9	some efficacious donors	10000	100	100	100	100

S3 Table

Sensitivity of simulated clinical trial power to parameterization of the myopic Bayesian algorithm parameters. Starting from the parameter described in Table 1, 10,000 trials using the myopic Bayesian donor allocation were simulated for each of six different parameterizations of the Bayesian algorithm. "Accurate" means that the prior is centered around the true value (i.e., that $\frac{A}{A+B}$ equals the true value); "inaccurate" means that the prior is centered at approximately double the true value (but that $A + B$ has been held constant). "Strong" means that every hyperparameter is ten-fold greater; "weak" means that every hyperparameter is ten-fold smaller. "Uniform" means a uniform prior was used for all parameters. All 95% confidence intervals are within 1 percentage point of the reported value and are not shown.

	Hyperparameter values						
Hyperparameterization	A_{placebo}	B_{placebo}	A_{peff}	B_{peff}	A_{feff}	B_{feff}	Power (%)
From Table 1	2	35	7	11	1	5	41.1
Weak, accurate	0.2	3.5	0.7	1.1	0.1	0.5	41.2
Strong, accurate	20	350	70	110	10	50	40.4
Weak, inaccurate	0.4	3.3	1.4	0.4	0.2	0.4	40.9
Strong, inaccurate	40	330	140	40	20	40	34.4
Uniform	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	39.6

S4 Table

Sensitivity of simulated clinical trial power to urn parameterization. Using the parameter set described in Table 1, 10,000 trials using the urn-based donor allocation were simulated for each of several combinations of the parameters for the urn model as described in [18]. We also vary whether drawn balls are replaced or not. All 95% confidence intervals are within 1 percentage point of the reported value and are not shown.

w	α	β	replace?	power (%)
1	3	0	True	28.5
1	3	0	False	35.6
1	3	1	True	13.9
1	3	1	False	14.2
1	6	0	True	33.2
1	6	0	False	41
1	6	1	True	16.2
1	6	1	False	16.4
3	3	0	True	19.7
3	3	0	False	29.1
3	3	1	True	13.2
3	3	1	False	13.1
3	6	0	True	24.8
3	6	0	False	32.6
3	6	1	True	15.6
3	6	1	False	15.3

S5 Table

Dependence of simulated clinical trial power on number of available donors. For each parameter value, trials were simulated using 1, 3, 5, 10, 15, and 30 donors. The number of patients was fixed at 30. All 95% confidence intervals are within 1 percentage point of the reported value and are not shown.

Parameters				Power (%)			
f_{eff}	p_{eff}	p_{placebo}	N_{donors}	random	block	urn	Bayesian
0.15	0.4	0.05	1	15.5	15.3	15.5	14.8
0.15	0.4	0.05	3	12.4	12.2	28.4	30.9
0.15	0.4	0.05	5	9.87	9.67	33.8	38
0.15	0.4	0.05	10	7.49	6.93	38.2	44.3
0.15	0.4	0.05	15	6.91	6.38	34.6	45.9
0.15	0.4	0.05	30	6.17	5.26	16.6	45.9
0.15	0.4	0.25	1	6.08	6.15	5.89	5.58
0.15	0.4	0.25	3	4.64	4.21	6.7	6.89
0.15	0.4	0.25	5	4.65	4.7	7.28	7.1
0.15	0.4	0.25	10	4.37	4.18	6.2	7.52
0.15	0.4	0.25	15	4.48	4.09	5.41	7.07
0.15	0.4	0.25	30	4.25	4.3	4.67	6.89
0.15	0.95	0.05	1	15.2	15.1	15	14.9
0.15	0.95	0.05	3	34.4	37.2	39	39.2
0.15	0.95	0.05	5	35	36.1	56.3	56.5
0.15	0.95	0.05	10	33.7	33.3	80.1	80.4
0.15	0.95	0.05	15	33.5	32.6	91	90.6
0.15	0.95	0.05	30	31.6	30.4	87.4	93.9
0.15	0.95	0.25	1	17.1	17.3	17	16.8
0.15	0.95	0.25	3	23.1	22.6	39	39.9
0.15	0.95	0.25	5	21.3	20.9	55.2	55.5
0.15	0.95	0.25	10	18.8	17.7	69.7	69.7
0.15	0.95	0.25	15	17.4	16.8	65.4	69.9
0.15	0.95	0.25	30	16.4	15.4	43.2	69.4
0.9	0.4	0.05	1	84.9	84.7	84.6	84.3

continued on next page

Parameters				Power (%)			
f_{eff}	p_{eff}	p_{placebo}	N_{donors}	random	block	urn	Bayesian
0.9	0.4	0.05	3	85.9	85.8	91.8	92.4
0.9	0.4	0.05	5	87.4	87.2	92.4	92.4
0.9	0.4	0.05	10	88.4	88.9	92.3	92.8
0.9	0.4	0.05	15	88	89	91.8	92.9
0.9	0.4	0.05	30	89.1	89.3	91.1	92.9
0.9	0.4	0.25	1	22	22.3	22.5	22.1
0.9	0.4	0.25	3	21.8	21.2	23.1	22.7
0.9	0.4	0.25	5	21.7	21.3	23.1	23.2
0.9	0.4	0.25	10	21.7	21.3	22.7	23.7
0.9	0.4	0.25	15	22.2	21.5	22.5	24
0.9	0.4	0.25	30	21.9	21.6	22.1	23.5
0.9	0.95	0.05	1	90.1	90.1	90.1	90.4
0.9	0.95	0.05	3	99.5	99.7	99.9	99.9
0.9	0.95	0.05	5	100	100	100	100
0.9	0.95	0.05	10	100	100	100	100
0.9	0.95	0.05	15	100	100	100	100
0.9	0.95	0.05	30	100	100	100	100
0.9	0.95	0.25	1	90.4	90.4	90.4	90.5
0.9	0.95	0.25	3	97.7	98.1	99.9	99.9
0.9	0.95	0.25	5	99.1	99.4	100	100
0.9	0.95	0.25	10	99.8	99.9	100	100

References

1. Cohen SH, Gerding DN, Johnson S, Kelly CP, Loo VG, McDonald LC, et al. Clinical Practice Guidelines for *Clostridium difficile* Infection in Adults. Infect Control Hosp Epidemiol. 2010;31(5):431–455. doi:10.1086/651706.

2. van Nood E, Vrieze A, Nieuwdorp M, Fuentes S, Zoetendal EG, de Vos WM, et al. Duodenal Infusion of Donor Feces for Recurrent Clostridium difficile. N Engl J Med. 2013;368(5):407–415. doi:10.1056/NEJMoa1205037.

3. Kelly CP, LaMont JT. *Clostridium difficile* – More Difficult Than Ever. N Engl J Med. 2008;359(18):1932–1940. doi:10.1056/NEJMra0707500.

4. Alang N, Kelly CR. Weight Gain After Fecal Microbiota Transplantation. Open Forum Infect Dis. 2015;2(1). doi:10.1093/ofid/ofv004.

5. Li SS, Zhu A, Benes V, Costea PI, Hercog R, Hildebrand F, et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. Science. 2016;352(6285):586–589. doi:10.1126/science.aad8852.

6. Kassam Z, Lee CH, Yuan Y, Hunt RH. Fecal Microbiota Transplantation for *Clostridium difficile* Infection: Systematic Review and Meta-Analysis. Am J Gastroenterol. 2013;108(4):500–508. doi:10.1038/ajg.2013.59.

7. Khoruts A, Sadowsky MJ. Understanding the mechanisms of faecal microbiota transplantation. Nat Rev Gastroenterol Hepatol. 2016;doi:10.1038/nrgastro.2016.98.

8. Smits LP, Bouter KEC, de Vos WM, Borody TJ, Nieuwdorp M. Therapeutic Potential of Fecal Microbiota Transplantation. Gastroenterology. 2013;145(5):946–953. doi:10.1053/j.gastro.2013.08.058.

9. Sadowsky MJ, Khoruts A. Faecal microbiota transplantation is promising but not a panacea. Nat Microbiol. 2016;1. doi:10.1038/nmicrobiol.2016.15.

10. Moayyedi P, Surette MG, Kim PT, Libertucci J, Wolfe M, Onischi C, et al. Fecal Microbiota Transplantation Induces Remission in Patients With Active Ulcerative Colitis in a Randomized Controlled Trial. Gastroenterology. 2015;149(1):102–109.e6. doi:10.1053/j.gastro.2015.04.001.

11. Cook TD, DeMets DL. Introduction to Statistical Methods for Clinical Trials. Chapman & Hall/CRC; 2008.

12. Bartlett RH, Roloff DW, Cornell RG, Andrews AF, Dillon PW, Zwischenberger JB. Extracorporeal Circulation in Neonatal Respiratory Failure: A Prospective Randomized Study. Pediatrics. 1985;76(4):479–487.

13. Chow SC, Chang M. Adaptive Methods in Clinical Trials. Chapman & Hall/CRC; 2007.

14. Hung JHM, O'Neill RT, Wang SJ, Lawrence J. A Regulatory View of Adaptive/Flexible Clinical Trial Design. Biom J. 2006;48(4):565–573. doi:10.1002/bimj.200610229.

15. Wei LJ, Smythe RT, Lin DY, Park TS. Statistical Inference with Data-Dependent Treatment Allocation Rules. J Am Stat Assoc. 1990;85(409):156–162. doi:10.1080/01621459.1990.10475319.

16. Wei LJ, Durham S. The Randomized Play-the-Winner Rule in Medical Trials. J Am Stat Assoc. 1978;73(364):840–843. doi:10.1080/01621459.1978.10480109.

17. Zelen M. Play the Winner Rule and the Controlled Clinical Trial. J Am Stat Assoc. 1969;64(325):131–146. doi:10.1080/01621459.1969.10500959.

18. Wei LJ. The Generalized Polya's Urn Design for Sequential Medical Trials. Ann Stat. 1979;7(2):291–296.

19. Berry DA, Fristedt B. Bandit Problems: Sequential Allocation of Experiments. Springer Netherlands; 1985.

20. Hahn T. Cuba – a library for multidimensional numerical integration. Comput Phys Commun. 2005;168(2):78–95. doi:10.1016/j.cpc.2005.01.010.

21. Tange O. GNU Parallel – The Command-Line Power Tool. ;login: The USENIX Magazine. 2011;36(1):42–47. doi:10.5281/zenodo.16303.

22. Rossen NG, Fuentes S, van der Spek MJ, Tijssen JG, Hartman JHA, Duflou A, et al. Findings From a Randomized Controlled Trial of Fecal Transplantation for Patients With Ulcerative Colitis. Gastroenterology. 2015;149(1):110–118. doi:10.1053/j.gastro.2015.03.045.

23. Su C, Lichtenstein GR, Krok K, Brensinger CM, Lewis JD. A meta-analysis of the placebo rates of remission and response in clinical trials of active Crohn's disease. Gastroenterology. 2004;126(5):1257–1269. doi:10.1053/j.gastro.2004.01.024.

24. Su C, Lewis JD, Goldberg B, Brensinger C, Lichtenstein GR. A Meta-Analysis of the Placebo Rates of Remission and Response in Clinical Trials of Active Ulcerative Colitis. Gastroenterology. 2007;132(2):516–526. doi:10.1053/j.gastro.2006.12.037.

25. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404–413. doi:10.1093/biomet/26.4.404.

Chapter 5

Discussion

5.1 Summary and limitations of reported work

This thesis describes contributions to the interpretation of microbial ecology sequence data and to the design of clinical trials. These contributions each have limitations that restrict their validity and applicability.

In Chapter 2, I introduced `texmex`, a tool designed to quantify the dynamics of microbial taxa in microbial ecology experiments that use amplicon sequence data, use pre-tests, and have few or no replicates. I expect this approach will be helpful when researchers want to analyze a pilot experiment, the environmental inoculum is difficult to acquire, or the experimentation is particularly onerous. Ideally, a researcher would perform many replicate experiments and use that information for a rigorous statistical inference that does not require any special consideration of the ecological structure of the data in question, thus obviating the need for a technique like `texmex`. Because the method is not statistical, however, it will never supplant methods that are designed to determine whether two sets of measurements are meaningfully different from a statistical standpoint.

In Chapter 3, I introduced the operational ecological unit (OEU) and the inferred biomass interpretative framework to link taxonomic survey data, an ecosystem-level metabolic model, and the results from a single-cell genetic assay. The model itself is conceptual and intentionally simple; it therefore lacks the ability to describe compli-

cated features of the lake ecosystem it models. For example, the model could never predict the hypolimnetic oxygen minimum observed in the survey data. Operational ecological units are essentially statistical and not necessarily functional, so it is not straightforward to confirm or disprove their "existence". The utility of the OEU concept could be evaluated by comparing an analysis of OEUs with a large database of known ecological interactions. The inferred biomass framework makes concrete, verifiable claims about microbial community function that could be compared with metagenomic data and, ultimately, verified or disproved by comparison with an exhaustive, *in situ* survey of ecosystem function. The results of the study are overall very suggestive, but they are not experimentally verified and would require extensive co-culturing or perturbative, *in situ* metabolic experiments to validate.

In Chapter 4, I introduced a model of differences in donors' stool with respect to its probability to cause patients to respond to treatment with fecal microbiota transplant. The model makes concrete predictions about the utility of clinical trial designs, but the structure of the model is based on a small amount of clinical trial data and would require extensive clinical trial data to verify. This study is in a catch-22: it aims to improve the probability of finding the statistically-significant clinical trial data that would provide the only way to assess the validity of the model.

5.2 Potential extensions of reported work

5.2.1 Rank-abundance distributions and small data

In a narrow sense, `texmex` is a software package that converts OTU tables into tables of values related to the initial counts and provides convenience functions for manipulating and selecting interesting OTUs based on those transformed values. More broadly, that work makes two contributions that should be helpful for future efforts to improve the interpretability of DNA sequence data in microbial ecology.

First, I was surprised that I could find no studies that directly examined or utilized the rank-abundance distribution of microbial ecology sequence data. (I found only

one paper that fit a rank-abundance distribution to microbial ecology data [1], and I describe what I perceive as deficiencies in the logic of its application in Chapter 2.) I believe that there are many applications that will emerge from using such rank-abundance distributions, just as z-scores make normally-distributed data more tractable for analysis. In particular, I expect that any attempt to compare OTUs across samples could benefit from the kind of "normalization" that `texmex` does. The standard statistical approach, in which an OTU's counts across samples are modeled as variates of a single random variable, seems like a weak approach compared to focusing on the ecological processes that cause the entire community to assemble.

This sort of sample-wise approach should also be helpful for understanding some of the more confusing aspects of microbiome data, particularly the zeros and the effects of rarefaction. It is becoming clearer that micro- and macroecology can share their methods [2], so microbial ecologists should, in some cases, pay closer attention to the methods and approaches used in traditional ecology.

Second, `texmex` starts from a very different place from many other analytical methods: it assumes a paucity of data rather than an abundance. As DNA sequencing has become cheaper, it is tempting to believe that microbial ecology is now limited only by the cleverness of the algorithms used to generate the data or the cleverness of the scientists who decide what questions to investigate. In fact, sample acquisition is still, in many cases, a limiting factor in microbial ecology, as has been my experience in the project described in Chapter 2 (as well as in a separate project in coordination with the Department of Energy). If a study is not comfortably in the regime of big data, I believe it is wiser to relegate yourself to the regime of small data. Although you can "do statistics" with three samples, if you cannot get twenty samples, it might be wiser to collect two and use a small-data technique for the first experiment. As reviewers of the manuscript pointed out, it is always better, *ceteris paribus*, to have more replicates. My point is that having more replicates always come at some cost, and the added replicates might deliver a p-value without any additional scientific insight.

I was impressed that a recent paper studying oil degradation pathways in samples

collected from the Deepwater Horizon spill—and which used an impressive combination of isotope labeling and metagenomic sequencing—identified similar organisms as my algorithm did [3], showing the power that small data and wise analytics can have.

There were extensions of this work that were outside the scope of this thesis. Are there datasets that are sufficiently well-resolved to be able to distinguish the rank-abundance distribution of microbial ecosystems? Is that distribution the Poisson-lognormal distribution or something else? Is it different for different ecosystems? What does that tell us, theoretically, about the structure of those communities? Can we use rank-abundance models to avoid the problems that compositional data present for analysis? Can we use rank-abundance distributions to infer more rational models of the behavior of taxa across samples? Relatedly, can we use rank-abundance models and timeseries data to draw inferences about the dynamical behaviors of individual taxa and entire bacterial communities? Can we better explain the overdisperse and apparently noisy behavior of taxa through time?

5.2.2 Modeling, consortia, and combinations of methodologies

Like `texmex`, the methods used for the project described in Chapter 3 also aimed to extract the maximum amount of insight from limited data. This project integrated the results from multiple methods to yield a single, biologically-interpretable discovery. This integration carries some lessons of its own.

First, models of microbial communities should aim for an optimum between complexity and falsifiability. Because the data generated by DNA sequencing are so massive and so complicated, it is tempting to make a complicated model of their behavior. However, even if such a model were made and even if it correctly recapitulated the system's behavior, are we better off for having it? For example, the model reported in this project used abstract categories (e.g., sulfate reduction) to describe microbes' behavior. The identity of the microbes that seemed to belong in that abstract category was determined separately, and it is the link between the microbes' identity and the abstract behavior in the model that was useful. If the model had perfectly described the behavior of every microbial species, then we would have pro-

duced a system exactly as complicated as the natural one, which would not advance our ability *interpret* the system.

Indeed, one of the strengths of the model presented in that work is that it was *a priori* unclear if it would even remotely recapitulate the lake's chemical dynamics. If it did not, then it would be immediately clear that our mental picture of the processes that shape the lake's behavior was largely incomplete. Adding another process into the model (e.g., the interaction between iron and sulfur) would hold the model and the associated data to a much more stringent measure: it would require a great enough precision in the data to be able to distinguish between the cases in which the iron-sulfur interactions are included or not. Given the year-to-year variability in this system, an ecosystem-wide model is not the appropriate tool for making that discrimination. The simple fact that the model worked at all—and that, if it had not worked, we would have learned something too—is its major contribution. A model that is not interesting if it fails is one that should not be considered interesting if it succeeds.

There is probably great opportunity for modeling in microbial systems. The model used in this project was based on a pre-next-generation-sequencing model of microbes in a groundwater aquifer responding to pollutants, which also highlights the fact that literature from before 1990 can be full of interesting insights and thorny questions that we now have the tools to explore more deeply.

For example, despite direct measurements of bacterial growth in zebrafish [4] and a bacterium genetically engineered to answer questions about the rate of division in the gut [5], I have not seen a model of division and colonization in the mammalian gut that accounts for its directionality: how does the unidirectional transport of vast quantities of microbes from the "top" to the "bottom" affect the microbial composition in the gut? Are downstream populations less diverse than upstream populations because every microbial species present at the bottom must have once been near the top?

Second, this project makes an early, unrefined estimate of the prevalence of microbial consortia in natural environments. Before nucleotide sequencing, bacterial species were distinguished based on their appearances or tests of their metabolisms.

This process was finicky and low throughput, so we had vastly underestimated the diversity of microbes. I believe we are on the cusp of a similar revelation about microbial consortia. I expect that theoretical arguments would show that a large number of cooperative species should be expected, and this contrasts against the very small number of consortia that are known and studied. The possibility that there are large numbers of consortia in many ecosystems is probably the most scientifically interesting and important result in this thesis.

Third, this project shows the potential that combinations of methods can play in understanding microbial systems. Surveys on their own do little to address microbial function; models on their own can seem like intellectual playgrounds unconnected to reality; high-throughput screens on their own can generate large amounts of data with small amounts of insight. In particular, I expect that combinations of models, surveys, and metabolite measurements will provide interesting and useful information about the interactions between microbial species (and hosts if they have them).

5.2.3 Decision-making in microbiome science

The third project in this thesis is an outlier: it describes a simple model—like Chapter 3 does—but it uses the model for an entirely different purpose. Rather than developing information about the possible behavior of a system, it uses a model and data to make a decision in face of a question. (This distinction is reminiscent of the difference in interpretations of the p-value [6] between Fisher, who originally formulated it as a method to discern truth [7] and Neyman and Pearson, who saw it as a way to decide actions [8].) I will venture to say that most models in the world are, like this one, operational models: they are designed to integrate data to inform a decision.

DNA sequencing is already being used in medicine to, for example, diagnose infections, and there is hope that more sophisticated, rapid, point-of-care diagnostics will be useful to, say, use information about the genetics of the pathogen to decide which antibiotic to administer to a patient. However, the role of modeling in decision science for microbiome science, as such, remains unclear. In what cases could a large collection of information about the microbes inhabiting a person's gut be useful for

making a decision? What decision would be made?

There are some appealing answers. Measurements of the microbiome could be used to diagnose a disease that is otherwise difficult or invasive to diagnose [9], to quantify the risk that a patient will develop a disease, or to help stratify patients based on the probability that they will respond to certain drugs [10, 11, 12]. I expect a "middle" way will also be profitable. A model that combines a simple treatment of a system (e.g., as in Chapter 4, each donor is considered efficacious or not) and a more complex one (e.g., it is asserted that the presence or absence of some microbial taxon in the donor determines the probability of patient response) could recommend decisions that are nearly optimal with respect to the simple, operational model while deriving greater benefit for the more complex, mechanistic model. This operational half of the approach might get complex hypothesis-testing into the clinic, since the simple half of the algorithm could be relied upon to make sensible decisions even if it became clear that the complex, mechanistic model was completely incorrect.

In general, I caution microbiome scientists against interpreting too much from 16S sequencing data. The fact that DNA sequencing is a less-biased way to enumerate communities than traditional culture-based methodologies may have reduced the emphasis on the problems that DNA sequencing presents: the microbiome appears to be a dynamic, noisy system; extraction and preparation methodologies greatly affect the output signal; different bioinformatic techniques can lead to different scientific conclusions; and proper methods of statistical analysis for these data are still under debate. Targeted questions with large sample sizes and perturbative techniques are the best avenue for conclusions; small experiments with decidedly exploratory analytical methods are the best avenue for developing avenues for fuller investigation.

> κτῆμά τε ἐς αἰεὶ μᾶλλον ἢ ἀγώνισμα
> ἐς τὸ παραχρῆμα ἀκούειν ξύγκειται.
>
> Thucydides, *History*, 1.22.4

Bibliography

[1] S. W. Kembel, M. Wu, J. A. Eisen, J. L. Green, and C. von Mering. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comp Biol*, 8(10), 2012.

[2] J. B. Hughes, J. J. Hellmann, T. H. Ricketts, and B. J. M. Bohannan. Counting the uncountable: Statistical approaches to estimating microbial diversity. *Appl Environ Microbiol*, 67(10):4399–4406, 2001.

[3] N. Dombrowski, J. A. Donaho, T. Gutierrez, K. W. Seitz, A. P. Teske, and B. J. Baker. Reconstructing metabolic pathways of hydrocarbon-degrading bacteria from the Deepwater Horizon oil spill. *Nat Microbiol*, 1:16057, 2016.

[4] M. Jemielita, M. J. Taormina, A. R. Burns, J. S. Hampton, A. S. Rolig, K. Guillemin, and R. Parthasarathy. Spatial and temporal features of the growth of a bacterial species colonizing the zebrafish gut. *mBio*, 5(6), 2014.

[5] C. Myhrvold, J. W. Kotula, W. M. Hicks, N. J. Conway, and P. A. Silver. A distributed cell division counter reveals growth dynamics in the gut microbiota. *Nat Commun*, 6, 2015.

[6] S. N. Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Ann Intern Med*, 130(12):995–1004, 1999.

[7] R. Fisher. *Statistical Methods in Scientific Inference*. Macmillan, New York, 3rd edition, 1973.

[8] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond A*, 231(694-706):289–337, 1933.

[9] E. Papa, M. Docktor, C. Smillie, S. Weber, S. P. Preheim, D. Gever, G. Giannoukos, D. Ciulla, D. Tabbaa, J. Ingram, D. B. Schauer, D. V. Ward, J. R. Korzenik, R. J. Xavier, A. Bousvaros, and E. J. Alm. Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. *PLoS ONE*, 7(6):e39242, 2012.

[10] R. A. Koeth, Z. Wang, B. S. Levison, J. A. Buffa, E. Org, B. T. Sheehy, E. B. Britt, X. Fu, Y. Wu, L. Li, J. D. Smith, J. A. DiDonato, J. Chen, H. Li, G. D. Wu, J. D. Lewis, M. Warrier, J. M. Brown, R. M. Krauss, W. H. W. Tang, F. D.

Bushman, A. J. Lusis, and S. L. Hazen. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat Med*, 19(5):576–585, 2013.

[11] A. Sivan, L. Corrales, N. Hubert, J. B. Williams, K. Aquino-Michaels, Z. M. Earley, F. W. Benyamin, Y. Man Lei, B. Jabri, M.-L. Alegre, E. B. Chang, and T. F. Gajewski. Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti–PD-L1 efficacy. *Science*, 350(6264):1084–1089, 2015.

[12] M. Vétizou, J. M. Pitt, R. Daillère, P. Lepage, N. Waldschmitt, C. Flament, S. Rusakiewicz, B. Routy, M. P. Roberti, C. P. M. Duong, V. Poirier-Colame, A. Roux, S. Becharef, S. Formenti, E. Golden, S. Cording, G. Eberl, A. Schlitzer, F. Ginhoux, S. Mani, T. Yamazaki, N. Jacquelot, D. P. Enot, M. Bérard, J. Nigou, P. Opolon, A. Eggermont, P.-L. Woerther, E. Chachaty, N. Chaput, C. Robert, C. Mateus, G. Kroemer, D. Raoult, I. G. Boneca, F. Carbonnel, M. Chamaillard, and L. Zitvogel. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science*, 350(6264):1079–1084, 2015.

Appendix A

Supplementary Information for Chapter 3

Part I
Biogeochemical model

1 Model scope

We treat the hypolimnion as a mostly closed system of cycling chemical species. Compartments are linked by transport processes, and reactions within compartments interconvert chemical species. Interaction with the environment outside the hypolimnion is represented by just two processes: input of oxidizable carbon in the upper compartments and input of methane from the sediment. The sediment and metalimnion are otherwise ignored.

The model also treats the processes in the lake as temporally symmetrical. In other words, the rules that govern the evolution of the lake's geochemistry are fixed through time. In a real lake, season-long trends like changes in sunlight and atmospheric temperature will affect properties of the lake, notably, the depth of the thermocline. The model presented here assumes that the size and behavior of the hypolimnion is fixed, which simplifies its construction but might limit its ability to recognize season-long trends that emerge on account of those altered properties.

1.1 The model only treats the lake's hypolimnion

In order to develop a model that captured as much of the lake's biogeochemical dynamics as possible while still remaining simple and conceptual, we developed a model that only treats the hypolimnion. We limited the model in this way because most of the lake's vertical distance is in the hypolimnion (17 out of 22 meters), because modeling the epilimnion presents very challenges from modeling the hypolimnion, and because modeling the two parts of the lake together is even more complicated.

- the epilimnion and hypolimnion are somewhat decoupled by the resistance to mixing across the thermocline
- a successful treatment of the two parts of the lake will require a model of the thermal properties of the lake
- the effects of wind and precipitation are much stronger in the epilimnion

- rivers move water into and out of the lake's epilimnion
- light penetrates in the epilimnion and delivers different amounts of energy to its different depths
- microbial growth is not overall limited by energy in the epilimnion like it is in the hypolimnion

2 Model

2.1 Chemical species

We intended to create a general model of the lake's seasonal chemical and biological dynamics, and we therefore simplified the number of chemical species used in the model. We aimed to simplify the model so that

- the relevant parameters and chemical species were better matched to the variations expected in dimictic lakes, rather than in aquifers,
- the dynamics were easier to interpret and match with biological measurements,
- the model avoided making predictions about chemical species that we did not or could not measure,
- the model had fewer parameters, but also so that
- the model remained a faithful representation of the key biogeochemical processes occurring the lake.

Supplementary Table 1 lists all chemical species simulated. Relative to the original model (1), we made the following changes:

- We used a single carbon species. The original model allows for two types of carbon, DOC and POC. Each type of carbon can consist of up to four species, each with their own degradation kinetic constant. Although we do expect that there are many kinds of carbon in this ecosystem (e.g., particulate biomass, humics from rainwater runoff, photosynthetic algae), some simulations in the original publication use only one carbon species. We also found that a single carbon species was sufficient to reproduce the expected dynamics. To minimize the possibility for overfitting, we used only one.

- We simplified the sulfur compounds from five species to two be eliminating S^0 and FeS and combining HS^- and S^{2-}. We could not directly measure either of the eliminated species, and our measurements did not distinguish between the two reduced species. The process producing S^0 has a kinetic constant two orders of magnitude smaller than the kinetic constant for process that competes for H_2S, and S^0 is inert in the original model, so we expected its removal would not dramatically affect the dynamics. FeS, aside from being unmeasured, is produced in the original model by one of a set of non-redox mineral precipitation-dissolution reactions that we exclude for other reasons, as described below. The two reduced sulfur species interconvert in the original model in one of a set of acid dissociation reactions that we exclude for other reasons, as described below.

- We eliminated manganese compounds. Manganese, a terminal electron acceptor intermediate between nitrate and iron, is present at low abundances in Mystic Lake. We therefore expect it is unimportant to the dynamics.

- We eliminated calcium species. In the original model, calcium species participate in the precipitation-dissolution and acid dissociation reactions that we excluded, as described below. We also had no measurements for calcium.

- We simplified the nitrogen species by eliminating an adsorbed variety of ammonia, since we have measurements that distinguish between the two varieties and because we did not expect that the adsorption dynamics would be important to the biological dynamics. We also ignored the nitrogen gas produced by denitrification and iron oxidation on nitrate because nitrogen fixation is typically not a prominent process in the hypolimnion.

- We make the approximation the carbon dioxide is, for the purposes of methanogenesis, ubiquitous and abundant, as is generally true in eutrophic dimictic lakes.

In some cases, there is a clear correspondence between a single species in the model and a single species in nature (e.g., M for CH_4). In other cases, a single species in the model might stand for multiple species in nature (e.g., S^- for HS^- and S^{2-}).

3 Chemical reactions

As mentioned above, we also made simplifications to the chemical processes used in the original aquifer-specific model. The set of reactions used here is show in Supplementary Table 3. Relative to the original model, we made the following changes:

- We added iron oxidation on nitrate. Previous research in this lake (2) had shown that this process is important to the lake's biogeochemistry.
- We eliminated the pH-dependent acid dissolution reactions because the pH in Mystic Lake's hypolimnion only varies between about 6 and 7.
- We also eliminated the non-redox mineral precipitation-dissolution reactions. Those reactions are important mostly because they are involved in acid-base buffering, which is important in groundwater systems but not in the biogeochemistry of Mystic Lake.
- We eliminated the adsorption reaction, as described above.
- We eliminated the reactions that included carbon dioxide and manganese because they were, as argued above, not important to the lake's biogeochemistry.

4 Caveats to these modifications

It is not our intent to assert that, because a feature is not included in the model (e.g., S^0 as an electron acceptor, sulfate reduction in aerobic zones) it is not happening in the lake or is not important in other ecosystems. We aimed instead to identify the few processes that were most critical to the development of the observed chemical gradients in the lake. We hope that the loss of interesting biological complexity is balanced by the increased clarity of a simpler model with fewer processes and parameters. For example, including POC in the model would involve creating five new chemical processes (one each for the primary oxidation half-reactions), one new transport process (since POC will have different transport dynamics than DOC), and the relevant parameters for all those processes.

5 Mechanics: Transport and reactions

The rate of change in the concentration of a chemical species X at a depth i is

$$\frac{\partial X_i}{\partial t} = (\text{transport terms}) + (\text{reaction terms}) + (\text{source terms}) \,, \qquad (1)$$

where i refers to depth in meters, i.e., low i means vertically higher in the water column. The simulation proceeds in N compartments, which we spaced at one meter to be comparable to the collected chemical and biological data. The initial concentrations are set and the simulation proceeds for a time T, during which the chemical species concentrations and reaction rates are recorded. This time roughly corresponds to the period between the movement of the thermocline up the water column in spring and the breakdown of stratification in fall.

5.1 Transport: Diffusion and settling

Most chemical species are treated as dissolved in the water column. In the time and length scales relevant to the hypolimnion ecosystem, molecular diffusion is slow compared to bulk transport processes like vertical eddy diffusion. To model these bulk transport processes, most chemical species are transported by simple diffusion with rate $D\,(X_{i-1} - X_i) + D\,(X_{i+1} - X_i)$, where the diffusion constant D is the same for all chemical species, since it represents a bulk transport process. To account for the boundaries at the metalimnion and sediment, the first term is excluded in the uppermost simulation compartment; in the lowermost compartment, the second is excluded.

To simulate the settling of particulate carbon and oxidized iron species, C and Fe^+ settle in the model. A parameter p, where $0 < p < 1$, determines the balance between vertical eddy diffusion and settling for these chemical species so that the transport rate is

$$(1+p)D\,(X_{i-1} - X_i) + (1-p)D\,(X_{i+1} - X_i)\,. \qquad (2)$$

Since $p > 0$, these species tend to move down the water column and accumulate above the sediment. As with other species, the first term in excluded in the top compartment; the second term in the bottom compartment.

5.2 Reactions

5.2.1 Biotically-catalyzed reactions: Primary oxidations

The oxidation of carbon uses a chain of progressively less energetically-favorable terminal electron acceptors. Here, we follow the formulation laid out by Hunter *et al.* (ref. 1, especially equations 3 and 4).

The total rate of carbon degradation in a compartment follows first order kinetics:

$$R^{\mathrm{C}} \equiv k^{\mathrm{C}}\mathrm{C}; \quad \left(\frac{\partial \mathrm{C}}{\partial t}\right)_{\text{reaction}} = -R^{\mathrm{C}}, \tag{3}$$

where k^{C} is a first-order rate constant. The fraction of carbon taken up by oxidation on each of the terminal electron acceptors is determined by the abundance and relative metabolic merit of the electron acceptors. The j-th electron acceptor is consumed at a rate

$$R_j = \frac{f_j}{e_j} R^{\mathrm{C}}, \tag{4}$$

where e_j is the number of electrons neutralized per electron acceptor molecule and f_j is determined by successive applications of the formula

$$f_j = \left(1 - \sum_{k=1}^{j-1} f_k\right) \max\left\{1, \frac{[\mathrm{EA}_j]}{[\mathrm{EA}_{\mathrm{lim},j}]}\right\} \tag{5}$$

for $j \in \{1, 2, 3, 4\}$. If the j-th electron acceptor's concentration $[\mathrm{EA}_j]$ is greater than some constant limiting concentration $[\mathrm{EA}_{\mathrm{lim},j}]$, then that electron acceptor gets all the remainder of the carbon; otherwise, it gets a fraction of what is left determined by the ratio of the two concentrations.

The electron acceptors and their e_j are listed in Supplementary Table 2. Methanogenesis corresponds to $j = 5$, and gets all remaining carbon so that $f_5 = 1 - \sum_{k=1}^{4} f_k$. All the carbon allocated by R^{C} gets used up (i.e., $\sum_{j=1}^{5} f_j = 1$), but each electron acceptors accepts electrons according to a different stoichiometry (i.e., $\sum_{j=1}^{5} R_j \neq R^{\mathrm{C}}$).

5.3 Secondary oxidations

We model secondary oxidations, the oxidation of compounds other than carbon compounds, using second-order mass action kinetics as per Hunter *et al.*

(ref. 1's Table 4). For the transformation of substrates S_1, S_2 into a product P according to $a_1 S_1 + a_2 S_2 \to bP$, the reaction rate is $r \equiv k[S_1][S_2]$ and the reaction terms are

$$\left(\frac{\partial[P]}{\partial t}\right)_{\text{reaction}} = br \tag{6}$$

$$\left(\frac{\partial[S_i]}{\partial t}\right)_{\text{reaction}} = -a_i r \quad (i = 1, 2) \tag{7}$$

$$\tag{8}$$

with rate constant k. As per Hunter *et al.*, we do not adjust the rate according to the reaction's stoichiometry.

Primary and secondary oxidations are listed in Supplementary Table 3.

5.4 Source terms

Interactions between the hypolimnion and the outside world are modeled by simple source terms. Oxygen and carbon are added at the thermocline. Methane can be produced by primary oxidation in the water column, but methanogenesis also proceeds in the sediment, whence it is transported upward and consumed by methanotrophy. We model this process by a point source of methane in the sediment. All methane in our model is consumed before reaching the thermocline, so we omit the mechanics for emission of methane into the metalimnion.

5.5 Parameterization

A list of parameters and their values is included in Supplementary Table 4. Where possible, parameters related to the reaction rates were borrowed from Hunter *et al.* (1) and, in some cases, adjusted by hand. The parameters related to transport, source terms, and initial concentrations were drawn from published data where possible. Other values were adjusted by hand to match the observed data.

6 Implementation

The model was implemented in Matlab, and the ODE solutions were computed using the command `ode15s` with all chemical species restricted to

nonnegative values (command `odeset`).

7 Inferred biomass

In the compartment at depth d there are n_r biotically-catalyzed reactions with rates $R_i(d)$, where $i \in \{1, \ldots, n_r\}$. We define the *relative rate* of the i-th reaction at depth d as

$$r_i(d) \equiv \frac{R_i(d)}{\sum_{j=1}^{n_r} R_j(d)}. \tag{9}$$

In the inferred biomass framework, we assert that the biomass $b_i(d)$ of the organisms catalyzing the i-th process across depths d is proportional to the relative rates $r_i(d)$, that is,

$$b_i(d) = \alpha_i r_i(d) \tag{10}$$

for all d, where α_i is some constant of proportionality that relates biomass to relative rate that varies with i (the organisms catalyzing different rates) but not with depth. Because the α_i are unknown, we never infer the relationship between abundances of different biomasses, even at the same depth. Furthermore, $b_i(d)$ is the biomass of *all* organisms catalyzing process i—the inferred biomass framework does not provide information about the combined biomass of individual taxa, only the biomass of all taxa catalyzing a modeled process.

In comparisons with the survey data, we used relative rates $r_i(d)$ rather than absolute rates $R_i(d)$ because survey count data show relative, not absolute, abundances. For example, consider a process i whose absolute rate is higher at depth x than at depth y, i.e., $R_i(x) > R_i(y)$. Other processes, however, are much more active at x than at y so that although i's absolute rate is higher at x, its relative rate is higher at y, i.e., $r_i(x) < r_i(y)$. In this case, we would expect the biomass of organisms performing process i to have higher absolute abundance at x but higher relative abundance at y.

References

[1] Hunter K, Wang Y, Van Cappellen P (1998) Kinetic modeling of microbially-driven redox chemistry of subsurface environments: coupling transport, microbial metabolism and geochemistry. *J Hydrol (Amst)* 209:53–80.

[2] Senn DB, Hemond HF (2002) Nitrate controls on iron and arsenic in an urban lake. *Science* 296(5577):2373–2376.

[3] US Geological Survey (2014) National Water Information System (USGS Water Data for the Nation).

[4] Wetzel R (2001) *Limnology.* (Academic Press), 3rd edition.

[5] Senn D (2001) Ph.D. thesis (MIT).

[6] Benoit G, Hemond H (1996) Vertical eddy diffusion calculated by the flux gradient method: Significance of sediment-water heat exchange. *Limnol Oceanogr* 41:157–168.

[7] Varadharajan C, Hemond H (2012) Time-series analysis of high-resolution ebullition fluxes from a stratified, freshwater lake. *J Geophys Res* 117(G2).

[8] Peterson E (2005) Master's thesis (MIT).

[9] Oh H et al. (2011) Complete genome sequence of strain IMCC9063, belonging to SAR11 subgroup 3, isolated from the Arctic Ocean. *J Bacteriol* 193(13):3379–3380.

[10] Weon H et al. (2009) *Solitalea koreensis* gen. nov., sp. nov. and the reclassification of [*Flexibacter*] *canadensis* as *Solitalea canadensis* comb. nov. *Int J Syst Evol Microbiol* 59(8):1969–1975.

[11] Finneran K, Johnsen C, Lovley D (2003) *Rhodoferax ferrireducens* sp. nov., a psychrotolerant, facultatively anaerobic bacterium that oxidizes acetate with the reduction of Fe(III). *Int J Syst Evol Microbiol* 53(3):669–673.

[12] Nevin KP et al. (2005) *Geobacter bemidjiensis* sp. nov. and *Geobacter psychrophilus* sp. nov., two novel fe(III)-reducing subsurface isolates. *Int J Syst Evol Microbiol* 55(4):1667–1674.

[13] Balk M, Altınbaş M, Rijpstra W, Damsté J, Stams A (2008) *Desulfatirhabdium butyrativorans* gen. nov., sp. nov., a butyrate-oxidizing, sulfate-reducing bacterium isolated from an anaerobic bioreactor. *Int J Syst Evol Microbiol* 58(1):110–115.

[14] Purkhold U, Wagner M, Timmermann G, Pommerening-Röser A, Koops H (2003) 16S rRNA and *amoA*-based phylogeny of 12 novel betaproteobacterial ammonia-oxidizing isolates: extension of the dataset and proposal of a new lineage within the nitrosomonads. *Int J Syst Evol Microbiol* 53(5):1485–1494.

[15] Alawi M, Lipski A, Sanders T, Pfeiffer E, Spieck E (2007) Cultivation of a novel cold-adapted nitrite oxidizing betaproteobacterium from the Siberian Arctic. *ISME J* 1(3):256–264.

[16] Blöthe M, Roden E (2009) Composition and activity of an autotrophic Fe(II)-oxidizing, nitrate-reducing enrichment culture. *Appl Environ Microbiol* 75(21):6937–6940.

[17] Emerson D, Moyer C (2002) Neutrophilic Fe-oxidizing bacteria are abundant at the loihi seamount hydrothermal vents and play a major role in Fe oxide deposition. *Appl Environ Microbiol* 68(6):3085–3093.

[18] Lapidus A et al. (2011) Genomes of three methylotrophs from a single niche reveal the genetic and metabolic divergence of the *Methylophilaceae*. *J Bacteriol* 193(15):3757–3764.

[19] Wartiainen I, Hestnes A, McDonald I, Svenning M (2006) *Methylobacter tundripaludum* sp. nov., a methane-oxidizing bacterium from Arctic wetland soil on the Svalbard islands, Norway (78 degrees N). *Int J Syst Evol Microbiol* 56(1):109–113.

Part II
Supplementary Tables and Figures

symbol	name	representative compounds
O	dissolved oxygen	O_2
C	oxidizable carbon	cyanobacteria biomass, glucose, acetate
N^+	oxidized nitrogen	nitrate, nitrite
N^-	reduced nitrogen	ammonia
Fe^+	oxidized iron	Fe(III) compounds
Fe^-	reduced iron	Fe(II)
S^+	oxidized sulfur	sulfate compounds
S^-	reduced sulfur	sulfide compounds
M	methane	CH_4

Supplementary Table 1: Chemical species included in the model.

j	EA	e_j
1	O	4
2	N^+	5
3	Fe^+	1
4	S^+	8
5	$\varnothing$	8

Supplementary Table 2: Electron acceptors in the primary oxidation reactions. $j = 5$ corresponds to methanogenesis.

Primary oxidations	rate	
$\mathrm{C} \rightarrow a\mathrm{N}^- + e\mathrm{e}^-$	R^{C}	primary oxidation half-reaction
$\mathrm{O} \rightarrow \varnothing$	R_1	aerobic heterotrophy
$\mathrm{N}^+ \rightarrow \varnothing$	R_2	denitrification
$\mathrm{Fe}^+ \rightarrow \mathrm{Fe}^-$	R_3	iron reduction
$\mathrm{S}^+ \rightarrow \mathrm{S}^-$	R_4	sulfate reduction
$\varnothing \rightarrow \mathrm{M}$	R_5	methanogenesis
Secondary oxidations	**rate constant**	
$2\mathrm{O} + \mathrm{N}^- \rightarrow \mathrm{N}^+$	k_1	ammonia oxidation
$2\mathrm{O} + \mathrm{S}^- \rightarrow \mathrm{S}^+$	k_2	sulfide oxidation
$\mathrm{N}^+ + 5\mathrm{Fe}^- \rightarrow 5\mathrm{Fe}^+$	k_3	iron oxidation on nitrate
$\mathrm{M} + 2\mathrm{O} \rightarrow \varnothing$	k_4	methanotrophy on oxygen
$\mathrm{M} + \mathrm{S}^+ \rightarrow \mathrm{S}^-$	k_5	methanotrophy on sulfate
$\frac{1}{4}\mathrm{O} + \mathrm{Fe}^- \rightarrow \mathrm{Fe}^+$	k_6	iron oxidation

Supplementary Table 3: Reactions simulated in the model.

[1]The cited database was queried for river gauge USGS 01102500 (Aberjona River, Winchester, MA; the Aberjona drains into Mystic Lake) parameter P00681 ("Organic carbon, water, filtered, milligrams per liter") during 1999-2000. The average value was 4.91 mg L^{-1}. The cited text's Table 12-4 lists C:N = 15.1 for carbon concentrations just above this in Wisconsin lakes, thus N:C = 1/15.1 = 0.066.

[2]The cited work reports an initial rate for iron oxidation on nitrate as 2.4 μM day^{-1} with initial nitrate concentration 30 μM and initial iron concentrations 10–50 μM, which corresponds to the shown range when assuming a second-order rate form.

[3]The cited paper collates reports of vertical eddy diffusion constants 0.002–0.05 cm^2 s^{-1} for lakes with depths comparable to Mystic Lake's. Diffusion constants are typically written as (time)/(length)^2, but here the compartment height, 1 meter, sets the length scale.

[4]The cited work reports an effective settling rate 0.024 m day^{-1} (v_{eff}, Table 3B). The shown value is computed by equating v_{eft} to comparable to $D \times p_{\text{Fe}}$ and using the shown value of D.

[5]The cited text gives annual organic carbon input for Wingra Lake, a polluted urban lake, as 691 g C m^{-2} yr^{-1} (Table 23-12) and for Lawrence Lake as 130.6 g C m^{-2} yr^{-1} (Table 23-13). These two values correspond to the reported range if the carbon is assumed delivered to a one-meter-high compartment. One should note that the carbon inputs reported in the literature should not directly correspond to C in the model, which is a simplified biomass.

[6]The cited work reports 1.3–4.0 mmol m^{-2} d^{-1}, which corresponds to the shown values if the methane is delivered to a well-mixed one-meter high lowest compartment as assumed in the model.

[7]Ref. 4 reports that eutrophic lakes have median total organic carbon around 12.0 mg L^{-1} = 10^3 μM (Table 23-1). The carbon concentrations later in the simulation are more similar to this value.

[8]The total concentrations of nitrogen, iron, and sulfur species were chosen to correspond with the total amount of these chemical observed in the cited work's appendix tables.

parameter	value	unit	source & reported value
General parameters			
T	0.4	yr	Asserted to set time scale
N	17	—	Asserted for 1 m compartments
Primary oxidation parameters			
k^{C} (rate constant)	1.0	yr^{-1}	(1) ($k^{\mathrm{DOC}} = 3 \times 10^{-5}$—$3 \times 10^{1}$ yr^{-1})
a (N:C ratio)	0.1	—	(3) and (4)[1] (0.066)
$[\mathrm{O}_{\mathrm{lim}}]$	20.0	μM	(1) (20.0 μM)
$\left[\mathrm{N}^{+}_{\mathrm{lim}}\right]$	5.0	μM	(1) (5.0 μM)
$\left[\mathrm{Fe}^{+}_{\mathrm{lim}}\right]$	0.1	μM	(1) (60 μmol dm^{-3})
$\left[\mathrm{S}^{+}_{\mathrm{lim}}\right]$	30.0	μM	(1) (30.0 μM)
Secondary oxidation parameters			
k_1 (ammonia oxidation)	5.0	$\mu\mathrm{M}^{-1}$ yr^{-1}	(1) ($k_4^{\mathrm{sr}} = 5 \times 10^{6}$ M^{-1} yr^{-1})
k_2 (sulfide oxidation)	0.16	$\mu\mathrm{M}^{-1}$ yr^{-1}	(1) ($k_5^{\mathrm{sr}} = 1.6 \times 10^{5}$ M^{-1} yr^{-1})
k_3 (iron oxidation, nitrate)	1.0	$\mu\mathrm{M}^{-1}$ yr^{-1}	(5)[2] (0.6–3 $\mu\mathrm{M}^{-1}$ yr^{-1})
k_4 (methanotrophy, oxygen)	10^{4}	$\mu\mathrm{M}^{-1}$ yr^{-1}	(1) ($k_9^{\mathrm{sr}} = 10^{10}$ M^{-1} yr^{-1})
k_5 (methanotrophy, sulfate)	10^{-2}	$\mu\mathrm{M}^{-1}$ yr^{-1}	(1) ($k_{10}^{\mathrm{sr}} = 10^{4}$ M^{-1} yr^{-1})
k_6 (iron oxidation)	10^{4}	$\mu\mathrm{M}^{-1}$ yr^{-1}	(1) ($k_2^{\mathrm{sr}} = 10^{7}$ M^{-1} yr^{-1})
Transport parameters			
D	50	yr^{-1}	(6)[3] (6–158 yr^{-1})
p_{Fe} (settling for oxidized iron)	0.3	—	(5)[4] (0.18)
p_{C} (settling for biomass)	0.3	—	Manually adjusted
Source rates			
s_{C}	9.4×10^{4}	$\mu\mathrm{M}$ yr^{-1}	(4)[5] (1.3×10^{4}—6.9×10^{4} $\mu\mathrm{M}$ yr^{-1})
s_{O}	6.6×10^{3}	$\mu\mathrm{M}$ yr^{-1}	Manually adjusted
s_{M}	2830	$\mu\mathrm{M}$ yr^{-1}	(7)[6] (475–1460 $\mu\mathrm{M}$ yr^{-1})
Initial concentrations			
[O]	50	μM	Manually adjusted
[C]	200	μM	Manually adjusted[7]
$[\mathrm{N}^{+}] + [\mathrm{N}^{-}]$	100	μM	(8)[8]
$[\mathrm{N}^{+}]/[\mathrm{N}^{-}]$	10	—	Manually adjusted
$[\mathrm{Fe}^{+}] + [\mathrm{Fe}^{-}]$	60	μM	(8)
$[\mathrm{Fe}^{+}]/[\mathrm{Fe}^{-}]$	10	—	Manually adjusted
$[\mathrm{S}^{+}] + [\mathrm{S}^{-}]$	250	μM	(8)
$[\mathrm{S}^{+}]/[\mathrm{S}^{-}]$	10	—	Manually adjusted
[M]	0	μM	Manually adjusted

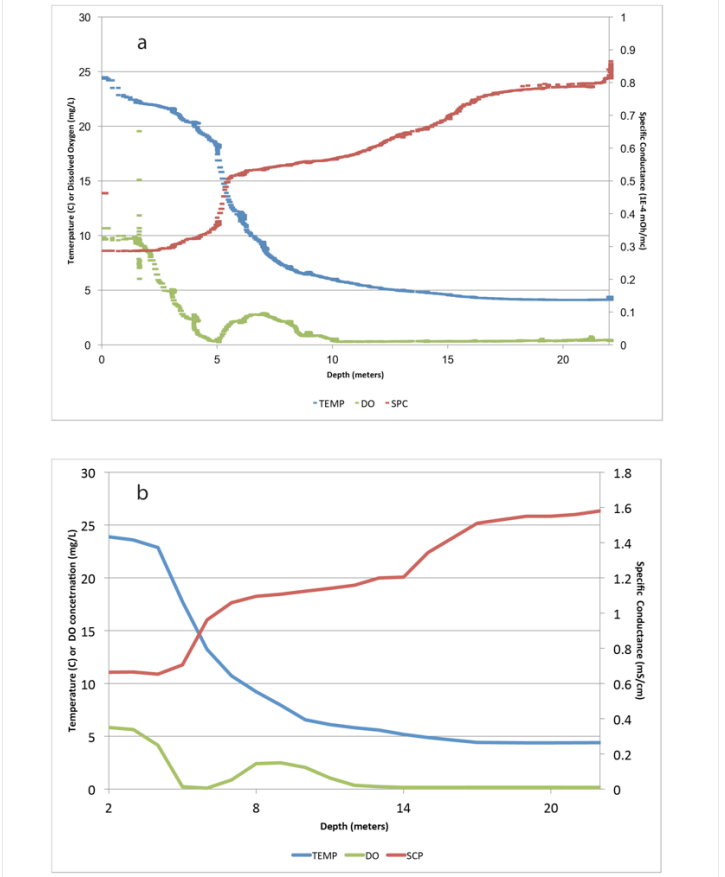
Supplementary Table 4: Parameter values and sources.

comparison	pairs in real data	pairs in shuffled data				
		mean	std	min	max	p-value
timepoints	190	68.2	8.4	49	91	10^{-109}
sample prep	777	101.4	9.3	77	129	$\sim$0
OEU callers	2564	131.3	11.9	96	172	$\sim$0

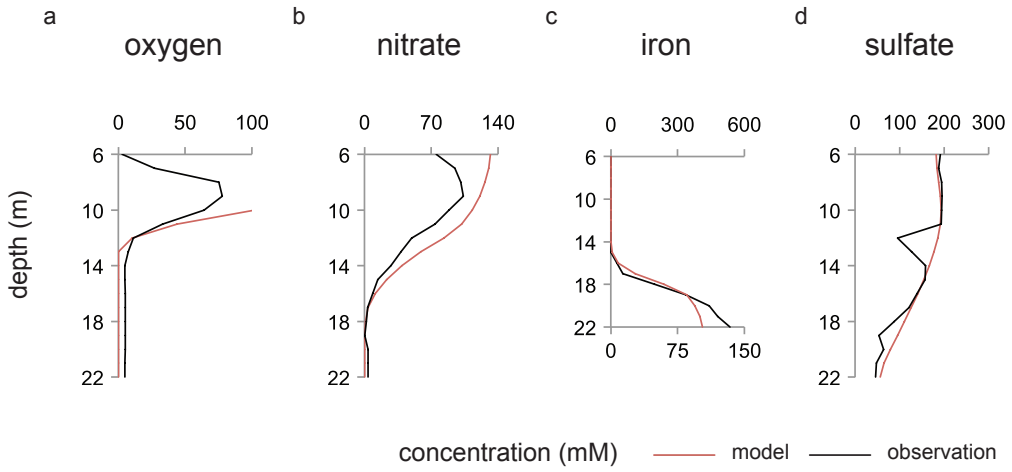
Supplementary Table 5: Numbers of pairs of OTUs that were in the same OEU across datasets. Values for shuffled data represent 1000 random shufflings of the assignments of OTUs to OEUs in one of the two datasets in the comparison. The p-value represent the results of one-sided z-tests using the mean and standard deviations of the number of OTU pairs from the shuffled datasets. (The $\sim$0 indicates values much less than 10^{-109}.)

function (OTU ID in 2008; 2013)	closest genome or type strain (reference)	identity (%)	clone GI (accession)
Aerobic heterotrophy (31; 1)	*Candidatus Pelagibacter* sp. IMCC9063 (9)	91	444189431 (KC192422)
Denitrification (298; 188)	*Solitalea canadensis* DSM 3403 (10)	90	444189414 (KC192405)
Iron reduction (11; 12)	*Rhodoferax ferrireducens* T118 (11)	99	444189437 (KC192428)
Iron reduction (13; 228)	*Geobacter psychrophilus* strain P35 (12)	96	444189534 (KC192525)
Sulfate reduction (88; 106)	*Desulfatirhabdium butyrativorans* strain HB1 (13)	92	444189499 (KC192490)
Ammonia oxidation (16-2; 39)	*Nitrosospira briensis* (14)	98	444189385 (KC192376)
Nitrite oxidation (141-2; 29)	*Candidatus Nitrotoga arctica* (15)	99	444189434 (KC192425)
Sulfide and iron oxidation (104; 6)	*Sideroxydans lithotrophicus* ES-1 (16)	95	444189494 (KC192485)
Sulfide and iron oxidation (125; 223)	*Sideroxydans lithotrophicus* ES-1 (17)	99	444189496 (KC192487)
Methane oxidation (99; 68)	*Methylotenera versatilis* 301 (18)	94	444189407 (KC192398)
Methane oxidation (276; 8)	*Methylobacter psychrophilus* (19)	99	444189514 (KC192505)

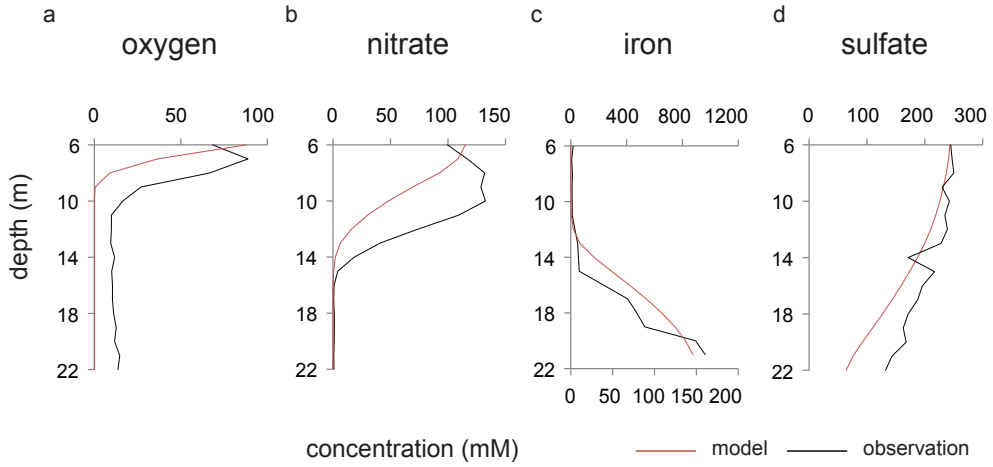
Supplementary Table 6: Reference OTUs.



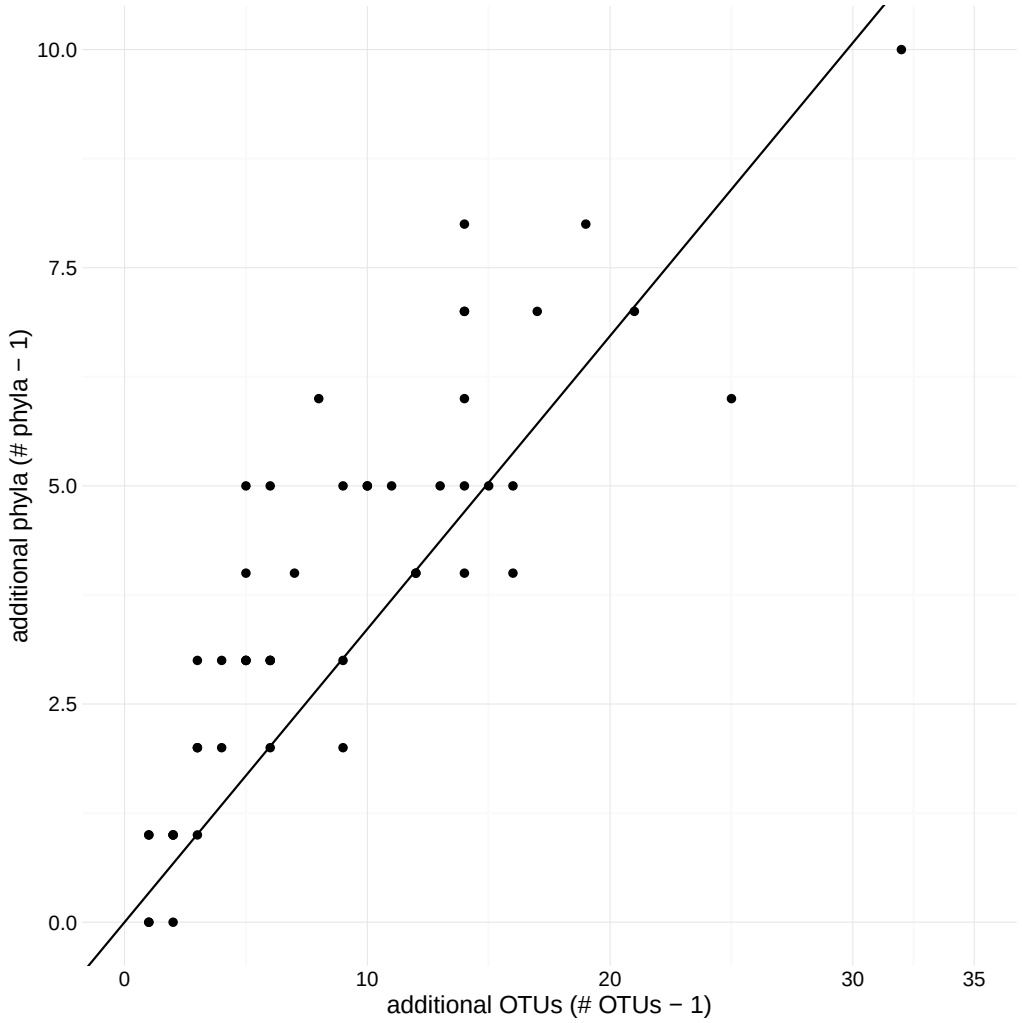
Supplementary Figure 1: *In situ* measurements taken during sample collection on (*a*) Aug. 13, 2008 and (*b*) Aug. 15, 2013. Temperature (°C, blue), dissolved oxygen (DO, mg/L; green), and specific conductance (SCP mS/cm; red) are shown.



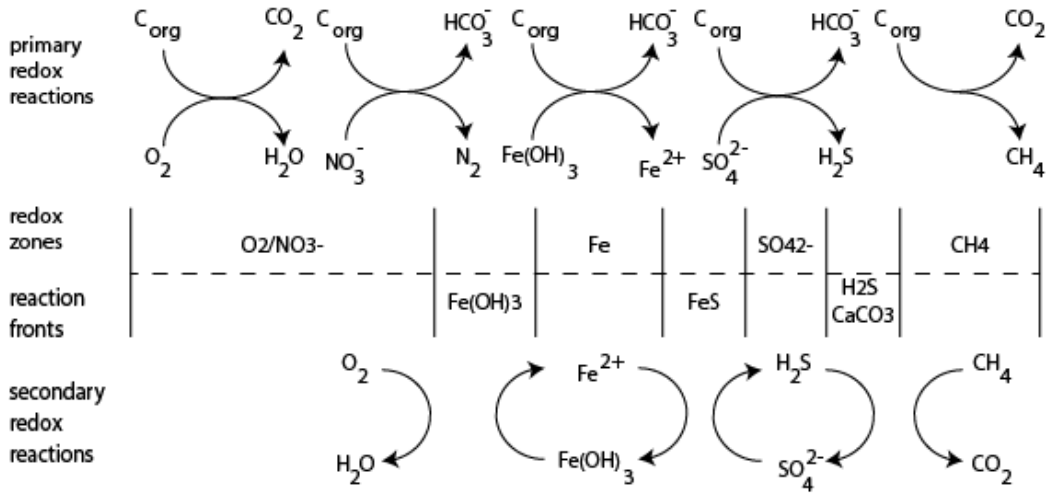
Supplementary Figure 2: Correspondence between 2013 chemical observations and model predictions after the model was calibrated to match the chemical observation. Modeled concentrations (red) are on the same scale as observations (black) except for iron (top axis, observed; lower axis, modeled).



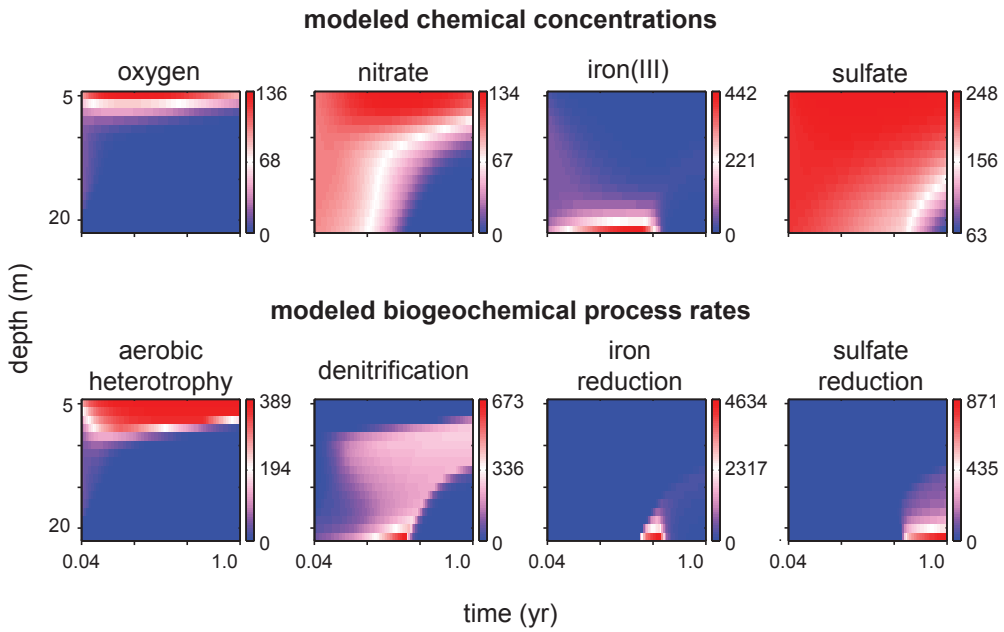
Supplementary Figure 3: Correspondence between 2008 chemical observations and model predictions after the model was calibrated to match the chemical observation. Modeled concentrations (red) are on the same scale as observations (black) except for iron (top axis, observed; lower axis, modeled).



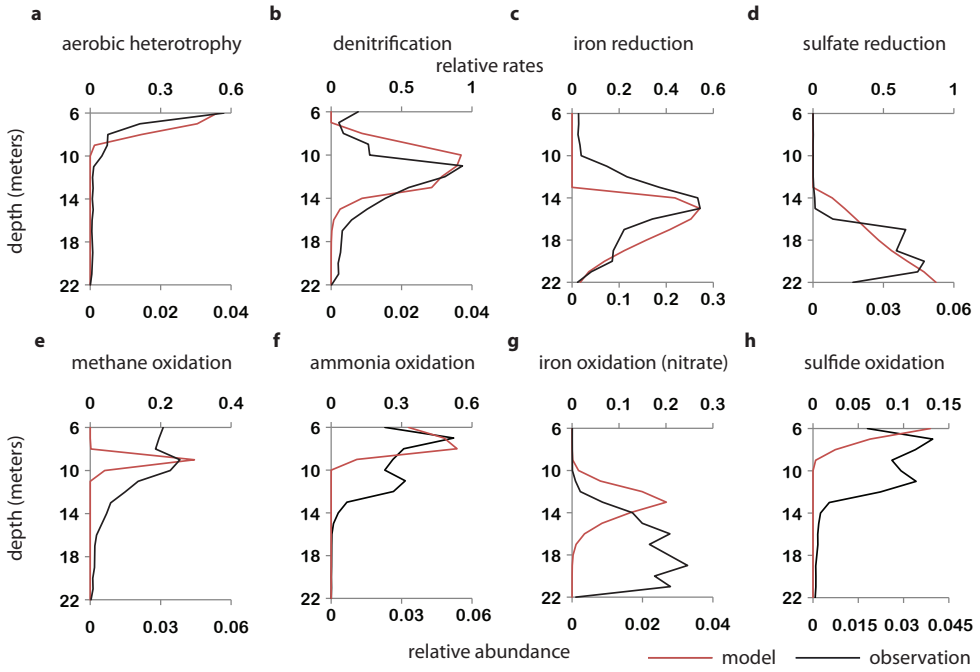
Supplementary Figure 4: The number of phyla within OEUs increased linearly with the number of OTUs in the OEU. The regression was constrained so that 1 OTU in an OEU necessarily produced 1 phylum in that OEU.



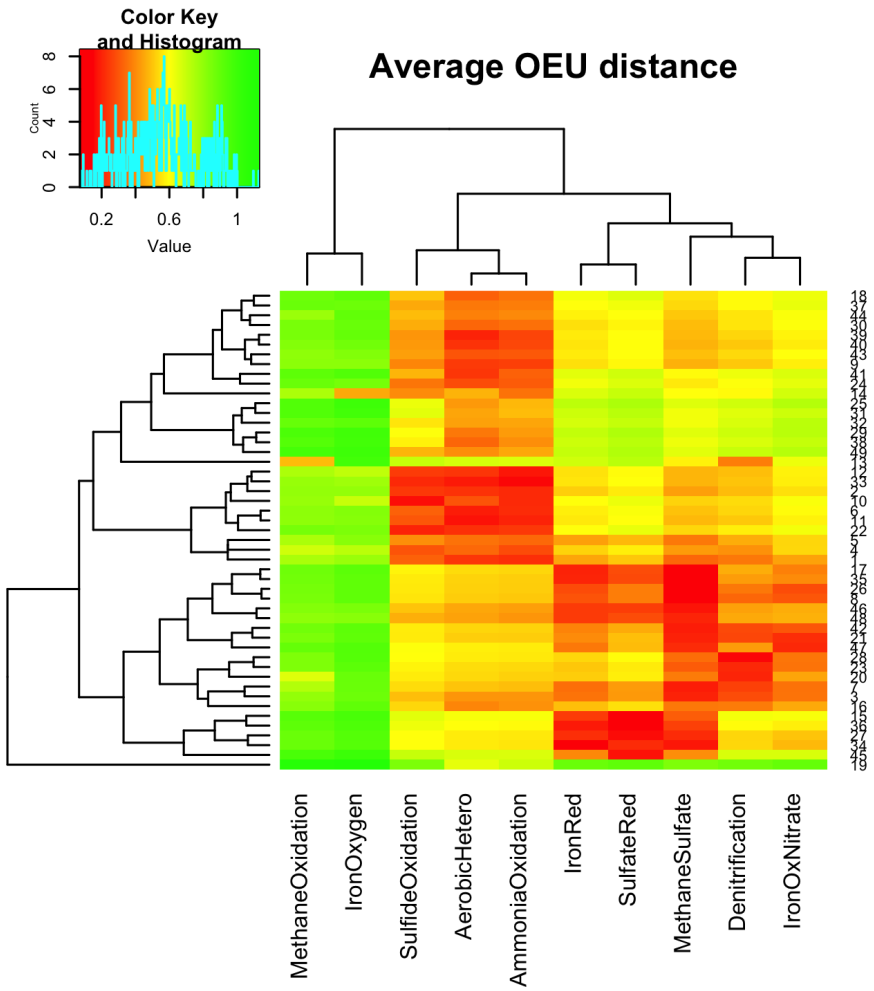
Supplementary Figure 5: Primary and secondary oxidation-reduction (redox) reactions used in the biogeochemical model, along with the typical redox zones and reaction fronts. Adapted from (1).



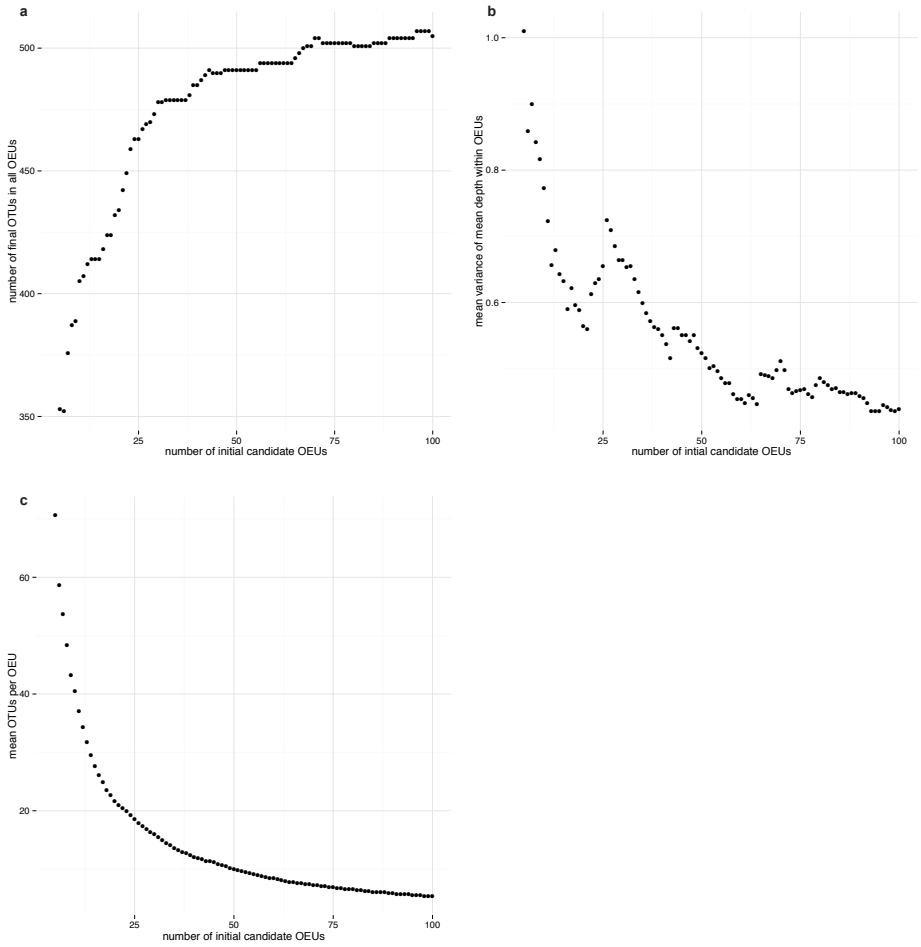
Supplementary Figure 6: The model predicts the spatial and temporal dynamics of the lake's chemistry as well as the places and times in the lake that favor specific biogeochemical processes. The model starts in a homogeneous state, representing an idealized, fully-mixed dimictic lake in the spring of 2008. Concentrations (top) of electron acceptors are measured in μM; the rates of biogeochemical processes (bottom) are measured in μM yr^{-1}.



Supplementary Figure 7: Observed distribution of key populations (black, bottom x-axis, measured in relative abundance) and their correspondence with modeled processes (red, top x-axis, measured in relative rate) from 2008. The observations, which reflect the abundance of organisms, and the model results, which reflect the predicted prevalence of a metabolic process, have peaks at similar locations and their distributions roughly correspond, which suggests that the distribution of these organisms is largely determined by the favorability of the corresponding metabolic process.



Supplementary Figure 8: Average Euclidean distance across OTUs within each OEU (rows; 2013 data) to the modeled processes (columns; 2013 model), demonstrating the relationship of OEUs with processes. Dendrograms show the clustering patterns of modeled processes (top) and OEUs (left). OEUs cluster largely by their relationship to aerobic and anaerobic processes.



Supplementary Figure 9: (a) The number of OTUs remaining in the final analysis after OEUs are called increases with the number of initial candidate OEUs. (b) Each OTU is distributed across the lake's depth, so each OTU has a mean depth. For every OEU, the variance in mean depths was computed. For each set of OEUs produced by a different number of initial candidate OEUs, the mean of those variances was computed. The mean (across OEUs) of variances (within an OEU) of the mean depth (of each OTU) is a measure of overall cluster quality. The mean variance decreases with increasing initial OEU number. (c) The mean number of OTUs in each OEU decreases with increasing initial OEU number.

Appendix B

Supplementary Information for Chapter 4

Part I
Model design

In the main text, we mentioned three main parameters: the placebo rate p_{placebo} (hereafter shortened to p_{pl}), the efficacy of efficacious stool p_{eff}, and the frequency of efficacious donors f_{eff}.

As will be discussed below, correctly relating p_{pl} and p_{eff} requires a hidden parameter. This parameter is not important for the simulations in the main text, but it is important when drawing random variates $(p_{\text{pl}}, p_{\text{eff}})$ from a distribution.

1 Placebo rate and treatment efficacy

In our model, a patient in the treatment arm can respond to the treatment under one of three scenarios:

1. the donor was inefficacious and the patient responded because of the placebo effect,
2. the donor was efficacious and the patient would have responded even if they had been given a placebo, or
3. the donor was efficacious and the patient would *not* have responded if they had been given only the placebo.

We account for these three possibilities by treating the two effects—an effect from efficacious stool and the placebo effect—as independent.

Specifically, we say that efficacious stool has some "active ingredient" that, in the absence of a placebo effect, would cause a fraction p_{ing} of patients to respond. The efficacy rate p_{eff} of treatment with efficacious stool will be higher than this active ingredient efficacy p_{ing} because a patient can respond to the placebo. If a fraction p_{pl} of patients would respond to the placebo alone, then:

$$1 - p_{\text{eff}} = (1 - p_{\text{ing}})(1 - p_{\text{pl}}), \tag{1}$$

that is, when a patient is administered efficacious stool, the patient does not respond only if they do not respond to the active ingredient *and* they do not respond to the placebo. Eq. (1) is equivalent to

$$p_{\text{eff}} = p_{\text{ing}} + p_{\text{pl}} - p_{\text{ing}}p_{\text{pl}}. \tag{2}$$

2 Drawing random variates

2.1 Distribution of parameters values

In the simplest use case for the model, the model parameters are fixed point estimates. In a more complex use, the parameters are each drawn from a probability distribution, allowing for some uncertainty about the exact values of the parameters while simulating. "Drawing from" a distribution could either mean drawing random sets of parameters or specifying a Bayesian prior distribution for the parameter.

In this study, we used the beta distribution for the parameters because:

- The parameters are all probabilities of Bernoulli trials (i.e., a "coin flip" that answers, for example, whether a patient responds to treatment) and the beta distribution is the conjugate prior for the binomial distribution.

- The hyperparameters of the beta distribution are easy to interpret. $\text{Beta}(s, f)$ corresponds to s Bernoulli trial successes and f failures. It has a "strength" of $s+f$ (i.e., the prior data and the collected data will have equal weight in determining the posterior after $s + f$ "coin flips" have been observed in the experimental data) and places the mean estimate of the parameter at $s/(s + f)$.

In clinical practice, it is the patient who responds to treatment or not. For the purposes of the mathematics, it is more convenient to say that the *donor* succeeded (i.e., a patient treated with stool from that donor responded) or failed (i.e., the patient did not respond to treatment).

For example, consider the clinical trial results [3] presented in main text, where 1 of 6 donors appeared efficacious. The point estimate for f_{eff} is just 1/6, but finding 1 of 6 donors efficacious would be inconsistent with, say, $f_{\text{eff}} = 1/3$. It could be more conservative, then, to run simulations in which:

- f_{eff} is drawn from $\text{Beta}(A_{\text{feff}}, B_{\text{feff}})$, where A_{feff} is the number of efficacious donors previously observed and B_{feff} is the number of inefficacious donors

- p_{pl} is drawn from $\text{Beta}(A_{\text{pl}}, B_{\text{pl}})$, where A_{pl} is the number of patients in the placebo arm who responded and B_{pl} is the number who did not respond, and

Clinical trial result	Parameter	Point estimate	A	B
2 of 37 in placebo arm responded	p_{placebo}	$2/37 = 0.054$	2	35
1 of 6 donors appeared efficacious	f_{eff}	$1/6 = 0.17$	1	5
7 of 18 patients allocated to the efficacious donor responded	p_{eff}	$7/18 = 0.39$	7	11

Table 1: Point estimates and hyperparameters for model parameters using clinical data [3]. Compare against main text Table 1. A and B are hyperparameters for a beta distribution.

- p_{eff} is drawn from $\text{Beta}(A_{\text{peff}}, B_{\text{peff}})$, where A_{peff} is the number of patients in the *treatment* arm who responded and B_{peff} is the number who did not respond.

Specifically, we define the distribution on the model parameters as:

$$P_{\text{pl,ing}}(p_{\text{pl}}, p_{\text{ing}}) = \text{Beta}(p_{\text{pl}}; A_{\text{pl}}, B_{\text{pl}}) \times \text{Beta}(p_{\text{eff}}; A_{\text{peff}}, B_{\text{peff}}) \tag{3}$$

$$P_{\text{feff}}(f_{\text{eff}}) = \text{Beta}(f_{\text{eff}}; A_{\text{feff}}, B_{\text{feff}}), \tag{4}$$

where the hyperparameters A and B are shown in Table 1.

2.2 Dependence of parameters

Note that, although f_{eff} is treated as independent from p_{pl} and p_{ing}, the priors on p_{pl} and p_{eff} constitute a joint distribution on p_{pl} and p_{ing}. It is *not* true that p_{pl} can be drawn from $\text{Beta}(A_{\text{pl}}, B_{\text{pl}})$ and p_{eff} separately drawn from $\text{Beta}(A_{\text{peff}}, B_{\text{peff}})$ because this will violate the requirement that $p_{\text{eff}} > p_{\text{pl}}$. We sampled $(p_{\text{pl}}, p_{\text{eff}})$ using rejection sampling:

1. Draw p_{pl}, p_{ing}, and a threshold T from the uniform distribution on $[0, 1]$.
2. Compute $p_{\text{eff}} = p_{\text{pl}} + p_{\text{ing}} - p_{\text{pl}}p_{\text{ing}}$.
3. Accept the pair $(p_{\text{pl}}, p_{\text{ing}})$ if

$$\text{Beta}(p_{\text{pl}}; A_{\text{pl}}, B_{\text{pl}}) \times \text{Beta}(p_{\text{eff}}; A_{\text{peff}}, B_{\text{peff}}) < T. \tag{5}$$

In our implementation, we speed up this sampling by using numerical optimization to find the maximum M of the product of the two beta distributions and then draw T from $[0, M]$.

3 Incorporating inefficacious donors as placebos

Here we articulated a prior on the placebo rate p_{pl} using only the results from the placebo arm. We could also have lumped the patients assigned to inefficacious donors, who we assert respond to efficacious treatment with probability p_{pl}, in with patients from the the placebo arm, who also respond with probability p_{pl}. In the clinical trial whose results are summarized in Table 1, 2 of 20 patients assigned to the 5 apparently inefficacious donors responded [3], so the prior on p_{pl} would be informed by $2 + 2 = 4$ successful and $35 + 18 = 53$ failed placebo-or-inefficacious treatments.

Part II
Utility & decision theory

4 Theory

The model of differences in donor stool efficacy that we laid out can be used as a framework for optimizing non-adaptive trial designs with respect to some *utility function*, i.e., a single number that encodes the desirability of some outcome. In the main text, we reported on statistical powers, which are the expected utilities of a utility function that assigns a utility 1 to trials with $p < 0.05$ and a utility 0 to other trials. A researcher might be interested in optimizing trial design with respect to some other utility, say, the number of efficacious donors identified.

5 Methods

We used the model to optimize the design of a forthcoming two-stage Phase II clinical trial that uses FMT to treat ulcerative colitis. In this trial, the researchers were limited to a non-adaptive trial design in the trial's first stage, but the results of the first stage could be used to assign donors in the second stage. Therefore, the trial's researchers aimed to optimize the probability that they would be able to identify and discover an efficacious donor in the

first stage so that they could use that donor in the second stage. [swo: cite Russell?]

There were 30 patients in each arm. We evaluated the following allocations: use one donor for all patients, use one donor for every two patients, one for every three, and so forth, up to using a new donor for every patient. Using the fixed model parameters shown in Table 1, 10,000 simulations were performed for each allocation and the number of trials in which the the donor with the best ratio of successes to failures was tabulated. (In the case of a tie among donors, one of the best donors was selected at random.)

To evaluate the robustness of these results, the same simulations were performed but, instead of using point estimates for the model parameters, drawing the model parameters from the distributions shown in Table 1.

6 Results

The trial designs most likely to produce a "best" donor (i.e., with the best ratio of responding patients to patients treated) assigns 2 or 3 patients to each donor (i.e., evenly distributes patients across 10 or 15 donors; Table 2, column "point estimates"). The difference between the best strategy (2 or 3 patients per donor) and the worst strategy (one single donor for all patients) was large: the best strategy was 62% likely to provide a best donor who was also an efficacious donor, while the worst strategy was only 17% likely. Using a new donor for every patient produced intermediate results (58% likely to identify an efficacious donor).

When including this uncertainty in the model parameters, the ranking of strategies remains about the same (e.g., 3 patients per donor is optimal), but the probability that the best donor is an efficacious one decreases for most strategies (Table 2, column "distributions").

Part III
Bayesian assignment strategies

Adaptive donor assignment strategies aim to use the information derived from the patients' outcomes—and possibly some *a priori* beliefs about the values of the model parameters—to make decisions about how to assign

	Probability (%)	
N_{donors}	point estimates	distributions
30	58	32
15	62	39
10	62	40
6	56	38
5	54	37
3	41	31
2	32	25
1	17	17

Table 2: Probability that a donor allocation yields a best donor that is actually efficacious. All confidence intervals were within 1% of the reported value.

donors. The donor assignment problem, then, is amenable to a Bayesian treatment. For example, the myopic Bayesian algorithm reported in the main text (and described in more detail below) assigns donors such that the patient has the highest probability of responding to treatment given our prior expectations about donors' efficacies and the data so far accumulated during the trial. In Bayesian statistics, this "updated" probability is a posterior predictive probability.

7 Predictive posterior distributions

7.1 Derivation of predictive posterior

To derive the posterior predictive probability, we first need to articulate the posterior probability on the core parameters p_{pl}, p_{ing}, and f_{eff} as well as a new parameter that indicates whether each donor is efficacious or not. Let $\boldsymbol{q}$ be a vector of entries "efficacious" or "inefficacious" with a length equal to the number of donors. Let q_i be called the *quality* of the i-th donor. The posterior probability is then

$$P(p_{\text{pl}}, p_{\text{ing}}, f_{\text{eff}}, \boldsymbol{q}|\boldsymbol{X}) \propto \underbrace{P(\boldsymbol{X}|p_{\text{pl}}, p_{\text{ing}}, f_{\text{eff}}, \boldsymbol{q})}_{\text{likelihood}} \times \underbrace{P(p_{\text{pl}}, p_{\text{ing}}, f_{\text{eff}}, \boldsymbol{q})}_{\text{prior}}, \quad (6)$$

where $\boldsymbol{X}$ is the data: the number of patients who responded (or not) for each donor. More specifically, we say that donor i's stool was administered to n_i patients, of which s_i patients responded.

As per Section 2, we assume that the prior can be separated into parts:

$$P(p_{\text{pl}}, p_{\text{ing}}, f_{\text{eff}}, \boldsymbol{q}) = P_{\text{pl,ing}}(p_{\text{pl}}, p_{\text{ing}}) P_{\text{feff}}(f_{\text{eff}}) P(\boldsymbol{q}|f_{\text{eff}}), \tag{7}$$

that is, that the placebo rate p_{pl} and active ingredient efficacy p_{ing} are independent of the frequence of efficacious donors f_{eff}, which is in turn independent of the qualities $\boldsymbol{q}$ of the particular donors. The prior on $\boldsymbol{q}$ can be broken up by donor because the donors' qualities are all independent of one another:

$$P(\boldsymbol{q}|f_{\text{eff}}) = \prod_{i=1}^{D} \begin{Bmatrix} f_{\text{eff}} & \text{if } q_i \text{ is efficacious} \\ 1 - f_{\text{eff}} & \text{if not} \end{Bmatrix}, \tag{8}$$

where D is the number of donors. We assume that the outcomes of the patients are independent, so the likelihood of the data $\boldsymbol{X}$ is probability of the observed combinations of successes and failures for each donor:

$$P(\boldsymbol{X}|p_{\text{pl}}, p_{\text{ing}}, f_{\text{eff}}, \boldsymbol{q}) = \prod_{i=1}^{D} \begin{Bmatrix} \text{Bin}(s_i; n_i, p_{\text{eff}}) & \text{if } q_i \text{ is efficacious} \\ \text{Bin}(s_i; n_i, p_{\text{pl}}) & \text{if not} \end{Bmatrix}, \tag{9}$$

where $\text{Bin}(s_i; n_i, p_{\text{eff}})$ is the probability mass function of the binomial distribution with s_i successes out of n_i trials with probability of success p_{eff}. Equations (6), (8) and (9) can be combined:

$$\begin{aligned} P(p_{\text{pl}}, p_{\text{ing}}, f_{\text{eff}}, \boldsymbol{q}|\boldsymbol{X}) \propto & \\ \prod_{i=1}^{D} & \begin{Bmatrix} f_{\text{eff}} \times \text{Bin}(s_i; n_i, p_{\text{eff}}) & \text{if } q_i \text{ is ``efficacious''} \\ (1 - f_{\text{eff}}) \times \text{Bin}(s_i; n_i, p_{\text{pl}}) & \text{if not} \end{Bmatrix} \\ & \times P_{\text{pl,ing}}(p_{\text{pl}}, p_{\text{ing}}) P_{\text{feff}}(f_{\text{eff}}). \end{aligned} \tag{10}$$

To simplify the notation, we will write the core parameters and their prior as:

$$\boldsymbol{\pi} \equiv (p_{\text{ing}}, p_{\text{pl}}, f_{\text{eff}}) \tag{11}$$

$$P_{\pi}(\boldsymbol{\pi}) \equiv P_{\text{pl,ing}}(p_{\text{pl}}, p_{\text{ing}}) P_{\text{feff}}(f_{\text{eff}}). \tag{12}$$

It will be convenient to treat $\boldsymbol{\pi}$ and $\boldsymbol{q}$ separately.

To compute the posterior predictive probability that the next patient will respond to treatment with stool from donor i, we marginalize over the parameters $\boldsymbol{\pi}$ (i.e., integrate over p_{pl}, p_{ing}, and f_{eff}) and the qualities $\boldsymbol{q}$ (i.e., sum over the 2^D possibilities for the q_i). Let σ_i represent the event where the next patient responds when treated with stool from donor i. Then the posterior predictive probability is

$$P(\sigma_i|\boldsymbol{X}) = \int \sum_{\boldsymbol{q}} P(\sigma_i|\boldsymbol{\pi}, \boldsymbol{q}) \underbrace{P(\boldsymbol{\pi}, \boldsymbol{q}|\boldsymbol{X})}_{\text{posterior}} \, \mathrm{d}\boldsymbol{\pi} \tag{13}$$

The probability that the next patient responds to treatment is p_{eff} if the donor is efficacious or p_{pl} if not:

$$P(\sigma_i|\boldsymbol{\pi}, \boldsymbol{q}) = \begin{Bmatrix} p_{\text{eff}} & \text{if } q_i \text{ is "efficacious"} \\ p_{\text{pl}} & \text{if not} \end{Bmatrix}. \tag{14}$$

Combining Eqs. (10), (13) and (14) yields a computable predictive posterior:

$$\begin{aligned} P(\sigma_i|\boldsymbol{X}) \propto \int & \underbrace{[p_{\text{eff}} f_{\text{eff}} \operatorname{Bin}(s_i; n_i, p_{\text{eff}}) + p_{\text{pl}}(1 - f_{\text{eff}}) \operatorname{Bin}(s_i; n_i, p_{\text{pl}})]}_{\text{term for donor } i} \\ & \times \prod_{j \neq i} \underbrace{[f_{\text{eff}} \operatorname{Bin}(s_j; n_j, p_{\text{eff}}) + (1 - f_{\text{eff}}) \operatorname{Bin}(s_j; n_j, p_{\text{pl}})]}_{\text{terms for other donors}} \times \underbrace{P_\pi(\boldsymbol{\pi})}_{\text{priors}} \, \mathrm{d}\boldsymbol{\pi} \end{aligned} \tag{15}$$

7.2 Computing the predictive posterior

Eq. (15) is unwieldy and there are some notational and computational simplifications that can make it easier to use. First define

$$\boldsymbol{s} \equiv \{s_1, s_2, \ldots, s_D\}, \tag{16}$$

where s_i is the number of "successes" for donor i (i.e., the number of patients who responded to treatment with stool from donor i) and

$$\boldsymbol{f} \equiv \{f_1, f_2, \ldots, f_D\}, \tag{17}$$

the number of "failures" for donor i. We then define a shorthand $Q(\boldsymbol{s}; \boldsymbol{f})$ for dealing with all these data and parameters:

$$\begin{aligned} Q(\boldsymbol{s}; \boldsymbol{f}) \equiv & \\ & \int \prod_{i=1}^{D} \left[f_{\text{eff}} p_{\text{eff}}^{s_i} (1 - p_{\text{eff}})^{f_i} + (1 - f_{\text{eff}}) p_{\text{pl}}^{s_i} (1 - p_{\text{pl}})^{f_i} \right] \times P_\pi(\boldsymbol{\pi}) \, \mathrm{d}\boldsymbol{\pi}. \end{aligned} \tag{18}$$

Lemma 1. *The posterior predictive probability that the next patient will respond to treatment with stool from donor 1 is*

$$\frac{Q(s_1 + 1, s_2, s_3, \ldots; \boldsymbol{f})}{Q(\boldsymbol{s}; \boldsymbol{f})}. \tag{19}$$

It could just as well been any donor i; no generality is lost.

Proof. Let σ_1 be the event where donor 1 cures the next patient. We are trying to show that

$$P(\sigma_1 | \boldsymbol{s}, \boldsymbol{f}) = \frac{Q(s_1 + 1, s_2, s_3, \ldots; \boldsymbol{f})}{Q(\boldsymbol{s}; \boldsymbol{f})}. \tag{20}$$

The predictive posterior probability requires integrating over $\boldsymbol{\pi}$ and summing over $\boldsymbol{q}$:

$$P(\sigma_1 | \boldsymbol{s}, \boldsymbol{f}) \propto \int \sum_{\boldsymbol{q}} P(\sigma_1 | \boldsymbol{\pi}, \boldsymbol{q})\, P(\boldsymbol{\pi}, \boldsymbol{q} | \boldsymbol{s}, \boldsymbol{f})\, \mathrm{d}\boldsymbol{\pi}. \tag{21}$$

The first term is easy, and follow directly from our model assumptions:

$$P(\sigma_1 | \boldsymbol{s}, \boldsymbol{f}) = \begin{Bmatrix} p_{\text{eff}} & \text{if donor 1 is efficacious} \\ p_{\text{pl}} & \text{otherwise} \end{Bmatrix} \tag{22}$$

The second probability is the posterior probability of the parameters:

$$P(\boldsymbol{\pi}, \boldsymbol{q} | \boldsymbol{s}, \boldsymbol{f}) \propto P(\boldsymbol{s}, \boldsymbol{f} | \boldsymbol{\pi}, \boldsymbol{q})\, P(\boldsymbol{\pi}, \boldsymbol{q}). \tag{23}$$

Each donor is independent of the others, and each patient's response to treatment is a Bernoulli trial, so the first probability (the likelihood) is a product of binomial probability densities:

$$P(\boldsymbol{s}, \boldsymbol{f} | \boldsymbol{\pi}, \boldsymbol{q}) = \prod_{i=1}^{D} \begin{Bmatrix} \text{Bin}(s_i; n_i, p_{\text{eff}}) & \text{if donor } i \text{ efficacious} \\ \text{Bin}(s_i; n_i, p_{\text{pl}}) & \text{if not} \end{Bmatrix} \tag{24}$$

The constants in the binomials depend on the $\boldsymbol{s}$ and $\boldsymbol{f}$, not on any of the varying parameters in the integral/sum, so they can be pulled out:

$$P(\boldsymbol{s}, \boldsymbol{f} | \boldsymbol{\pi}, \boldsymbol{q}) \propto \prod_{i=1}^{D} \begin{Bmatrix} p_{\text{eff}}^{s_i}(1 - p_{\text{eff}})^{f_i} & \text{if donor } i \text{ efficacious} \\ p_{\text{pl}}^{s_i}(1 - p_{\text{pl}})^{f_i} & \text{if not} \end{Bmatrix} \tag{25}$$

The prior $P(\boldsymbol{\pi}, \boldsymbol{q})$ in Eq. (23) broken up to reveal the more familiar prior $P_\pi(\boldsymbol{\pi})$:

$$P(\boldsymbol{\pi}, \boldsymbol{q}) = \prod_{i=1}^{D} \begin{Bmatrix} f_{\text{eff}} & \text{if donor } i \text{ is efficacious} \\ 1 - f_{\text{eff}} & \text{if not} \end{Bmatrix} \times P_\pi(\boldsymbol{\pi}) \tag{26}$$

Some algebra shows that the products in the definitions for $P(\boldsymbol{\pi}, \boldsymbol{q})$ and $P(\boldsymbol{s}, \boldsymbol{f}|\boldsymbol{\pi}, \boldsymbol{q})$ move nicely through the sum in the definition of the posterior probability, and the extra p_{eff} or p_{pl} from the predictive probability moves right inside the donor 1 term:

$$\begin{aligned} P(\sigma_1|\boldsymbol{s}, \boldsymbol{f}) \propto \int & \left[f_{\text{eff}} p_{\text{eff}}^{s_1+1} (1 - p_{\text{eff}})^{f_1} + (1 - f_{\text{eff}}) p_{\text{pl}}^{s_1+1} (1 - p_{\text{pl}})^{f_1} \right] \\ & \times \prod_{i=2}^{D} \left[f_{\text{eff}} p_{\text{eff}}^{s_i} (1 - p_{\text{eff}})^{f_i} + (1 - f_{\text{eff}}) p_{\text{pl}}^{s_i} (1 - p_{\text{pl}})^{f_i} \right] P(\boldsymbol{\pi}) \, \mathrm{d}\boldsymbol{\pi} \end{aligned} \tag{27}$$

But this is just Q as if donor 1 already had another success:

$$P(\sigma_1|\boldsymbol{s}, \boldsymbol{f}) \propto Q(s_1 + 1, s_2, s_3, \ldots; \boldsymbol{f}). \tag{28}$$

A little more algebra along these lines will show that the posterior of donor 1 causing a failure means replacing the p_{eff} and p_{pl} with $1 - p_{\text{eff}}$ and $1 - p_{\text{pl}}$, which shows that

$$Q(s_1 + 1, s_2, s_3, \ldots; \boldsymbol{f}) + Q(\boldsymbol{s}; f_1 + 1, f_2, f_3, \ldots) = Q(\boldsymbol{s}, \boldsymbol{f}), \tag{29}$$

and therefore that the denominator shown in the lemma statement is the right one. □

7.3 New donors in the predictive posterior

Donors that have no successes or failures (i.e., whose stool has not been administered to any patient) are "hiding in the background" in $Q(\boldsymbol{s}; \boldsymbol{f})$. A new donor with $s_i = f_i = 0$ does not contribute to the product over donors i shown in (18):

$$f_{\text{eff}} p_{\text{eff}}^{s_i} (1 - p_{\text{eff}})^{f_i} + (1 - f_{\text{eff}}) p_{\text{pl}}^{s_i} (1 - p_{\text{pl}})^{f_i} = f_{\text{eff}} + (1 - f_{\text{eff}}) = 1. \tag{30}$$

Thus, there is nothing special in the math about new donors.

7.4 Incorporating placebo arm data

Information from the placebo arm of an ongoing trial can be incorporated just like the successes and failures of each donor. In this way, the results from the placebo arm can be used to continuously refine the posterior distribution on p_{pl}.

The placebo arm is mathematically equivalent to a donor that we know is not efficacious. Whereas most donors get terms in $Q(\boldsymbol{s};\boldsymbol{f})$ that are equal to $[f_{\text{eff}}p_{\text{eff}}^{s_i}(1-p_{\text{eff}})^{f_i}+(1-f_{\text{eff}})p_{\text{pl}}^{s_i}(1-p_{\text{pl}})^{f_i}]$, the placebo arm "donor" gets a term $p_{\text{pl}}^{s_{\text{pl}}}(1-p_{\text{pl}})^{f_{\text{pl}}}$, where s_{pl} is the number of patients in the placebo arm who responded and f_{pl} is the number who did not.

8 The myopic Bayesian strategy

A donor selection *strategy* determines what donor should be used next in light of the history of the trial thus far (i.e., the number of successes and failures attributed to each donor), the priors on the model parameters, and the number of patients remaining. Technically speaking, the strategy is a function that takes those inputs and returns the identity of the donor to be used next.

In the *myopic* (or "greedy") strategy, each donor is selected because, at that moment, they appear to be the one that maximizes the probability that the next patient will respond to treatment. Thus, before each patient is assigned, Eq. (15), and the donor i that maximizes $P(\sigma_i|\boldsymbol{X})$ is used for that patient. This kind of algorithm is called "myopic" or "greedy" because it makes the best immediate choice, which is not necessarily the choice that will lead to the optimal outcome for the entire trial [2, 1]. The optimal strategy will be examined below.

The myopic strategy has a property that makes it intuitive in many scenarios: if some donor has at least as many successes as and no more failures than any other donor, then that donor is the myopic choice.

Theorem 1. *If $s_i \geq s_j$ and $f_i \leq f_j$, then $P(\sigma_i|\boldsymbol{X}) > P(\sigma_j|\boldsymbol{X})$.*

Proof. Without loss of generality, let $i = 1$ and $j = 2$. By Lemma 1, the difference in predictive posteriors is proportional to the difference of two Q values:

$$P(\sigma_1|\boldsymbol{X})-P(\sigma_2|\boldsymbol{X}) \propto Q(s_1+1,s_2,s_3,\ldots;\boldsymbol{f})-Q(s_1,s_2+1,s_3,\ldots;\boldsymbol{f}). \quad (31)$$

Eq. (18) shows that the difference between these two Q values will be a new integral. The terms in the integral corresponding to the other donors ($i > 2$) and the prior on $\boldsymbol{\pi}$ will remain the same. Only the product of the $i = 1$ and $i = 2$ terms will change. Those terms will be:

$$\begin{aligned}(1 \text{ term with an extra success}) \times (\text{normal 2 term}) - \\ (\text{normal 1 term}) \times (2 \text{ term with an extra success}).\end{aligned} \quad (32)$$

In gory detail:

$$\begin{aligned}\left[f_{\text{eff}} p_{\text{eff}}^{s_1+1}(1-p_{\text{eff}})^{f_1} + (1-f_{\text{eff}}) p_{\text{pl}}^{s_1+1}(1-p_{\text{pl}})^{f_1}\right] \times \\ \left[f_{\text{eff}} p_{\text{eff}}^{s_2}(1-p_{\text{eff}})^{f_2} + (1-f_{\text{eff}}) p_{\text{pl}}^{s_2}(1-p_{\text{pl}})^{f_2}\right] - \\ \left[f_{\text{eff}} p_{\text{eff}}^{s_1}(1-p_{\text{eff}})^{f_1} + (1-f_{\text{eff}}) p_{\text{pl}}^{s_1}(1-p_{\text{pl}})^{f_1}\right] \times \\ \left[f_{\text{eff}} p_{\text{eff}}^{s_2+1}(1-p_{\text{eff}})^{f_2} + (1-f_{\text{eff}}) p_{\text{pl}}^{s_2+1}(1-p_{\text{pl}})^{f_2}\right].\end{aligned} \quad (33)$$

This produces two terms with f_{eff}^2, two with $(1-f_{\text{eff}})^2$, and four with $f_{\text{eff}}(1-f_{\text{eff}})$. The f_{eff}^2 terms cancel: they both have $p_{\text{eff}}^{s_1+s_2+1}$. Similarly, the $(1-f_{\text{eff}})^2$ terms cancel: they both have $p_{\text{pl}}^{s_1+s_2+1}$. This leaves four terms:

$$\begin{aligned}f_{\text{eff}}(1-f_{\text{eff}}) \times \\ \Big[p_{\text{eff}}^{s_1+1}(1-p_{\text{eff}})^{f_1} p_{\text{pl}}^{s_2}(1-p_{\text{pl}})^{f_2} + p_{\text{eff}}^{s_2}(1-p_{\text{eff}})^{f_2} p_{\text{pl}}^{s_1+1}(1-p_{\text{pl}})^{f_2} - \\ p_{\text{eff}}^{s_1}(1-p_{\text{eff}})^{f_1} p_{\text{pl}}^{s_2+1}(1-p_{\text{pl}})^{f_2} - p_{\text{eff}}^{s_2+1}(1-p_{\text{eff}})^{f_2} p_{\text{pl}}^{s_1}(1-p_{\text{pl}})^{f_2}\Big].\end{aligned} \quad (34)$$

These terms can be rearranged into:

$$\begin{aligned}f_{\text{eff}}(1-f_{\text{eff}})(p_{\text{eff}}-p_{\text{pl}}) p_{\text{eff}}^{s_1}(1-p_{\text{eff}})^{f_1} p_{\text{pl}}^{s_2}(1-p_{\text{pl}})^{f_2} \times \\ \left[1 - \left(\frac{p_{\text{pl}}}{p_{\text{eff}}}\right)^{s_1-s_2} \left(\frac{1-p_{\text{eff}}}{1-p_{\text{pl}}}\right)^{f_2-f_1}\right].\end{aligned} \quad (35)$$

So long as $s_1 \geq s_2$ and $f_1 \leq f_2$, then all these terms are positive. Because every other term in the integral is positive, the difference in Eq. (31) will be positive, so $P(\sigma_i|\boldsymbol{X}) > P(\sigma_j|\boldsymbol{X})$. □

This is a theorem about the myopic Bayesian strategy. Even though donor A might be better than B (in the sense that $s_A \geq s_B$ and $f_A \leq f_B$), it could be that somehow, it is more favorable to the trial's final outcome to

assign the next patient to B. That possibility will be discussed below when comparing the myopic and optimal Bayesian strategies.

This theorem has a simple corollary that shows that there is a common situation when the best-choice donor is a new donor (one that has not yet been used to treat any patients).

Corollary 1. *A patient is more likely to respond to treatment with stool from a new donor than to treatment with stool from a donor with no successes and at least one failure.*

Thus, the myopic Bayesian strategy requires that, at the beginning of the trial, you use a new donor until you get at least one positive patient outcome.

9 The optimal Bayesian strategy

A strategy is *optimal* with respect to some *utility function* and some true model parameters (or distribution of model parameters) if there is no other strategy that has a higher expected utility averaged over the donor's random outcomes (or also averaged over values of the model parameters drawn from those true distributions). In other words, a strategy is the optimal one if, in performing many simulated trials, the strategy makes decisions about donor assignment that lead to the best average results.

9.1 Trial trees

The greedy strategy has the advantage that, when implemented, it need only compute the answer to a small number of questions: for each patient, where should that patient be assigned? The optimal strategy, on the other hand, needs to look ahead to all possible outcomes. It is easy to express these possibility in terms of a tree that represent the trial's possible outcomes.

The root of the tree (denoted as $\varnothing$) is the current trial state. The root has a number of child *donor nodes*, which represent the possibility of choosing to allocate the next patient to each donor. In the case where there are two available donors, call these nodes A and B. Each donor node has two *state nodes*. The left child represents the new trial state in which the patient responded to treatment (i.e., the donor "succeeded"); the right child represents the new trial state in which the patient did not respond. We label these nodes with the donor node label and either s or f to indicate success

or failure. (This is an adaptation of the notation used by Zelen [5]). A tree representing a trial with two donors A and B and two remaining patients is shown in Figure 1.

A strategy's expected utility is the expected utility of the root (trial state) node. A state node's expected utility is just the expected utility of the donor choice node corresponding to the donor that the strategy would select given that trial state. The expected utility of a donor choice node is the weighted average of the expected utilities of its two (trial state) children. For example, if the next patient will respond to stool from donor A with probability p and the expected utilities of the two outcome nodes As and Af as $U(As)$ and $U(Af)$, then the exected utility of the donor choice node A is $pU(As) + (1-p)U(Af)$.

The recursion ends at the leaves (the trial state in which there are no remaining patients), which are each assigned a utility by the utility function. One choice for utility is the number of patients that responded to treatment. Another choice would be to assign a utility 0 to a leaf if the outcome is not statistically significant and assign a utility 1 is if the outcome is significant.

9.2 Counting trial states

Because the optimal strategy will require recursion through the entire tree, it will eventually require computing the posterior predictive probability for every possible trial state. For modest numbers of patients (e.g., 30), the number of trial states is large (tens of millions).

How many unique trial states are there for a given number of patients N? We specify "unique" because two donors with the same number of successes and the same number of failures are indistinguishable. If donors were distinguishable, then the number of unique trial states would be the number of ways of distributing the indistinguishable N balls (i.e., the patients) among $2N$ bins (i.e., donor A successes, donor A failures, donor B successes, etc.). This value is described by the multiset number. In our case, the donors are indistinguishable, but they have two distinguishable "sub-bins": you can tell a donor who has 1 success and 0 failures from a donor who has 0 successes and 1 failure, but you can't tell apart two donors who each have 1 success and 1 failure.

First, we show a formula to compute the number of ways to put n indistinguishable balls into indistinguishable bins.

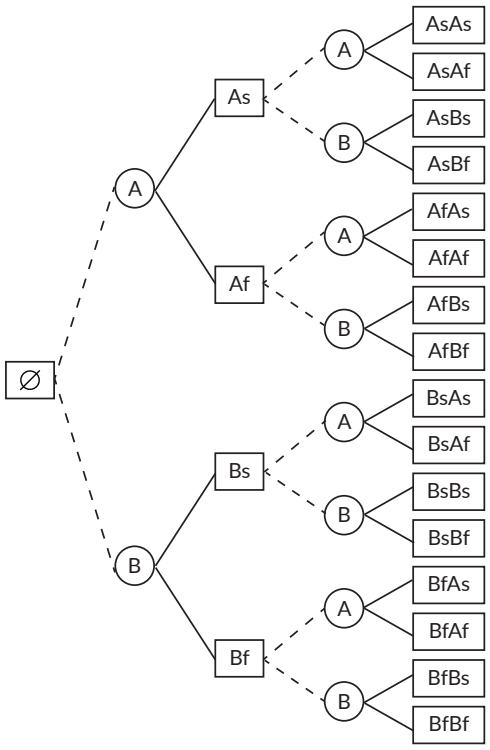


Figure 1: **A depth-2 trial tree.** Trial states are marked in square boxes; donor decision nodes in circles. Starting at the initial trials state $\varnothing$, donor A or B can be assigned to the next patient. If donor A is assigned, that patient can either experience a success (leading to trial state As) or a failure (Af). The solid line have probabilities associated with them: given the priors and the donors' histories, there is some probability associated with the transition between $\varnothing$ and As, i.e., the probability that the next patient will respond to treatment with stool from donor A. The dotted lines indicate a choice: different donor strategies can assign different donors to the next patient given the same trial state.

Lemma 2. *Let the number of ways to distribute n indistinguishable balls into indistinguishable bins with at most m balls per bin be written $W(n, m)$. The number of ways to distribute n indistinguishable balls into indistinguishable bins is $W(n, n)$, where $W(n, m)$ is defined by the recursion:*

$$W(0, m) = W(1, m) = 1 \tag{36}$$

$$W(n, 1) = 1 \tag{37}$$

$$W(n, m) = \sum_{i=0}^{\lfloor \frac{n}{m} \rfloor} W(n - im, m - 1) \tag{38}$$

Proof. Because the bins are only distinguished by the number of balls they contain, we can enumerate the distributions by writing the number of bins that have a certain number of balls. For example, for the $n = 4$ case, there are 5 ways to distribute the balls:

1. 1 bin has 4 balls.
2. 1 bin has 3 balls; 1 bin has 1 ball.
3. 2 bins have 2 balls.
4. 1 bin has 2 balls; 2 bins have 1 ball.
5. 4 bins have 1 ball.

This enumeration suggests a recursion: for some m, the number of bins that could have m balls is $i \in \left\{0, 1, \ldots, \lfloor \frac{n}{m} \rfloor\right\}$. For each of those situations, we ask how many ways there are to arrange the remaining $n - im$ balls among bins with at most $m - 1$ balls per bin.

For $n = 0$, there is only one arrangement: all the bins have 0 balls. For $n = 1$, there is also only one arrangement: one bin has one ball. If $m = 1$, there is only one arrangement: n bins having 1 ball. □

We can extend this logic to ask the number of ways to distibute the N patients among the (at most N) indistinguishable donors.

Theorem 2. *Let a double-bin be an ordered pair of integers (i.e., a bin with distinguishable "success" and "failure" sub-bins). Let $H(n, m)$ be the number of ways to put n indistinguishable balls into double-bins with at most m balls*

per double-bin. Then the number of ways to put n indistinguishable balls into double-bins is $H(n, n)$ where $H(n, m)$ is defined by the recursion:

$$H(0, m) = 1 \tag{39}$$
$$H(1, m) = 2 \tag{40}$$
$$H(n, 1) = n + 1 \tag{41}$$
$$H(n, m) = \sum_{i=0}^{\lfloor \frac{n}{m} \rfloor} \binom{m+i}{m} \times H(n - im, m - 1) \tag{42}$$

Proof. This proof is like the lemma except that there is an extra degree of freedom. Now the bins with m donors fall into $m + 1$ subcategories: they can have 0 in the left sub-bin and m in the right, 1 in the left and $m - 1$ in the right, and so forth, up to m in the left and 0 in the right. The number of ways to distribute the i bins that have m donors among the possible $m + 1$ subcategories is $\binom{m+i}{m}$.

For $n = 0$, there is only one way to distribute the balls: 0 successes and 0 failures. For $n = 1$, there are two ways to distribute the balls: either 1 success or 1 failure. For $m = 1$, there are $n + 1$ ways to distribute the balls: 0 successes and n failures, 1 success and $n - 1$ failures, and so forth, up to n successes and 0 failures. □

Note that this recursion relationship for $H(n, n)$ will yield the same number as an algorithmic approach:

1. Make a string of n stars and $2n - 1$ bars.
2. For every possible permutation of that string, make a list of the length of run of stars. (This is equivalent to enumerating all ways to write n non-negative integers that add up to n.)
3. Group that list of run lengths into pairs. (These are the left and right sub-bins.)
4. Sort the list. (This accounts for indistinguishability of the double-bins.)
5. Count the number of unique sorted lists of tuples.

This approach is slow because it requires enumerating all $(3n - 1)!$ permutations of the stars and bars.

# patients	# trial states
1	2
2	6
3	14
4	33
5	70
6	149
7	298
8	591
9	$1,132$
10	$2,139$
15	$39,894$
20	5.58×10^5
25	6.39×10^6
30	6.29×10^7
40	4.34×10^9
50	2.14×10^{11}
60	8.22×10^{12}

Table 3: The number of possible trial states for a given number of patients, assuming that the number of available donors is equal to the number of patients.

Some representative results are shown in Table 3. These results argue for the difficulty of an exact, optimal calculation: for a moderate-size trial with 30 donors, the tree will have 60 million unique leaf trial states.

9.3 Optimality of the myopic strategy

9.3.1 Utility in an optimal strategy

Is the myopic strategy optimal with respect to some utility function and priors? The myopic strategy locally optimizes the number of patients with successful outcomes, so we will compare the myopic strategy against the strategy that is optimal with respect to the number of patients with successful outcomes at the end of the trial.

For convenience, we will merge the notation $Q(\boldsymbol{s}; \boldsymbol{f})$ and the node labels

in Figure 1 to define:

$$Q(\varnothing) \equiv Q(\boldsymbol{s}; \boldsymbol{f}) \tag{43}$$
$$Q(As) \equiv Q(s_A + 1, s_B, s_C, \ldots; \boldsymbol{f}) \tag{44}$$
$$Q(AsAf) \equiv Q(s_A + 1, s_B, s_C, \ldots; f_A + 1, f_B, f_C, \ldots) \tag{45}$$

and so forth. For example, As can be read as "donor $\underline{A}$ has had a success." One line of algebra shows that the probability of donor A having two successes, for example, is $Q(AsAs)/Q(\varnothing)$.

As mentioned above, the optimal strategy propagates expected utilities up through the tree, weighting the utilities by the probability of the patient outcome. Thus, the utility of a donor node is:

$$\begin{aligned} U(A) &= P(As)U(As) + P(Af)U(Af) \\ &= \frac{1}{Q(\varnothing)} \left[Q(As)U(As) + Q(Af)U(Af) \right]. \end{aligned} \tag{46}$$

The utility of a state node is the utility of its child donor node that has the greatest utility:

$$U(\varnothing) = \max \left\{ U(A), U(B) \right\}. \tag{47}$$

The myopic strategy is optimal if $P(As)$ being greater than (or equal to) $P(Bs)$ for all other donors B implies that $U(A) \geq U(B)$ for all other donors B, that is, that no donor has a greater utility than the donor who is most likely to succeed.

We do not have a proof for the conjecture that the myopic strategy is optimal with respect to the stated utility function. That conjecture may indeed be false.

9.3.2 Optimality of the myopic strategy for two patients remaining

To simplify the question, we concentrate on the case in which there are two remaining patients and only two donors (A and B) to choose from.

Theorem 3. *In a trial where:*

- *there are two patients remainings*
- *there are two donor (where donor A has had s_A successes and f_A failures and donor B has had s_B successes and f_B failures),*

- $s_A \geq s_B$ *and* $f_A \leq f_B$, *and*
- *the utility is an increasing function of the number of patients with successful outcomes,*

then A is the optimal donor.

Proof. We need to show that $U(A) \geq U(B)$. (If $U(A) = U(B)$, then both donor choices are optimal.)

First, define the utilities $U_0 < U_1 < U_2$, where U_0 is the utility of final trial outcomes in which neither of the two patients responded, U_1 is the utility for one patient responding, and U_2 is the utility for two patients responding.

Next, expand the definitions of expected utility and substitute these utility values. Expanding Eq. (46),

$$\begin{aligned} U(A) = \frac{1}{Q(\varnothing)} \Big\{ & \max \Big[Q(AsAs)U(AsAs) + Q(AsAf)U(AsAf), \\ & \qquad Q(AsBs)U(AsBs) + Q(AsBf)U(AsBf) \Big] \\ & + \max \Big[Q(AfAs)U(AfAs) + Q(AfAf)U(AfAf), \\ & \qquad Q(AfBs)U(AfBs) + Q(AfBf)U(AfBf) \Big] \Big\}. \end{aligned} \tag{48}$$

This is the sum of two terms, each of which is the maximum of two other terms, so the entire value is the maximum over four terms, each with four subterms. Making that replacement and converting the, e.g., $U(AsAs)$ into U_2 yields:

$$\begin{aligned} U(A) = \frac{1}{Q(\varnothing)} \max \Big\{ & \\ & Q(AsAs)U_2 + 2Q(AsAf)U_1 + Q(AfAf)U_0, \\ & Q(AsAs)U_2 + [Q(AsAf) + QU(AfBs)]\, U_1 + Q(AfBf)U_0, \\ & Q(AsBs)U_2 + [Q(AsBf) + QU(AfAs)]\, U_1 + Q(AfAf)U_0, \\ & Q(AsBs)U_2 + [Q(AsBf) + QU(AfBs)]\, U_1 + Q(AfBf)U_0 \Big\}. \end{aligned} \tag{49}$$

Repeating the same algebra for $U(B)$ gives the same result, but with A and

B swapped:

$$
\begin{aligned}
U(B) = \frac{1}{Q(\varnothing)} \max \Big\{ & \\
& Q(BsBs)U_2 + 2Q(BsBf)U_1 + Q(BfBf)U_0, \\
& Q(BsBs)U_2 + \left[Q(BsBf) + QU(BfAs)\right] U_1 + Q(BfAf)U_0, \\
& Q(BsAs)U_2 + \left[Q(BsAf) + QU(BfBs)\right] U_1 + Q(BfBf)U_0, \\
& Q(BsAs)U_2 + \left[Q(BsAf) + QU(BfAs)\right] U_1 + Q(BfAf)U_0 \Big\}.
\end{aligned}
\tag{50}
$$

Note that, by Eq. (29), all the Q terms on each line in Eqs. (49) and (50) sum to $Q(\varnothing)$. Thus, each line in the two equations represents a different of U_2, U_1, and U_0.

To show that $Q(A) > Q(B)$, we need to be able to compare the Q coefficients in the two equations. We can use an approach like the one used in Theorem 1. In Eq. (35), a term $(p_{\text{eff}} - p_{\text{pl}})$ appeared. Performing the same algebra with different pairs of Q values will produce the same results but with that one term changed:

$$Q(As) - Q(Bs) \to (p_{\text{eff}} - p_{\text{pl}}) \tag{51}$$

$$Q(AsAs) - Q(BsBs) \to (p_{\text{eff}}^2 - p_{\text{pl}}^2) \tag{52}$$

$$Q(BfBf) - Q(AfAf) \to \left[(1 - p_{\text{pl}})^2 - (1 - p_{\text{eff}})^2\right] \tag{53}$$

By the same logic as was used in Theorem 1, because each of the terms on the right of the arrow is positive, the terms on the left are all positive. Thus, we have:

$$Q(AsAs) \geq Q(BsBs) \tag{54}$$

$$Q(AfAf) \leq Q(BfBf) \tag{55}$$

Using these two inequalities and the fact that the Q coefficients on each line in the big equations all sum to 1, we can see that the first line in Eq. (49) is at least as great as first line in Eq. (50). Similar comparisons shows that the second line in the first equation is at least as great as the second line in the second equation, that the third line in the first is at least as great as the third in the second equation, and that the fourth lines of both equations are equal to one another.

Thus, no matter which line in Eq. (50) is the maximum, there is a line in Eq. (49) that is at least as great, and therefore $U(A) \geq U(B)$. □

Thus, when there are only two patients remaining, in cases in which it is easy to assert that a donor is the myopic choice, that donor is also the optimal choice.

10 Multiple donor assignment

In FMT trials, the outcome from the patient is not immediate; some time will pass between the assignment of a patient to a donor and that patient's outcome. It can be that a patient will need to be assigned to a donor before the outcome from the previously-assigned patient is known.

Notable, the urn-based donor allocation strategy does not need any modification to work in these situations: balls are drawn from the urn to select donors, and balls are put into the urn once a patient outcome is known.

The tree formalism discussed in Section 9.1 assumes that only one patient is assigned at a time. In the case of donor being assigned in sequence before their outcomes are known, a few modifications are required. Say, for example, a patient has been assigned to donor A but the outcome is not yet known, and a new patient needs to be assigned to A or B. The same formalism described above can assigned expected utilities to the donor nodes under As and Af. Call AsA the donor node where the first patient was assigned to A, that patient was (or will be) successful, and the second patient is assigned to A as well. Similarly name AsB, AfA, and AfB. Then the expected utilities in question are the assignment of *two* donors, AA or AB, where AA's utility is a mixture over the utilities of AsA and AfA (and the mixing term is still A's probability of success) and AB's utility is, similarly, a mixture over AsB and AfB. This configuration is summarized in Figure 2.

For practical purposes, patients could be assigned in a way that optimizes the expected number of successes given the probabilities computed using the known outcomes. For example, if donor i has a probability of success $p_i \equiv P(\sigma_i|\boldsymbol{X})$ and donor j has $p_j \equiv P(\sigma_j|\boldsymbol{X})$, then the ratio of patients assigned to them should approach $\sqrt{p_i/p_j}$ (as per the calculation in [4], p. 195).

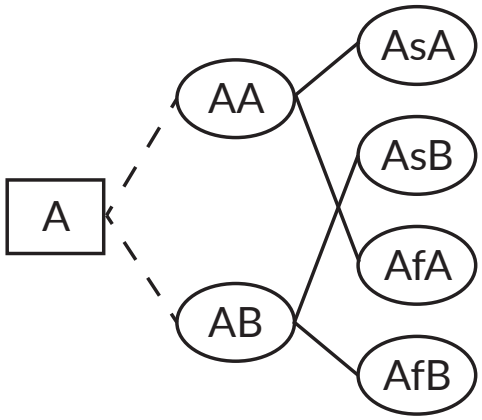


Figure 2: **A trial tree with two successive donor choices.** The current state of the trial, A, indicates that the next patient has been assigned to donor A but the outcome is not known. The dotted lines indicate that the expected utility of the current state A is the maximum over the utilities of the two donor choice nodes AA and AB. The solid lines indicate that the expected utility of AA is the weighted average of the utilities of AsA and AfA, whose utilities are computed as described previously.

References

[1] Donald A. Berry and Bert Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Springer Netherlands, 1985.

[2] Feifang Hu and William F. Rosenberger. *The Theory of Response-Adaptive Randomization in Clinical Trials*. Wiley, 2006.

[3] Paul Moayyedi, Michael G. Surette, Peter T. Kim, Josie Libertucci, Melanie Wolfe, Catherine Onischi, David Armstrong, John K. Marshall, Zain Kassam, Walter Reinisch, and Christine H. Lee. Fecal microbiota transplantation induces remission in patients with active ulcerative colitis in a randomized controlled trial. *Gastroenterology*, 149(1):102–109.e6, 2015.

[4] William F. Rosenberger and John M. Lachin. *Randomization in Clinical Trials*. Wiley, 2016.

[5] M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131–146, 1969.